THÉSE DE DOCTORAT DE L'UNIVERSITÉ PIERRE ET MARIE CURIE
École Doctorale de Sciences Mathématiques de Paris Centre

Méthodes numériques sur des grilles sparse appliquées à l'évaluation d'options en finance

présentée pour obtenir

LE GRADE DE DOCTEUR

DE L'UNIVERSITÉ PIERRE ET MARIE CURIE

Spécialité : Mathématiques

par

David Pommier

Soutenue le 28 novembre 2008 devant le jury composé de

M.	F. ABERGEL	(Examinateur)
M.	Y. ACHDOU	(Directeur de thèse)
M.	R. CONT	(Examinateur)
M.	B. LAPEYRE	(Rapporteur)
M.	T. LELIEVRE	(Examinateur)
M.	M. MASMOUDI	(Rapporteur)
M.	O. PIRONNEAU	(Examinateur)
Mme.	A. SULEM	(Examinateur)

Remerciements

J'aimerais exprimer ma profonde gratitude à Yves Achdou pour tout ce qu'il m'a apporté pendant ces quatre années. Je vous remercie pour votre disponibilité, votre gentillesse mais également pour la qualité de votre encadrement si sérieux et si rigoureux.

Je voudrais ensuite adresser mes remerciements à Agnès Sulem pour m'avoir accueilli au sein de l'équipe projet MATHFI et pour m'avoir chaleureusement soutenu.

Je remercie Bernard Lapeyre pour la lecture minutieuse de ce manuscrit et pour ses diverses suggestions, ainsi que Mohamed Masmoudi pour avoir accepté d'être rapporteur de ce travail.

J'exprime toute ma reconnaissance à Frédéric Abergel, Tony Lelièvre, Rama Cont et Olivier Pironneau pour l'honneur qu'ils me font en acceptant de faire partie du jury ainsi que pour les discussions que nous avons pu avoir.

Je tiens à saluer tous mes anciens collègues de l'équipe de recherche de BNP Parisbas Equities & Derivatives et, en particulier, ceux de l'équipe « EDP » : Xavier, Alexandre et Grégoire. Merci à chacun de vous pour votre bienveillance, la confiance que vous m'avez accordée et surtout nos discussions si instructives.

J'aimerais exprimer toute ma reconnaissance envers Stéphane Tyc pour m'avoir accueilli dans son équipe. Je remercie également Jean-Jacques et Bassam pour m'avoir fait profiter de leur expérience. J'ajouterais une mention spéciale à Julien, Love et Matthieu.

Je voudrais également remercier Antonino, Francesco, Peter, Fred et Aurélien d'avoir répondu à mes questions avec tant de gentillesse.

Merci à Nicolas, Maxime, Juliette, Céline et Jérome pour les discussions mathématiques, le soutien et les conseils éclairés. J'en profite pour remercier l'ensemble de mes amis dont le soutien et les encouragements m'ont été précieux.

Pour finir, je voudrais exprimer mon infinie gratitude à ma famille et, plus particulièrement, à mes parents pour tout ce qu'ils m'ont apporté.

Il ne me reste plus qu'à dire merci à Aurélie pour ton aide et ton soutien quotidien dans cette aventure.

Table des matières

Introduction

Cette thèse regroupe plusieurs travaux relatifs à la résolution numérique d'équations aux dérivées partielles et d'équations intégro-différentielles issues de la modélisation stochastique de produits financiers.

Les deux thématiques développées sont les méthodes de Sparse Grid appliquées à la résolution numérique d'équations en dimension supérieure à trois et les méthodes de raffinement de maillage appliquées à l'évaluation d'options américaines.

Ce document se décompose selon trois parties. Les principaux résultats concernant les méthodes de Sparse Grid sont présentés dans la première partie. Deux exemples d'application de ces méthodes à des problèmes posés en finance quantitative sont proposés dans la deuxième partie. La dernière partie est consacrée au développement d'indicateurs d'erreur a posteriori pour des problèmes de frontières libres ou d'inéquations variationnelles.

Méthodes numériques sur des Grilles Sparse

Les méthodes de Sparse Grid sont introduites par Korobov [Kor57] en 1957 pour l'approximation numérique d'intégrales en dimension grande. Ces travaux sont ensuite complétés par Babenko [Bab60] puis par Smolyak [Smo63]. Ce dernier définit la notion de produit tensoriel sparse. Les premiers résultats concernant l'interpolation d'une fonction régulière sur une base sparse sont énoncés dans l'article de Cavendish & al [CGH76]. Au début des années 90, Zenger [Zen91] publie les premiers travaux relatifs à la résolution d'équations aux dérivées partielles sur ces espaces sparse. Nos travaux s'intéressent à cette problématique.

La première partie s'articule selon deux chapitres. Les principaux résultats d'approximation de fonctions régulières sur les espaces sparse sont introduits au chapitre 1. Les différentes méthodes de résolution d'équations différentielles et intégro-différentielles appliquées à des problèmes elliptiques et paraboliques font l'objet du chapitre 2. Ces techniques sont récentes. Des résultats tels que la stabilité de la méthode des différences finies sparse ne sont pas, à notre connaissance, établis. Notre contribution consiste alors en l'élaboration d'une nouvelle démonstration de la consistance de l'opérateur de différences finies sparse. Cette approche permet notamment de définir un schéma de discrétisation consistant pour certains opérateurs intégraux.

Application des méthodes de Sparse Grid à la finance

Deux types de problèmes sont abordés. Le premier concerne l'évaluation d'options vanilles (call et put européens) dans un modèle à saut avec une volatilité stochastique multi-

facteurs. La résolution numérique de l'équation de valorisation est obtenue à l'aide d'une méthode de différence finie sparse. Celle-ci donne des résultats numériques intéressants et compétitifs par rapport à une méthode de Monte-Carlo. Le second problème traite de l'évaluation de produits multi sous-jacents. Il nécessite le recours à une méthode de Galerkin sur une base d'ondelettes sparse. Ces deux problèmes se distinguent par les propriétés de la fonction payoff.

Neuf chapitres constituent cette deuxième partie. Le chapitre 3 expose le cadre de ce travail : les problèmes posés en mathématiques financières y sont introduits ainsi que les méthodes usuelles employées pour les résoudre. Dans le chapitre 4, certains résultats connus concernant le lien entre les *équations différentielles stochastiques* (EDS) et les *équations aux dérivées partielles* (EDP) sont rappelés. Nous discutons, en particulier, de la formulation faible du problème déterministe associé à un processus de diffusion dégénéré. Le chapitre 5 est consacré au modèle à volatilité stochastique multi-facteurs. Des résultats d'existence et d'unicité sont démontrés à partir de la formulation faible du problème déterministe. Le chapitre 6 est, quant à lui, dédié aux expériences numériques pour lesquelles nous avons appliqué un algorithme de *différences finies sparse* au problème d'évaluation d'options. L'extension de ce modèle au cas d'un processus de diffusion à saut fait l'objet du chapitre 7. Nous formulons le problème déterministe associé au problème de valorisation d'options auquel nous appliquons une méthode de *différences finies sparse*. Le problème d'évaluation d'options multi sous-jacents est abordé au chapitre 8. Le problème déterministe initial est reformulé afin de pouvoir appliquer une *méthode de Galerkin* sur une *base d'ondelettes sparse*. Une implémentation de ces deux méthodes est proposée au chapitre 9.

Indicateurs d'erreur pour l'évaluation d'options américaines

Les méthodes d'adaptation de maillage sont des techniques éprouvées pour la résolution d'équations aux dérivées partielles. Le travail proposé ici concerne l'étude d'indicateurs d'erreur a posteriori pour des problèmes d'inéquations variationnelles. L'idée consiste à combiner les techniques connues pour les problèmes d'obstacle elliptiques et celles concernant les estimateurs d'erreur a posteriori mises en oeuvre pour les équations paraboliques. Les résultats obtenus pour l'évaluation d'une option américaine sur un panier de deux actifs sont présentés.

Cette partie est composée d'un unique chapitre.

Première partie

Méthodes numériques sur des grilles sparse

Chapitre 1

Approximation sur une base sparse

Ce chapitre est dédié aux propriétés d'approximation d'espaces sparse. Il s'agit d'espaces discrets de fonctions obtenus par tensorisation sparse d'espaces de dimension un.

La première partie introduit des espaces de Sobolev, notés *espaces des dérivées mixtes*, pour lesquels nous disposons de résultats d'approximation de fonctions régulières.

La deuxième partie constitue une étape préalable à la construction d'un produit tensoriel sparse. Les propriétés d'approximation des bases en dimension un, utilisées dans le produit tensoriel sparse, y sont présentées. Des résultats nécessaires pour la suite sont également énoncés : ils permettent, entre autres, d'établir la consistance des schémas de différences finies sur une grille sparse et de calculer la matrice de rigidité dans les méthodes de Galerkin. Les bases d'ondelettes définies sur $L^2(\mathbb{R})$ sont étendues à l'espace $L^2(\mathbb{R}^d)$ par application d'un produit tensoriel anisotrope.

La dernière partie est consacrée à la construction du produit tensoriel sparse proposée dans [GK00]. Ce procédé consiste à extraire un sous-ensemble de bases tensorielles anisotropes. Les résultats d'approximation sur les espaces sparse se déduisent des propriétés du produit tensoriel et des propriétés d'approximation des ondelettes en dimension un. Ces notions sont illustrées à travers l'étude d'une grille particulière, la *Sparse Grid* « classique ».

1.1 Espace de fonctions

1.1.1 Espace de Sobolev sur les ouverts de $\mathbb{R}^d$

Multi-indice Donnons quelques définitions et notations nécessaires à la présentation. Sur le compact $\Omega = [0,1]^d$, un point de Ω est noté $\boldsymbol{x} = (x_1, \ldots, x_d)$. Considérons une fonction u définie sur Ω à valeurs réelles. L'opérateur de dérivation est défini par :

$$D^{\mathbf{r}} u \stackrel{\text{def}}{=} \frac{\partial^{|\boldsymbol{r}|_1} u}{\partial^{r_1} x_1 \ldots \partial^{r_d} x_d}, \tag{1.1}$$

où $\boldsymbol{r} \in \mathbb{R}^d$ est le d-uplet $(r_1, \ldots, r_d)$ pour lequel deux normes sont définies

$$|\boldsymbol{r}|_1 = \sum_{i=1}^{d} r_i, \quad |\boldsymbol{r}|_\infty = \max_{1 \le i \le d} |r_i|. \tag{1.2}$$

L'ensemble des multi-indices est muni des opérations et relations suivantes :

– l'addition, la soustraction et le produit par un scalaire ;
– l'opération de puissance définie, pour tout réel a, par $a^{\boldsymbol{r}} = (a^{r_1}, \ldots, a^{r_d})$;
– la multiplication $\boldsymbol{r} \cdot \boldsymbol{q}$ définie pour deux multi-indices $\boldsymbol{r}$ et $\boldsymbol{q}$ comme le produit composante par composante :

$$\boldsymbol{r} \cdot \boldsymbol{q} \stackrel{\text{def}}{=} (\boldsymbol{r}_1 \boldsymbol{q}_1, \cdots, \boldsymbol{r}_d \boldsymbol{q}_d) ; \tag{1.3}$$

– une relation d'ordre, notée $\boldsymbol{r} \leq \boldsymbol{q}$ (*resp* $\boldsymbol{r} < \boldsymbol{q}$) et définie par

$$\boldsymbol{r} \leq \boldsymbol{q} \text{ si pour } 1 \leq i \leq d,\ \boldsymbol{r}_i \leq \boldsymbol{q}_i,\ (\textit{resp } \boldsymbol{r}_i \leq \boldsymbol{q}_i \text{ et } \exists\, i \text{ tel que } r_i < q_i). \tag{1.4}$$

Nous aurons également recours aux éléments particuliers de cet ensemble : $\mathbf{0} = (0, \ldots, 0)$, $\mathbf{1} = (1, \ldots, 1)$ et $\boldsymbol{e_i} = (0, \ldots, 0, 1, 0, \ldots, 0)$ le $i^{\text{ième}}$ vecteur unitaire de $\mathbb{R}^d$.

Espace de Sobolev d'ordre entier m Soient Ω un domaine borné de $\mathbb{R}^d$, i.e. un ouvert borné inclus dans $\mathbb{R}^d$, et $\partial\Omega$ sa frontière.

Pour un entier positif m, $\mathcal{C}^m(\Omega)$, (respectivement $\mathcal{C}^{m,1}(\Omega)$), représente l'espace des fonctions m fois différentiables sur Ω, et telles que, pour tout $|\boldsymbol{\alpha}|_1 \leq m$, $\partial^{\boldsymbol{\alpha}} v$ est continue (respectivement Lipschitz continue) sur Ω. L'espace $\mathcal{C}^m(\overline{\Omega})$ est l'espace des fonctions de $\mathcal{C}^m(\mathbb{R}^d)$ restreintes à Ω. L'espace $\mathcal{C}^m(\overline{\Omega})$ est un espace de Banach pour la norme :

$$\|v\|_{\mathcal{C}^m(\overline{\Omega})} = \max_{0 \leq |\boldsymbol{\alpha}|_1 \leq m} \sup_{\boldsymbol{x} \in \Omega} |\partial^{\boldsymbol{\alpha}} u(\boldsymbol{x})|.$$

Soient $\mathcal{D}(\Omega)$ (respectivement $\mathcal{D}(\mathbb{R}^d)$) l'espace des fonctions à valeurs réelles, indéfiniment différentiables, et de support compact inclus dans Ω (respectivement $\mathbb{R}^d$), et $\mathcal{D}(\overline{\Omega})$ l'espace des fonctions de $\mathcal{D}(\mathbb{R}^d)$ restreint à Ω.

Pour un réel $p \geq 1$, $L^p(\Omega)$ est l'espace des fonctions v à valeurs réelles, Lebesgue-mesurables sur Ω et telles que v^p est intégrable sur Ω. Nous définissons

$$\|v\|_{L^p(\Omega)} = \left(\int_\Omega |v|^p d\boldsymbol{x} \right)^{\frac{1}{p}}. \tag{1.5}$$

$L^\infty(\Omega)$ est l'espace des fonctions v à valeurs réelles, Lebesgue-mesurables sur Ω et essentiellement bornées sur Ω. Notons

$$\|v\|_{L^\infty(\Omega)} = \text{ess sup}\{|f(\boldsymbol{x})|; \boldsymbol{x} \in \Omega\}. \tag{1.6}$$

Pour $1 \leq p \leq \infty$, nous pouvons montrer que $\|.\|_{L^p(\Omega)}$ est une norme sur $L^p(\Omega)$.

Dans le cas $p = 2$, la norme $\|\cdot\|_{L^2(\Omega)}$ est une norme euclidienne et $L^2(\Omega)$ est un espace de Hilbert pour le produit scalaire

$$(u, v)_{L^2(\Omega)} = \int_\Omega u(\boldsymbol{x}) v(\boldsymbol{x}) d\boldsymbol{x}.$$

Définition 1.1 *Pour un nombre p tel que $1 \leq p \leq \infty$ et un entier positif m, l'espace de Sobolev $W^{m,p}(\Omega)$ est, par définition, l'espace des fonctions v dans $L^p(\Omega)$ dont toutes les dérivées partielles $\partial^{\boldsymbol{\alpha}} v, |\boldsymbol{\alpha}|_1 \leq m$ (prises au sens des distributions) appartiennent à $L^p(\Omega)$. Notons que $W^{0,p}(\Omega) = L^p(\Omega)$ et que, dans le cas $p = 2$, $W^{m,2}(\Omega)$ est noté*

$$H^m(\Omega) = W^{m,2}(\Omega).$$

Pour $v \in W^{m,p}(\Omega)$, nous définissons

$$\begin{aligned} \|v\|_{W^{m,p}(\Omega)} &= \Big(\sum_{0 \le |\boldsymbol{\alpha}|_1 \le m} \|\partial^{\boldsymbol{\alpha}} v\|^p_{L^p(\Omega)} \Big)^{\frac{1}{p}}, \quad si\ 1 \le p < \infty, \\ \|v\|_{W^{m,\infty}(\Omega)} &= \max_{0 \le |\boldsymbol{\alpha}|_1 \le m} \|\partial^{\boldsymbol{\alpha}} v\|_{L^\infty(\Omega)}, \end{aligned} \tag{1.7}$$

et $\|\|_{W^{m,p}(\Omega)}$ est une norme sur $W^{m,p}(\Omega)$. La norme $\|\|_{H^m(\Omega)}$ est une norme Euclidienne associée au produit scalaire

$$(v, w)_{H^m(\Omega)} = \int_\Omega \sum_{|\boldsymbol{\alpha}|_1 \le m} \partial^{\boldsymbol{\alpha}} v \partial^{\boldsymbol{\alpha}} w \, d\boldsymbol{x}.$$

Les semi-normes suivantes sont également définies pour $v \in W^{m,p}(\Omega)$ par

$$\begin{aligned} |v|_{W^{m,p}(\Omega)} &= \left(\sum_{|\boldsymbol{\alpha}|_1 = m} \|\partial^{\boldsymbol{\alpha}} v\|^p_{L^p(\Omega)} \right)^{\frac{1}{p}}, \quad \text{si } 1 \le p < \infty, \\ |v|_{W^{m,\infty}(\Omega)} &= \max_{|\boldsymbol{\alpha}|_1 = m} \|\partial^{\boldsymbol{\alpha}} v\|_{L^\infty(\Omega)}. \end{aligned} \tag{1.8}$$

L'espace $\mathcal{D}(\Omega)$ est inclus dans $W^{m,p}(\Omega)$. Soit $W_0^{m,p}(\Omega)$ l'adhérence de $\mathcal{D}(\Omega)$ dans $W^{m,p}(\Omega)$. L'inégalité de Poincaré-Friedrich indique que, pour p tel que $1 \le p \le \infty$, il existe une constante C telle que, pour tout $v \in W_0^{1,p}(\Omega)$,

$$|v|_{W^{1,p}(\Omega)} \le \|v\|_{W^{1,p}(\Omega)} \le C |v|_{W^{1,p}(\Omega)}. \tag{1.9}$$

Donc $|\cdot|_{W^{1,p}(\Omega)}$ est une norme sur $W_0^{1,p}(\Omega)$ équivalente à $\|\cdot\|_{W^{1,p}(\Omega)}$.

1.1.2 Espace de Sobolev d'ordre fractionnaire

Soient un réel p tel que $1 \le p < \infty$ et un réel positif s. s est supposé non entier : $m = \lfloor s \rfloor$ correspond à la partie entière de s et $\rho = s - m$. Soient $|\cdot|_{W^{s,p}(\Omega)}$ et $\|\cdot\|_{W^{s,p}(\Omega)}$ définies par

$$\begin{aligned} |v|_{W^{s,p}(\Omega)} &= \left(\sum_{|\boldsymbol{\alpha}|_1 = m} \int_{\boldsymbol{x} \in \Omega} \int_{\boldsymbol{y} \in \Omega} \frac{|\partial^{\boldsymbol{\alpha}} v(\boldsymbol{x}) - \partial^{\boldsymbol{\alpha}} v(\boldsymbol{y})|^p}{|\boldsymbol{x} - \boldsymbol{y}|^{d + \rho\, p}} d\boldsymbol{x} d\boldsymbol{y} \right)^{\frac{1}{p}}, \\ \|v\|_{W^{s,p}(\Omega)} &= \left(|v|^p_{W^{s,p}(\Omega)} + \|v\|_{W^{m,p}(\Omega)} \right)^{\frac{1}{p}}, \end{aligned} \tag{1.10}$$

si $p < \infty$, et (voir [GR86])

$$|v|_{W^{s,\infty}(\Omega)} = \max_{\alpha = m} \sup_{\boldsymbol{x}, \boldsymbol{y} \in \Omega, \boldsymbol{x} \ne \boldsymbol{y}} \frac{|\partial^\alpha v(\boldsymbol{x}) - \partial^\alpha v(\boldsymbol{y})|}{|\boldsymbol{x} - \boldsymbol{y}|^\rho},$$

$$\|v\|_{W^{s,\infty}(\Omega)} = \max \left(|v|_{W^{s,\infty}(\Omega)}, \|v\|_{W^{s,\infty}(\Omega)} \right).$$

Alors l'espace $W^{s,p}(\Omega)$ est, par définition, l'espace de Banach des fonctions v telles que $\|v\|_{W^{s,p}}$ est bornée. A nouveau, $H^s(\Omega) = W^{s,2}(\Omega)$ est un espace de Hilbert.

Le sous-espace $W_0^{s,p}(\Omega)$ est défini comme l'adhérence de $\mathcal{D}(\Omega)$ dans $W^{s,p}(\Omega)$. Soient p tel que $1 < p < \infty$, deux réels positifs s et r et un réel θ, tel que $0 \leq \theta \leq 1$. $W^{\theta s+(1-\theta)r,p}(\Omega)$ peut être obtenu en interpolant (interpolation réelle) $W^{s,p}(\Omega)$ et $W^{r,p}(\Omega)$, voir [LM61, Ada75, GR86].

Définition 1.2 (Méthode d'interpolation réelle de Lions-Peetre) *Soient :*
- *deux espaces de Banach X, Y tels que Y est inclus dans X avec injection dense ;*
- *la fonction K de $X \times \mathbb{R}^+$ à valeurs dans $\mathbb{R}^+$ et définie par*

$$K(v,t;X,Y) = \inf_{w \in Y} \left(\|v-w\|_X + t\|w\|_Y \right),$$

cette fonction est continue, non décroissante et concave par rapport à la variable t ;
- *$\theta \in (0,1)$ et $1 \leq p \leq \infty$.*

L'espace d'interpolation réelle $[X,Y]_{\theta,p}$ est l'espace des fonctions v de X telles que

$$\|v\|_{[X,Y]_{\theta,p}} = \left\| t^{-\theta} K(v,t;X,Y) \right\|_{L^p(\mathbb{R}^+, dt/t)} < \infty.$$

L'espace $L^p\left(\mathbb{R}^+, dt/t\right)$ est l'espace des fonctions L^p intégrables sur $\mathbb{R}^+$ par rapport à la mesure de Haar dt/t :

$$L^p\left(\mathbb{R}^+, dt/t\right) = \left\{ v \;\middle|\; \left(\int_{\mathbb{R}^+} v(t)^p \frac{dt}{t} \right)^{1/p} < \infty \right\}.$$

Cette définition permet d'écrire une seconde caractérisation des espaces de Sobolev d'ordre fractionnaire :

$$H^s(\Omega) = \left[L^2(\Omega), H^m(\Omega) \right]_{s/m,2}, \quad s \in (0,m). \tag{1.11}$$

Ce résultat reste valable pour s entier.

Elle nous permet également d'introduire les espaces de Besov $B^s_{p,q}(\Omega)$, définis par interpolation réelle des espaces L^p et $W^{r,p}$ où $r = \lfloor s \rfloor + 1$:

$$B^s_{p,q}(\Omega) = \left[L^p(\Omega), W^{r,p}(\Omega) \right]_{s/r,q}, \quad s \in (0,r). \tag{1.12}$$

Proposition 1.1 ([Tri92]) *Pour $1 < p < \infty$, $1 \leq q \leq \infty$ et $-\infty < s < \infty$, si $s = (1-\theta)s_0 + \theta s_1$ alors*

$$B^s_{p,q}(\Omega) = \left[W^{s_0,p}(\Omega), W^{s_1,p}(\Omega) \right]_{\theta,q}.$$

Cette proposition justifie, en particulier,(1.11).

Le lecteur trouvera dans [DeV98, Tri06] une description plus précise des espaces de Besov et des liens avec l'analyse multi-résolution. Nous ferons le lien entre ces espaces et les propriétés d'approximation non linéaire des bases d'ondelettes.

Nous donnons à présent une description partielle de la construction de l'interpolation par une méthode spectrale [Tri92]. Cette interprétation justifie l'introduction des espaces interpolés dans l'analyse des propriétés d'un opérateur elliptique.

Soient X et Y deux espaces de Hilbert tels que Y est inclus dans X avec injection dense. Ces espaces sont munis de leur norme respective $(\cdot,\cdot)_X$,$(\cdot,\cdot)_Y$. Nous supposons qu'il existe un opérateur linéaire non-borné A symétrique défini positif tel que :

– le domaine de A est l'espace Y, $A : D(A) = Y \to X$,
– A définit une norme équivalente sur Y : $\|u\|_Y^2 = \|u\|_X^2 + \|Au\|_X^2$.

Afin de simplifier l'exposé, nous supposerons que A possède un spectre discret $0 < a_1 \leq \cdots \leq a_n < \infty$ et des vecteurs propres $\{\nu_i\}$ qui forment une base orthogonale de X.

Proposition 1.2 *Si les hypothèses précédentes sont vérifiées, alors l'espace*

$$[X,Y]_{\theta,2} = [X,Y]_\theta = D\left(A^\theta\right) \quad 0 \leq \theta \leq 1,$$

est le domaine de l'opérateur A^θ, muni de la norme

$$\|u\|_{[X,Y]_\theta} = \left\|A^\theta u\right\|_X, \qquad où \ \left\|A^\theta u\right\|_X^2 = \sum_{i=1}^{\infty} a_i^{2\theta} \left(u, \nu_i\right)_X^2 .$$

Nous en déduisons le lemme suivant,

Lemme 1.3 *Pour tout $u \in Y$, $\|u\|_{[X,Y]_\theta} = C_\theta \|u\|_Y^\theta \|u\|_X^{1-\theta}$.*

Espace de Sobolev sur $\mathbb{R}^d$ L'espace de Sobolev $H^s\left(\mathbb{R}^d\right)$ peut être défini à l'aide de la transformée de Fourier.

Propriété 1.4 (Caractérisation par transformation de Fourier de $H^s\left(\mathbb{R}^d\right)$) *Soit un nombre réel s, la distribution w définie sur $\mathbb{R}^d$ appartient à $H^s\left(\mathbb{R}^d\right)$ si et seulement si sa transformée de Fourier $\hat{w}$ vérifie :*

$$\int_{\mathbb{R}^d} \left(1 + |\boldsymbol{\xi}|^2\right)^s |\hat{w}\left(\boldsymbol{\xi}\right)|^2 d\boldsymbol{\xi} < \infty. \tag{1.13}$$

L'espace $H^s\left(\mathbb{R}^d\right)$, muni du produit scalaire et de la norme

$$(u,v)_{H^s\left(\mathbb{R}^d\right)} = \int_{\mathbb{R}^d} \left(1 + |\boldsymbol{\xi}|^2\right)^s \hat{u}(\boldsymbol{\xi})\bar{\hat{v}}(\boldsymbol{\xi}) d\boldsymbol{\xi}, \quad \|u\|_{H^s(\mathbb{R})} = \sqrt{(u,u)_{H^s\left(\mathbb{R}^d\right)}}, \tag{1.14}$$

est un espace de Hilbert, et la norme définie par (1.14) est équivalente à celle définie par (1.10).

Pour les équations intégro-différentielles, il est possible d'écrire la formulation variationnelle au sens faible en utilisant des espaces de Sobolev à poids. L'existence et l'unicité de la solution sont démontrées par application du théorème de Lax-Milgram. Plusieurs choix d'espaces sont possibles afin d'obtenir des problèmes variationnels bien posés. Dans [FS06, Ach08], les auteurs proposent un exemple de ces espaces caractérisés à l'aide de la transformée de Fourier : pour s un nombre réel positif et ϕ une fonction continue strictement positive, l'espace $H^{\phi,s}$ est défini par

$$H^{\phi,s} = \left\{ w \in L^2\left(\mathbb{R}^d\right) \mid \int_{\mathbb{R}^d} \left(1 + \phi\left(\boldsymbol{\xi}\right)\right)^s |\hat{w}\left(\boldsymbol{\xi}\right)|^2 d\boldsymbol{\xi} < \infty \right\}. \tag{1.15}$$

Cet espace permet d'établir la formulation variationnelle dans le cas des *processus de Lévy*.

Le lecteur trouvera dans [AT02] un exemple d'espace de Sobolev à poids défini, cette fois, à partir des variables d'espaces. Celui-ci permet d'établir la formulation variationnelle dans le cas d'une équation parabolique dégénérée obtenue pour l'évaluation du prix d'options dans un modèle de diffusion à volatilité stochastique.

Espace de Hölder d'ordre fractionnaire Par la suite, nous exprimerons les erreurs de consistance des schémas de différences finies à travers des majorations mettant en jeu certaines normes de Hölder.

Soient $\boldsymbol{\alpha}$ appartenant à $\mathbb{R}_+^d$ et $[\boldsymbol{\alpha}]$ le multi-indice de $\mathbb{N}^d$ tel que la $i^{\text{ème}}$ composante admette la décomposition : $\alpha_i = [\alpha]_i + \{\alpha\}_i$, où $[\alpha_i]$ est la partie entière de α_i.

Soit $\mathcal{C}^{\boldsymbol{\alpha}}(\bar{\Omega})$ l'espace de Hölder des fonctions continues telles que, pour tout $\boldsymbol{\beta} \leq [\boldsymbol{\alpha}]$, $D^{\boldsymbol{\beta}}u$ est continue et

$$\sup\left\{\frac{|D^{[\boldsymbol{\alpha}]}u(\boldsymbol{x}+\boldsymbol{h}) - D^{[\boldsymbol{\alpha}]}u(\boldsymbol{x})|}{|h_1|^{\{\alpha_1\}}\dots|h_d|^{\{\alpha_d\}}}, \boldsymbol{x}, \boldsymbol{x}+\boldsymbol{h} \in \Omega, |h_i| > 0, i = 1, \dots, d\right\} < +\infty. \quad (1.16)$$

La dernière quantité correspond à la semi-norme sur $\mathcal{C}^{\boldsymbol{\alpha}}(\bar{\Omega})$, notée $|u|_{\mathcal{C}^{\boldsymbol{\alpha}}(\bar{\Omega})}$ (éventuellement $|u|_{\boldsymbol{\alpha}}$ lorsque le contexte ne présente aucune ambiguïté).

1.1.3 Espace de Sobolev avec des dérivées mixtes

A présent, nous introduisons un nouvel espace de Sobolev, noté $X^{m,p}(\Omega)$, sur lequel il est possible d'approcher une fonction sur une base sparse.

Définition 1.3 *Soient un réel p tel que $1 \leq p \leq \infty$ et m un entier positif. $X^{m,p}(\Omega)$ est l'espace des fonctions de $\Omega \subset \mathbb{R}^d$ dans $\mathbb{R}$ dont toutes les dérivées partielles d'ordre au plus m dans chacune des variables sont dans $L^p(\Omega)$,*

$$X^{m,p}(\Omega) = \left\{u \middle| D^{\boldsymbol{r}}u \in L^p(\Omega), |\boldsymbol{r}|_\infty \leq m\right\}. \quad (1.17)$$

Le sous-espace $X_0^{m,p}(\Omega)$, pour $m \geq 1$, est défini comme l'adhérence de $\mathcal{D}(\Omega)$ dans $X^{m,p}(\Omega)$,

$$X_0^{m,p}(\Omega) = \overline{\mathcal{D}(\Omega)}^{X^{m,p}(\Omega)}. \quad (1.18)$$

L'espace $\mathcal{H}^m(\Omega) = X^{m,2}(\Omega)$ est un espace de Hilbert, et

$$\mathcal{H}_0^m(\Omega) \stackrel{\text{def}}{=} \left\{u \in H_0^1(\Omega) \middle| D^{\boldsymbol{r}}u \in L^2(\Omega), |\boldsymbol{r}|_\infty \leq m\right\}. \quad (1.19)$$

Ce nouvel espace *des dérivées mixtes* vérifie :

$$H^m(\Omega) \subset \mathcal{H}^m(\Omega) \subset H^{d\,s}(\Omega). \quad (1.20)$$

Nous définissons, comme pour les espaces de Sobolev $W^{m,p}(\Omega)$ (1.8), les semi-normes pour $v \in X_0^{m,p}(\Omega)$:

$$\begin{aligned} |v|_{X^{m,p}(\Omega)} &= \left(\sum_{|\boldsymbol{\alpha}|_\infty = m} \|\partial^{\boldsymbol{\alpha}}v\|_{L^p(\Omega)}^p\right)^{\frac{1}{p}}, \quad \text{si } 1 \leq p < \infty, \\ |v|_{X^{m,\infty}(\Omega)} &= \max_{|\boldsymbol{\alpha}|_\infty = m} \|\partial^{\boldsymbol{\alpha}}v\|_{L^\infty(\Omega)}. \end{aligned} \quad (1.21)$$

Griebel [BG04] montre des équivalences de normes de type Poincaré-Friedrichs (1.9) avec les semi-normes définies par (1.21).

Supposons à présent que le domaine Ω est un hypercube *i.e.* $\Omega = \bigotimes_{i=1}^{d} \Omega_i$ où Ω_i est un intervalle de $\mathbb{R}$ (éventuellement $\mathbb{R}$ tout entier).

Considérons les espaces d'ordre fractionnaire : soit $\boldsymbol{r} \in (\mathbb{R}^+)^d$, l'espace de Hilbert $\mathcal{H}^{\boldsymbol{r}}(\Omega)$ est défini par :

$$\mathcal{H}^{\boldsymbol{r}}(\Omega) \stackrel{\text{def}}{=} H^{r_1}(\Omega_1) \otimes \cdots \otimes H^{r_d}(\Omega_d), \tag{1.22}$$

L'équation (1.22) généralise la définition 1.3 au cas des indices réels et à des espaces avec des régularités anisotropes. La définition de l'espace $\mathcal{H}^{t,s}(\Omega)$, introduit dans [GK00], est rappelée ci-dessous :

Définition 1.4 *Soit $t \in \mathbb{R}_0^+, t+s \geq 0$, alors*

$$\mathcal{H}^{t,s}(\Omega) \stackrel{\text{def}}{=} \mathcal{H}^{t\mathbf{1}+s\boldsymbol{e_1}}(\Omega) \cap \cdots \cap \mathcal{H}^{t\mathbf{1}+s\boldsymbol{e_n}}(\Omega). \tag{1.23}$$

Remarque 1.1 *L'espace de Sobolev des dérivées mixtes $\mathcal{H}^t(\Omega)$, $t \in \mathbb{R}_0^+$ correspond à l'espace $\mathcal{H}^{t,0}(\Omega)$. L'espace de Sobolev classique $H^s(\Omega)$, $s \in \mathbb{R}_0^+$ correspond à l'espace $\mathcal{H}^{0,s}(\Omega)$.*

Dans l'exemple suivant, une fonction de $H^1(\Omega)$ n'appartenant pas à $\mathcal{H}^1(\Omega)$ est exhibée.

Exemple 1.1 *Soit la fonction $h : \mathbb{R}^2 \to \mathbb{R}$ définie par $h(x_1, x_2) = \max(1 - (x_1 + x_2), 0)$, fonction payoff d'un put sur un panier de deux actifs. En dérivant, au sens des distributions, cette fonction, nous montrons que*

$$\frac{\partial h}{\partial x_i} = -1_{\{x_1+x_2 \leq 1\}} \in L^2(\Omega) \quad et \quad \frac{\partial^2 h}{\partial x_1 \partial x_2} = -\delta_{\{x_1+x_2=1\}} \notin L^2(\Omega), \tag{1.24}$$

alors

$$h \in H^1(\Omega) \quad et \quad h \notin \mathcal{H}^1(\Omega). \tag{1.25}$$

Proposition 1.5 *Soit u une fonction $\mathbb{R}^d \to \mathbb{R}$ à variables séparées,* i.e. *u se décompose en produit de fonctions uni-dimensionnelles, alors u appartient à $H^s(\Omega)$ si et seulement si u appartient à $\mathcal{H}^s(\Omega)$.*

Preuve Démontrons que si

$$u(\boldsymbol{x}) = u_1(x_1) \dots u_d(x_d), \quad \text{alors} \quad u \in H^s(\Omega) \Rightarrow u \in \mathcal{H}^s(\Omega). \tag{1.26}$$

La réciproque est triviale. Si u appartient à $H^s(\Omega)$, alors u_i appartient à $H^s(\Omega_i)$,

$$\frac{\partial^{\boldsymbol{r}} u}{\partial x^{r_1} \dots \partial x^{r_d}} = \prod_{i=1}^{d} \frac{\partial^{r_i} u_i}{\partial x^{r_i}}, \tag{1.27}$$

en appliquant le théorème de Fubini,

$$\left\| \frac{\partial^{\boldsymbol{r}} u}{\partial x^{r_1} \dots \partial x^{r_d}} \right\|_{L^2(\Omega)} = \prod_{i=1}^{d} \left\| \frac{\partial^{r_i} u_i}{\partial x^{r_i}} \right\|_{L^2(\Omega_i)}, \tag{1.28}$$

ce qui permet de conclure. ■

1.1.3.1 Espace de Hölder des dérivées mixtes

La théorie des *Sparse Grid* repose sur la notion des espaces des dérivées mixtes. Les résultats de convergence obtenus sur des espaces de Hölder pour la méthode « classique » des différences finies se généralisent sur les espaces de Hölder des dérivés mixtes pour la méthode de différences finies sparse.

Définition 1.5 *Soit $\mathcal{C}^m_{mix}(\bar{\Omega})$ l'espace de Hölder des dérivées mixtes, c.-à-d. l'ensemble des fonctions u continues telles que pour tout $|\boldsymbol{\beta}|_\infty \leq m$, $D^{\boldsymbol{\beta}}u$ est continue :*

$$\mathcal{C}^m_{mix}(\bar{\Omega}) \stackrel{\text{def}}{=} \left\{ u : D^{\boldsymbol{\beta}}u \in \mathcal{C}^0\left(\bar{\Omega}\right), |\boldsymbol{\beta}|_\infty \leq m \right\}. \tag{1.29}$$

Soit $\mathcal{C}^{m,K}_{mix}(\bar{\Omega})$ l'ensemble des fonctions telles que la norme de Hölder des dérivées mixtes soit bornée :

$$\mathcal{C}^{m,K}_{mix}(\bar{\Omega}) \stackrel{\text{def}}{=} \left\{ u \in \mathcal{C}^m_{mix}(\bar{\Omega}) \mid \quad \left\| D^{\boldsymbol{\beta}}u \right\|_{\mathcal{C}^0} \leq K \right\}. \tag{1.30}$$

Nous montrerons qu'il est possible d'obtenir sur ces espaces des résultats d'approximation pour un opérateur d'interpolation. Nous aurons alors recours au lemme suivant.

Lemme 1.6 *Soit $u \in \mathcal{C}^2_{mix}(\bar{\Omega})$, alors,*

$$\begin{aligned} u\left(x\right) = u(0,\ldots,0) + \sum_{i=1}^{d} x_i \left(\frac{\partial u}{\partial x_i}\right)_{\big|x_i=0}(x) - \sum_{i,j=1,\, i\neq j}^{d} x_i x_j \left(\frac{\partial^2 u}{\partial x_i \partial x_j}\right)_{\big|x_i=0, x_j=0}(x) + \ldots \\ + x_1 \ldots x_d \left(\frac{\partial^d u}{\partial x_1 \ldots \partial x_d}\right) u_{\big|x_1=0,\ldots,x_d=0}(x) + \sum_{i=1}^{d} \int_0^{x_i} \left(\frac{\partial^2 u}{\partial x_i^2}\right)_{\big|x_i=z_i}(0)(x_i - z_i)dz_i \\ + \cdots + \int_{0,\ldots,0}^{x_1,\ldots,x_d} \frac{\partial^{2d} u}{\partial x_1^2 \ldots \partial x_d^2}\left(z_1,\ldots,z_d\right)\left(x_1 - z_1\right)\ldots\left(x_d - z_d\right)dz_1\ldots dz_d, \end{aligned} \tag{1.31}$$

où $u_{\big|x_i=z}(y) = u\left(y_1,\ldots,y_{i-1},z,y_{i+1},\ldots,y_d\right).$

Preuve Nous démontrons ce résultat en dimension 2, nous admettrons la généralisation. Appliquons la formule de Taylor reste intégral à la variable x_1,

$$u\left(x_1,x_2\right) = u(0,x_2) + x_1\frac{\partial u}{\partial x_1}(0,x_2) + \int_0^{x_1} \frac{\partial^2 u}{\partial x_1^2}(z_1,x_2)(x_1 - z_1)dz_1, \tag{1.32}$$

puis sur la variable x_2, nous obtenons

$$u\left(z_1,x_2\right) = u(z_1,0) + x_2\frac{\partial u}{\partial x_0}(z_1,0) + \int_0^{x_2} \frac{\partial^2 u}{\partial x_2^2}(z_1,z_2)(x_2 - z_2)dz_2. \tag{1.33}$$

Le terme intégral de (1.32) devient

$$\begin{aligned}\int_0^{x_1} \frac{\partial^2 u}{\partial x_1^2}(z_1,x_2)(x_1-z_1)dz_1 = &\int_0^{x_1} \frac{\partial^2 u}{\partial x_1^2}(z_1,0)(x_1-z_1)dz_1 \\ &+ x_2 \int_0^{x_1} \frac{\partial^3 u}{\partial x_1^2 \partial x_2}(z_1,0)(x_1-z_1)dz_1 \\ &+ \int_0^{x_1}\int_0^{x_2} \frac{\partial^4 u}{\partial x_1^2 \partial x_2^2}(z_1,z_2)(x_1-z_1)(x_2-z_2)dz_1dz_2\end{aligned}$$

$$x_2 \int_0^{x_1} \frac{\partial^3 u}{\partial x_1^2 \partial x_2}(z_1,0)(x_1-z_1)dz_1 = x_2 \frac{\partial^2 u}{\partial x_1 \partial x_2}(0,0) + x_2 \frac{\partial u}{\partial x_2}(x_1,0) - x_2 \frac{\partial u}{\partial x_2}(0,0).$$

Il suffit alors de remplacer dans (1.32) le terme $u(0,x_2)$ par (1.33) avec $z_1 = 0$ et le terme intégral pour obtenir (1.31). ■

1.2 Approximation dans des bases d'ondelettes biorthogonales

Dans cette partie, les bases d'ondelettes sont décrites en insistant sur les propriétés d'approximation fondamentales dans la démonstration des propriétés d'approximation du produit tensoriel sparse. Nous donnons également certains résultats utilisés dans l'analyse des méthodes de résolution numérique d'équations aux dérivées partielles du chapitre suivant. La définition d'une analyse multi-résolution et ses propriétés sont rappelées avant d'introduire deux familles d'ondelettes : les ondelettes biorthogonales et les ondelettes interpolantes. Dans le dernier paragraphe, la définition des bases d'ondelettes obtenues en dimension 1 est généralisée au cas de la dimension d.

Les résultats énoncés dans les deux premiers paragraphes sont présentés de manière plus complète dans [Coh03], [Mal98] et dans [Mas99].

1.2.1 Approximation multi-résolution

Le concept d'échelle a été introduit par J.Morlet avec la notion d'analyse par ondelettes. Par la suite, Meyer [Mey90a], Mallat [Mal98] et Daubechies [Dau92] ont présenté la notion d'Analyse Multi-Résolution (AMR) qui a permis le développement de la théorie des bases d'ondelettes. Dans ce paragraphe, nous présentons le concept d'AMR sur $\mathbb{R}$. Le lecteur trouvera dans [CDD96, CM98b] une adaptation au cas $L^2([0,1])$. Les changements sont techniques et concernent essentiellement les fonctions de base aux bords : ils n'apportent pas d'éléments indispensables à l'exposé.

1.2.1.1 Analyse multi-échelles

Toutes les méthodes multi-échelles reposent fondamentalement sur la notion d'espaces d'approximation emboîtés.

$$V_\ell \subset V_{\ell+1},\ \ell \in \mathbb{N}. \tag{1.34}$$

Une approximation fine $v_{\ell+1}$ appartenant à $V_{\ell+1}$ est décomposée en un terme d'approximation grossière v_ℓ appartenant à V_ℓ et en un terme de détail w_ℓ appartenant à W_ℓ, un espace supplémentaire de V_ℓ dans $V_{\ell+1}$:

$$V_{\ell+1} = V_\ell \oplus W_\ell, \tag{1.35}$$

$$v_{\ell+1} = v_\ell + w_\ell. \tag{1.36}$$

L'itération de la décomposition à deux échelles conduit à une décomposition multi-échelles

$$v = v_0 + \sum_{\ell \in n} w_\ell,$$

où les composantes w_ℓ apparaissent comme des corrections successives de l'approximation initiale.

1.2.1.2 Définition d'une AMR de $L^2(\mathbb{R})$

Les propriétés d'une AMR reposent sur la notion de base de Riesz.

Définition 1.6 (Base de Riesz) *Soit H un espace de Hilbert. Une famille de vecteurs $\{e_n\}_{n\in\mathbb{N}}$ est appelée base de Riesz de H, si elle est linéairement indépendante et si il existe deux constantes positives A et B telles que, pour tout $f \in H$, il est possible de trouver une suite de réels $(f_n)_{n\in\mathbb{N}}$ avec*

$$\lim_{n\to\infty} \left\| f - \sum_{n=0}^{N} f_n e_n \right\|_H = 0 \tag{1.37}$$

et

$$\frac{1}{B}\|f\|_H^2 \leq \sum_{n\in\mathbb{N}} |f_n|^2 \leq \frac{1}{A}\|f\|_H^2 \tag{1.38}$$

Proposition 1.7 ([Mal98]) *Soit $\{e_n\}_{n\in\mathbb{N}}$ une base de Riesz de H, alors il existe une base de Riesz $\{\widetilde{e}_n\}_{n\in\mathbb{N}}$ de H telle que, pour tout f appartenant à H,*

$$f = \sum_{n\in\mathbb{N}} \langle f, \widetilde{e}_n\rangle e_n = \sum_{n\in\mathbb{N}} \langle f, e_n\rangle \widetilde{e}_n. \tag{1.39}$$

De plus, nous obtenons une relation de biorthogonalité entre les bases :

$$\langle e_p, \widetilde{e}_n\rangle = \delta_p^n. \tag{1.40}$$

Preuve Le théorème de représentation de Riesz permet de montrer qu'il existe une famille $\{\widetilde{e}_n\}_{n\in\mathbb{N}} \in H$ telle que, pour tout f appartenant à H, $f_n = \langle f, \widetilde{e}_n\rangle$. De plus,

$$\frac{1}{B}\|f\|_H^2 \leq \sum_{n\in\mathbb{N}} |\langle f, \widetilde{e}_n\rangle|^2 \leq \frac{1}{A}\|f\|_H^2. \tag{1.41}$$

Nous en déduisons (1.39) et

$$A\|f\|_H^2 \leq \sum_{n\in\mathbb{N}} |\langle f, e_n\rangle|^2 \leq B\|f\|_H^2. \tag{1.42}$$

Ceci permet de montrer, d'une part, que la famille duale $\{\widetilde{e}_n\}_{n\in\mathbb{N}}$ est linéairement indépendante et, d'autre part, que la base $\{\widetilde{e}_n\}_{n\in\mathbb{N}}$ est également une base de Riesz pour l'espace H. ■

Définition 1.7 (Analyse multi-résolution) *Une analyse multi-résolution (AMR) est une séquence de sous-espaces fermés de l'espace $L^2(\mathbb{R})$, vérifiant les propriétés suivantes :*

(i) Les sous-espaces sont emboîtés :

$$\forall \ell \in \mathbb{Z}, \quad V_\ell \subset V_{\ell+1} \subset L^2(\mathbb{R}). \tag{1.43}$$

(ii) Leur union est dense dans $L^2(\mathbb{R})$:

$$\lim_{\ell\to\infty} V_\ell = \left(\overline{\bigcup_{\ell\in\mathbb{N}} V_\ell}^{L^2}\right) = L^2(\mathbb{R}), \tag{1.44}$$

(iii) L'intersection des espaces est réduite à la fonction nulle.

$$\lim_{\ell\to-\infty} V_\ell = \{0\}, \tag{1.45}$$

(iv) Les espaces sont reliés entre eux par une relation qui traduit une invariance par dilatation :

$$u \in V_\ell \;\Leftrightarrow\; u(2\cdot) \in V_{\ell+1}. \tag{1.46}$$

(v) Il existe une fonction $\varphi \in V_0$ appelée fonction d'échelle *telle que la famille $\{\varphi(\cdot - k),\ k \in \mathbb{Z}\}$ forme une base de Riesz de V_0.*

La relation d'invariance par dilatation (iv) et la propriété (v) permettent de montrer que la famille

$$\left\{\varphi_{\ell,\imath} = 2^{\ell/2}\varphi\left(2^\ell \cdot - \imath\right)\right\}, \tag{1.47}$$

est une base de Riesz de V_ℓ.

La représentation de φ dans l'espace V_1 implique la **relation d'échelle** :

$$\varphi(x) = \sqrt{2}\sum_n h(n)\varphi(2x - n), \tag{1.48}$$

avec $(h(n))_{n\in\mathbb{Z}} \in \ell^2(\mathbb{Z})$. Les $(h(n))_{n\in\mathbb{Z}}$ forment le *filtre* de φ. Ce filtre caractérise la fonction d'échelle φ.

La proposition 1.7 permet d'extraire une base de Riesz de générateur $\widetilde{\varphi}$ biorthogonale à la base générée par φ. La fonction d'échelle *primale* φ et la fonction d'échelle *duale* $\widetilde{\varphi}$ vérifient la relation de biorthogonalité :

$$\langle \varphi(\cdot - \imath), \widetilde{\varphi}(\cdot - j)\rangle = \delta_{i,j}. \tag{1.49}$$

Plusieurs propriétés de cette fonction *d'échelle duale* s'en déduisent. La fonction $\widetilde{\varphi}$ vérifie également une équation d'échelle de la forme (1.48), le *filtre dual* est noté $\widetilde{h}$. La fonction *d'échelle duale* n'est pas unique et, en général, n'appartient pas à V_0. Nous supposons que la fonction *d'échelle duale* $\widetilde{\varphi}$ engendre une *AMR duale* $\left(\widetilde{V}_\ell\right)_{\ell\in\mathbb{Z}} \in L^2(\mathbb{R})$. Les

AMR $(V_\ell)_{\ell\in\mathbb{Z}}$ et $\left(\widetilde{V}_\ell\right)_{\ell\in\mathbb{Z}}$ sont, par définition, des AMR biorthogonales si les fonctions d'échelles de ces AMR le sont.

Les fonctions de base de l'espace $\widetilde{V}_\ell$ vérifient (1.47). Nous pouvons définir l'opérateur de projection biorthogonale P_ℓ de $L^2(\mathbb{R})$ sur V_ℓ par :

$$\forall u \in L^2(\mathbb{R}), \quad P_\ell(u) = \sum_{i\in\mathbb{Z}} \langle u, \widetilde{\varphi}_{\ell,\imath}\rangle \, \varphi_{\ell,\imath}, \tag{1.50}$$

et, de façon symétrique, l'opérateur de projection biorthogonale $\widetilde{P}_\ell$ sur $\widetilde{V}_\ell$.

Dans le cas de fonctions *d'échelles primale* et *duale* à support compact, Cohen [Coh03] met en évidence une condition nécessaire et suffisante pour que la projection de u converge en norme L^2 vers la fonction u :

$$\forall u \in L^2(\mathbb{R}), \quad \lim_{\ell\to\infty} \|u - P_\ell(u)\|_{L^2(\mathbb{R})} = 0, \tag{1.51}$$

si et seulement si

$$\left(\int_{\mathbb{R}} \widetilde{\varphi}(w)\, dw\right) \sum_{\imath\in\mathbb{Z}} \varphi(x-\imath) = 1, \tag{1.52}$$

presque partout sur $\mathbb{R}$.

Afin de vérifier la condition (1.52), nous choisissons de normaliser les fonctions d'échelles pour qu'elles soient de moyenne 1 ; ceci induit la normalisation des filtres suivante :

$$\sum_{n\in\mathbb{Z}} h(n) = \sum_{n\in\mathbb{Z}} \widetilde{h}(n) = \sqrt{2}. \tag{1.53}$$

1.2.1.3 Construction des bases d'ondelettes

Appliquons la méthode de décomposition en échelles du § 1.2.1.1 à l'AMR associée à la fonction φ.

Proposition 1.8 *Soit la famille de fonctions définie par*

$$\left\{\psi_{\ell,\imath} = 2^{\ell/2}\psi\left(2^\ell \cdot -\imath\right), \quad \imath\in\mathbb{Z}\right\}, \tag{1.54}$$

où ψ est la fonction d'ondelette primale *et $\widetilde{\psi}$ la fonction* d'ondelette duale*. Nous supposons que ces fonctions vérifient les* équations d'ondelettes primale et duale *:*

$$\psi = \sqrt{2}\sum_n g(n)\varphi(2\cdot -n) \quad \textit{et} \quad \widetilde{\psi} = \sqrt{2}\sum_n \widetilde{g}(n)\widetilde{\varphi}(2\cdot -n), \tag{1.55}$$

où le filtre d'ondelette primal g *(*resp.dual $\widetilde{g}$*) est donné par*

$$g(n) = (-1)^{1-n}\,\widetilde{h}(1-n) \quad \text{resp.} \quad \widetilde{g}(n) = (-1)^{1-n}\,h(1-n). \tag{1.56}$$

Alors cette famille forme une base de l'espace de détails W_ℓ défini par (1.35).

De plus, cette base de W_ℓ complète la base de Riesz de V_ℓ de sorte que l'union des deux bases soit une base de Riesz de $V_{\ell+1}$.

Les résultats suivants sont extraits de [CDF92]. Ils nous permettent de définir la notion de base d'ondelettes. Remarquons que la relation de biorthogonalité (1.49) équivaut à la propriété sur les *filtres d'échelles* :

$$\sum_{n\in\mathbb{Z}} h\left(2\imath+n\right)\widetilde{h}\left(n\right)=2\delta_{\imath}. \tag{1.57}$$

La relation (1.56) implique une relation similaire sur les *filtres d'ondelettes*

$$\sum_{n\in\mathbb{Z}} g\left(2\imath+n\right)\widetilde{g}\left(n\right)=2\delta_{\imath}. \tag{1.58}$$

Cette dernière relation implique la biorthogonalié des fonctions d'échelles ψ et $\widetilde{\psi}$. Ces relations sur les filtres permettent également de montrer que l'espace W_0 (*resp.* $\widetilde{W_0}$) est orthogonal à l'espace $\widetilde{V_0}$ (*resp.*V_0).

Ces propriétés impliquent le résultat suivant :

Définition 1.8 (Base d'ondelettes) *L'union des bases des sous-espaces W_ℓ :*

$$\bigcup_{\ell\in\mathbb{N}}\left\{\psi_{\ell,\imath}=2^{\ell/2}\psi\left(2^{\ell}\cdot-\imath\right),\quad \imath\in\mathbb{Z}\right\},$$

forme une base de Riesz de $L^2\left(\mathbb{R}\right)$, que nous appellerons base d'ondelettes de $L^2\left(\mathbb{R}\right)$.

De plus, pour toute fonction $u\in L^2\left(\mathbb{R}\right)$,

$$u=P_0(u)+\sum_{\ell=0}^{\infty}Q_\ell(u),$$

où $Q_\ell=P_{\ell+1}-P_\ell$ est la projection biorthogonale de $L^2\left(\mathbb{R}\right)$ sur l'espace de détails W_ℓ.

Les relations entre les espaces W_ℓ impliquent que tout espace d'échelle V_ℓ est une somme directe des espaces de détails :

$$V_\ell=V_0\oplus W_0\oplus\cdots\oplus W_{\ell-1}.$$

En reprenant les notations introduites au § 1.2.1.1, $w_\ell=Q_\ell(u)$ et nous obtenons un critère de stabilité qui s'avère être essentiel pour la construction d'un produit tensoriel sparse : $\forall v\in L^2\left(\mathbb{R}\right)$,

$$\|v\|^2_{L^2(\mathbb{R})}\approx\|P_0(v)\|^2_{L^2(\mathbb{R})}+\sum_{\ell=0}^{\infty}\|w_\ell\|^2_{L^2(\mathbb{R})}. \tag{1.59}$$

Intérêt d'une AMR biorthogonale Nous discutons le choix de la base d'ondelettes basée sur un couple d'AMR biorthogonales. En particulier, nous reviendrons sur la propriété remarquable de filtres à supports finis.

Supposons que les propriétés (iv) et (v) de la définition 1.7 soient remplacées par : la famille des $\left(\varphi_{\ell,\imath}\right)_{\imath\in\mathbb{Z}}$ forme une base de Riesz de V_ℓ avec

$$\varphi_{\ell,\imath}=\sum_{k\in\Gamma_{\ell+1}}h^{\ell}_{i,k}\,\varphi_{\ell+1,k},\quad \psi_{\ell,\imath}=\sum_{k\in\Gamma^{\star}_{\ell+1}}g^{\ell}_{i,k}\,\varphi_{\ell+1,k}. \tag{1.60}$$

La base $(\psi_{\ell,\imath})_{\ell\in\mathbb{Z},\imath\in\mathbb{Z}}$ ainsi construite est une base d'ondelettes. Le filtre d'échelle $h^\ell_{i,k}$, $(i,k) \in (\Gamma_\ell, \Gamma_{\ell+1})$ et le filtre d'ondelette $g^\ell_{i,k}$, $k \in \Gamma^\star_{\ell+1}$ sont supposés à supports finis (*i.e.*$\#\Gamma_\ell \in \mathbb{N}$). La projection d'une fonction u sur l'espace d'échelle V_ℓ et celle sur l'espace de détails W_ℓ sont données par

$$v_\ell = \sum_{i\in\mathbb{Z}} v^\imath_\ell \, \varphi_{\ell,\imath}, \quad \text{et} \quad w_\ell = \sum_{i\in\mathbb{Z}} w^\imath_\ell \, \psi_{\ell,\imath}.$$

Un algorithme récursif permet de reconstruire la fonction, c.-à-d. de calculer les coefficients $v^\imath_{\ell+1}$ en fonction de $v^\imath_\ell$ et $w^\imath_\ell$.

$$v^k_{\ell+1} = \sum_{i\in\Gamma_\ell} h^\ell_{i,k} \, v^\imath_\ell + \sum_{i\in\Gamma^\star_\ell} g^\ell_{i,k} \, w^\imath_\ell. \tag{1.61}$$

Cette relation admet une représentation matricielle. En notant (H_ℓ) et (G_ℓ) les matrices de taille $\#\Gamma_\ell \times \#\Gamma_{\ell+1}$ et $\#\Gamma^\star_\ell \times \#\Gamma^\star_{\ell+1}$ telles que $(H_\ell)_{i,k} = h^\ell_{i,k}$ et $(G_\ell)_{i,k} = g^\ell_{i,k}$, la relation de reconstruction (1.61) devient

$$\begin{pmatrix} \boldsymbol{v}_\ell \\ \boldsymbol{w}_\ell \end{pmatrix} = H^T_\ell \boldsymbol{v}_\ell + G^T_\ell \boldsymbol{w}_\ell,$$

où $M_\ell \overset{\text{def}}{=} \left(H^T_\ell, G^T_\ell\right)$. La matrice M_ℓ est creuse, si les blocs H_ℓ et G_ℓ le sont.

Inversement, nous disposons d'un algorithme de décomposition multi-échelle rapide si la matrice de décomposition M^{-1}_ℓ est creuse. Or l'inverse d'une matrice creuse est en général pleine. La difficulté de construction de « bonnes bases » d'ondelettes consiste à trouver des *filtres* tels que M_ℓ et M^{-1}_ℓ soient des matrices creuses. La matrice inverse admet elle aussi une décomposition par bloc : $M^{-1}_\ell = \begin{pmatrix} \widetilde{H}_\ell \\ \widetilde{G}_\ell \end{pmatrix}$, où les *filtres* $\widetilde{H}_\ell$ et $\widetilde{G}_\ell$ sont les *filtres duaux*. La relation de décomposition consiste donc à appliquer ces *filtres duaux* :

$$M^{-1}_\ell \boldsymbol{v}_{\ell+1} = \begin{pmatrix} \boldsymbol{v}_\ell \\ \boldsymbol{w}_\ell \end{pmatrix} \Rightarrow \boldsymbol{v}_\ell = \widetilde{H}_\ell \boldsymbol{v}_{\ell+1}, \quad \boldsymbol{w}_\ell = \widetilde{G}_\ell \boldsymbol{v}_{\ell+1}. \tag{1.62}$$

La difficulté se résume à trouver des *filtres* $\widetilde{H}_\ell$ et $\widetilde{G}_\ell$ creux. Pour une AMR donnée V_ℓ, $\ell \in \mathbb{N}$, il existe une infinité de choix d'espaces supplémentaires W_ℓ. Le choix le plus naturel est certainement le complémentaire orthogonal. Il conduit cependant à des *filtres* duaux pleins.

Dans le cas d'un couple d'AMR biorthogonales, les filtres ne dépendent pas du niveau ℓ. Ils sont donnés par les relations

$$h^\ell_{i,k} = h(k-2\imath), \quad g^\ell_{i,k} = g\,(k-2\imath) \quad \text{et} \quad \widetilde{h}^\ell_{i,k} = h(\imath-2k), \quad \widetilde{g}^\ell_{i,k} = g\,(\imath-2k)\,. \tag{1.63}$$

Si le couple d'AMR est choisi tel que les fonctions d'échelles sont à support compact, alors les filtres sont finis et les relations de décomposition et de reconstruction sont obtenues par des convolutions discrètes.

1.2.1.4 Propriétés d'une AMR

Propriétés des *filtres* : La proposition suivante s'avère très utile pour le calcul des matrices de rigidité d'une méthode de Galerkin sur une base d'ondelettes. Le résultat est une application d'un résultat plus général sur les ondelettes biorthogonales qui définissent une famille stable pour l'intégration et la dérivation.

Proposition 1.9 (Dérivation des ondelettes biorthogonales) *Le filtre sur la* relation d'échelle *se conserve après une dérivation*

$$\varphi'_{\ell,\imath}(x) = \sum_n h(n)\varphi'_{\ell+1,2\imath+n}(x)\,. \tag{1.64}$$

Ce résultat s'applique également à la relation sur les ondelettes

$$\psi'_{\ell,\imath}(x) = \sum_n g(n)\varphi'_{\ell+1,2\imath+n}(x)\,. \tag{1.65}$$

Preuve Remarquons, tout d'abord, que (1.47) et la *relation d'échelle* (1.48) impliquent

$$\varphi_{\ell,\imath}(x) = \sum_n h(n)\varphi_{\ell+1,2\imath+n}(x)\,. \tag{1.66}$$

La proposition résulte du calcul suivant : à partir de la relation sur la fonction mère, nous appliquons l'opération de dérivation,

$$\begin{aligned} \varphi'_{\ell,\imath}(x) &= 2^{3/2\ \ell}\varphi'\left(2^\ell x - \imath\right) \quad \text{avec} \quad \varphi'(x) = 2^{3/2\ \ell}\sum_n h(n)\varphi'(2x-n) \\ &= 2^{3/2(\ell+1)}\sum_n h(n)\varphi'\left(2^{\ell+1}x - (2\imath+n)\right) = \sum_n h(n)\varphi'_{\ell+1,2\imath+n}(x) \end{aligned} \tag{1.67}$$

■

Ordre d'une AMR

Définition 1.9 L'ordre p *d'une AMR est le plus grand entier positif tel que* $\mathbb{P}_{p-1}$ *soit inclus dans* V_0*, autrement dit*

$$\forall q = 0,\ldots,p-1, \quad x^q = \sum_{\imath\in\mathbb{Z}} \alpha_\imath^q \varphi(x-\imath)\,. \tag{1.68}$$

Notons $\left(\varphi^p, \widetilde{\varphi}^{p,\tilde{p}}\right)$ *le générateur du couple d'AMR biorthogonales tel que l'AMR primale est d'ordre* p *et l'AMR duale est d'ordre* $\tilde{p}$*.*

Si φ est la fonction d'échelle d'une AMR, à support compact et dans $H^m(\mathbb{R})$ alors l'ordre de l'AMR est supérieur au sens large à m. En effet, φ exprimée dans l'espace de Fourier est solution d'une équation fonctionnelle. Cette propriété implique que φ vérifie une condition de Strang-Fix équivalente à la propriété des moments (1.68), voir [Coh03].

L'orthogonalité entre les espaces $\widetilde{W}_0$ et V_0 et la définition 1.9 impliquent la proposition suivante :

Proposition 1.10 *Si l'AMR primale est d'ordre p (*resp. *duale est d'ordre $\widetilde{p}$) l'ondelette duale possède $p-1$ moments nuls :*

$$\forall q=0,\ldots,p-1,\quad \int_{\mathbb{R}} x^q\widetilde{\psi}(x)dx=0,\quad \text{resp.}\quad \forall q=0,\ldots,\widetilde{p}-1,\quad \int_{\mathbb{R}} x^q\psi(x)dx=0. \tag{1.69}$$

Cette proposition caractérise la propriété attendue de la représentation sur une base d'ondelettes : si la fonction u est régulière, le produit scalaire de u avec l'ondelette duale est petit. Cette intuition est précisée par les deux lemmes suivants.

Lemme 1.11 (Estimation directe) *Si l'AMR primale est d'ordre p et si les fonctions d'échelles primale et duale sont dans $L^2(\mathbb{R})$ alors*

$$\|u-P_\ell(u)\|_{L^2(\mathbb{R})}\leq C2^{-\ell p}\,|u|_{H^p(\mathbb{R})}\,,\quad \forall u\in H^p(\mathbb{R})\,.$$

Lemme 1.12 (Estimation inverse) *Sous les hypothèses précédentes, si φ la fonction d'échelle primale est dans $H^r(\mathbb{R})\,, r\leq p$ alors, pour toute fonction $v\in V_\ell$,*

$$\|v\|_{H^s(\mathbb{R})}\leq C2^{\ell s}\|v\|_{L^2(\mathbb{R})},\quad \forall s=0,\ldots,r.$$

Ces deux estimations permettent de montrer l'équivalence de norme suivante :

Proposition 1.13 *Si l'AMR primale est d'ordre p et si la fonction d'échelle primale φ (*resp.*duale $\widetilde{\varphi}$) appartient à $H^r(\mathbb{R})\,,\ r\leq p$ (*resp.*$L^2(\mathbb{R})$) alors, pour toute fonction u dans $H^s(\mathbb{R})\,,\ s\leq r\leq p$,*

$$\forall\ 0\leq s\leq r\leq p,\quad \|u\|^2_{H^s(\mathbb{R})}\approx\|P_0(u)\|^2_{L^2(\mathbb{R})}+\sum_{\ell\geq 1}\sum_{\imath\in\mathbb{Z}}2^{2\ell s}\left\langle u,\widetilde{\psi}_{\ell,\imath}\right\rangle^2. \tag{1.70}$$

Cette équation peut s'écrire

$$\forall\ 0\leq s\leq r\leq p,\quad \|u\|^2_{H^s(\mathbb{R})}\approx\|P_0(u)\|^2_{L^2(\mathbb{R})}+\sum_{\ell\geq 1}2^{2\ell s}\|w_\ell\|^2_{L^2(\mathbb{R})}.$$

La régularité des fonctions d'échelles s'effectue à partir d'une analyse dans l'espace de Fourier. Les résultats établis dans [Dau92] permettent de s'assurer qu'une famille d'ondelettes vérifie les hypothèses de la proposition 1.13.

Remarque 1.2 *Dans le cas d'un domaine Ω borné, la proposition 1.13 reste vrai. Ce résultat est une conséquence du théorème* 33.4 *page 630 dans [Coh00], en prenant les paramètres $(p=q=2, t=s, s=r, n=p)$.*

1.2.1.5 Compléments sur les ondelettes

Résumé des propriétés des ondelettes utiles pour la résolution numérique d'équations aux dérivées partielles. Pour conclure, nous donnons quelques propriétés de l'analyse multi-résolution utiles pour la résolution d'équations aux dérivées partielles. Sur une AMR, nous disposons :

(i) de critères de compression pour la représentation matricielle en base d'ondelettes d'opérateurs dont le noyau intégral possède certaines propriétés de décroissance. Il s'agit plus précisément des opérateurs de type *Calderon-Zygmund* dont la définition est rappelée dans [Mey90b]. En finance, ces résultats sont appliqués au cas des processus à saut dans les travaux de Schwab[MPS04],

(ii) d'un préconditionnement diagonal efficace dans le cas d'opérateurs linéaires,

(iii) de méthodes d'approximation adaptatives.

Approximation non linéaire Les propriétés des ondelettes permettent d'affiner les résultats précédents en considérant les espaces de Besov (1.12). Les résultats rappelés ici n'ont pas été mis en oeuvre. Ils semblent cependant constituer une approche complémentaire à celle du produit tensoriel sparse, en proposant une stratégie adaptative. Cette méthode réduit la complexité du problème en tenant compte de la régularité locale de la solution. Les travaux de Schwab [SS08] reprennent cette approche.

Afin d'expliquer la notion d'approximation non linéaire, nous considérons un espace vectoriel X muni d'une base $\Psi = \{\psi_\lambda\}_{\lambda\in\Lambda}$.

La méthode d'*approximation non linéaire à N termes* (« N-term approximation ») consiste, par exemple, à approcher v dans l'ensemble V de toutes les combinaisons linéaires de fonctions de Ψ à au plus N termes non nuls. En notant

$$S(\Lambda_N) = \left\{ v \in X \;\middle|\; v = \sum_{\lambda_i\in\Lambda_N} \alpha_i \psi_{\lambda_i} \right\},$$

nous obtenons

$$V = \bigcup_{\#\Lambda_N \leq N} S(\Lambda_N).$$

Nous admettrons pour l'instant l'existence d'un couple d'AMR biorthogonales définies sur $\mathbb{R}^d$. Sous certaines hypothèses sur l'ordre des AMR et sur la régularité des fonctions d'échelles, nous pouvons montrer que

$$\|u\|_{B^s_{p,q}(\Omega)} \approx \left(\sum_{\ell\geq 0} 2^{\ell q(s+d(1/2-1/p))} \left(\sum_{\imath\in\mathbb{Z}} \left| \left\langle u, \widetilde{\psi}_{\ell,\imath} \right\rangle \right|^p \right)^{q/p} \right)^{1/q}, \tag{1.71}$$

où l'espace de Besov $B^s_{p,q}(\Omega)$ est défini par (1.12). Ce résultat permet de démontrer une estimation directe de la forme du lemme 1.11 dans le cas d'une méthode d'approximation non linéaire :

soit $P_{\Lambda_N}(u)$ la meilleure approximation à N termes de u. Si $u \in B^{sd+t}_{\tau,\tau}$ avec $\tau = \left(s + \frac{1}{2}\right)^{-1}$, alors

$$\|u - P_{\Lambda_N}(u)\|_{H^t(\Omega)} \leq C N^{-s}.$$

Il faut noter qu'une fonction de $B^{sd+t}_{\tau,\tau}$ n'est pas forcément uniformément régulière.

1.2.2 Quelques familles d'ondelettes

1.2.2.1 Ondelettes B-spline biorthogonales

Le lecteur trouvera dans [CDF92] une description précise de l'ensemble des propriétés de ces ondelettes. Les principales définitions sont données ci-dessous.

Définition 1.10 (B−spline) *Soit $\chi_{[0,1]}$ la fonction caractéristique de $[0,1]$. Les B-splines d'ordre p (ou de degré $(p-1)$) sont obtenues par p convolutions successives de la fonction $\chi_{[0,1]}$:*

$$\varphi^{(p)}(\omega) = \chi^{\star(p)}_{[0,1]} \left(x + \left\lceil \frac{p}{2} \right\rceil \right). \tag{1.72}$$

La transformée de Fourier de $\varphi^{(p)}$ est donnée par

$$\widehat{\varphi}^{(p)}(\omega) = \left(\frac{\sin(\omega/2)}{\omega/2}\right)^p \exp\left(\frac{-i\epsilon\omega}{2}\right), \tag{1.73}$$

où, si p est impair, $\epsilon = 1$ et φ a son support centré en $x = 1/2$, si p est pair, $\epsilon = 0$ et φ est symétrique par rapport à $x = 0$.

Proposition 1.14 (Base de splines) *La base des splines polynomiales d'ordre p est engendrée par les translatées entières des B-splines du même ordre. Ces fonctions sont globalement $(p-2)$ fois continûment dérivables et coïncident avec un polynôme de degré $(p-1)$ sur les intervalles $[i2^\ell, (\imath+1)2^\ell]$, pour $i \in \mathbb{Z}$.*

Elements de preuve La réplication des polynômes se déduit d'une condition de Strang Fix [CDF92] vérifiée par la fonction d'échelle définie par (1.72). La régularité se déduit de l'analyse dans l'espace de Fourier de cette même fonction d'échelle. ■

D'après son caractère polynomial et son support compact, $\varphi^{(p)}$ constitue un choix privilégié de fonction d'échelle pour une AMR $(V_\ell)_{\ell\in\mathbb{Z}}$.

La construction de la fonction d'échelle duale et des bases d'ondelettes primales et duales associées à l'AMR se déduisent de la représentation dans l'espace de Fourier des relations d'échelles (1.48) et de (1.72). Le lecteur trouvera dans [Mal98] les formules des *filtres* et des ondelettes primale et duale ψ et $\widetilde{\psi}$.

Définition 1.11 *Une AMR B-spline biorthogonale est une AMR biorthogonale engendrée par le couple de fonctions d'échelle $\left(\varphi^p, \widetilde{\varphi}^{p,\tilde{p}}\right)$ (voir la définition 1.9) où la fonction φ^p est une fonction B-spline d'ordre p.*

Exemple 1.2 *Si $p = 1$, les fonctions d'échelle de l'AMR sont constantes par morceaux, les ondelettes obtenues sont les ondelettes de Haar.*

Exemple 1.3 *Si $p = 2$, $\varphi^{(2)}$ est la fonction chapeau ; la base des fonctions d'échelle est la base nodale des techniques d'éléments finis P_1.*

Le lecteur trouvera dans [CDD96] les propriétés d'une famille d'ondelettes biorthogonales pour les générateurs biorthogonaux $\left(\varphi^p, \widetilde{\varphi}^{p,\widetilde{p}}\right)$ tels que $p + \widetilde{p}$ soit pair. Les auteurs démontrent en particulier la condition vérifiée par $\widetilde{p}$, à p fixé, pour que $\widetilde{\varphi}^{p,\widetilde{p}} \in L^2(\mathbb{R})$. Des résultats plus fins sur la régularité de $\widetilde{\varphi}^{p,\widetilde{p}}$ sont également présentés. Remarquons que, pour $p = 2$, la plus petite valeur de $\widetilde{p}$ pour laquelle $\widetilde{\varphi}^{p,\widetilde{p}}$ appartient à $L^2(\mathbb{R})$ correspond à $\widetilde{p} = 2$. Nous utiliserons ces ondelettes dans des méthodes de Galerkin Sparse. Une brève description en est donnée ci-dessous.

Remarque 1.3 *Le choix de la base d'ondelettes pour les méthodes de Galerkin répond à deux critères :*

- *vérifier les hypothèses de la proposition 1.13 implique de prendre $\tilde{p} \geq 2$;*
- *limiter le support des ondelettes pour réduire le nombre de coefficients à priori non nuls de la matrice de rigidité.*

Exemple 1.4 (Ondelettes générées par la fonction d'échelle $\widetilde{\varphi}^{2,2}$) *La fonction d'échelle φ^2 est la fonction chapeau. Les filtres d'échelles primal h et dual $\tilde{h}$ sont donnés par*

$$h\left(-1,0,1\right)=\left(\frac{1}{2},1,\frac{1}{2}\right),\ \widetilde{h}\left(-2,-1,0,1,2\right)=\left(\frac{-1}{4},\frac{1}{2},1,\frac{1}{2},\frac{-1}{4}\right).$$

L'ondelette $\psi^{(2,2)}$ vérifie l'équation d'ondelette

$$\psi\left(x\right)=\sqrt{2}\left[\frac{-1}{4}\left(\varphi\left(2x-1\right)+\varphi\left(2x+3\right)\right)+\frac{-1}{2}\varphi\left(2x\right)+\varphi\left(2x+2\right)+\varphi\left(2x+1\right)\right].$$

Cette fonction est à support compact inclus dans $[\frac{-1}{2},\frac{3}{2}]$. Elle est représentée figure 1.1

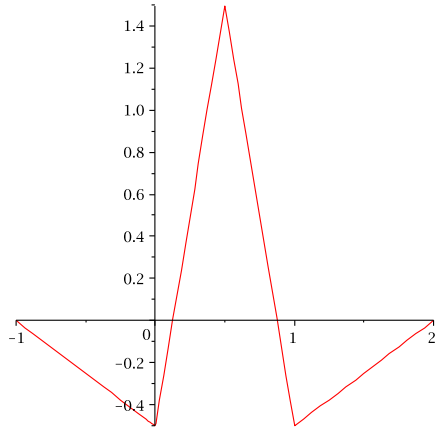

FIG. 1.1 – Ondelette $\psi^{2,2}$

1.2.2.2 Base d'ondelettes interpolantes ou base hiérarchique

Les ondelettes présentées ici nous permettront d'interpréter la méthode de différences finies sur une grille sparse comme une méthode de collocation. La classe d'analyse multi-résolution dite d'interpolation est présentée ci-dessous.

Le terme d'ondelette d'interpolation peut prêter à confusion. En effet, ce n'est pas à proprement parler l'ondelette qui est une fonction d'interpolation, mais la fonction d'échelle. Celle-ci vérifie la propriété d'interpolation

$$\phi\in L^2\left(\mathbb{R}\right)\cap\mathcal{C}^r\left(\mathbb{R}\right),\quad\text{avec}\quad\phi(k)=\delta_{k,0}. \tag{1.74}$$

A partir de cette fonction, l'approche proposée par D L. Donoho [Don92] définit la fonction d'ondelette par :

$$\psi\left(x\right)=\phi\left(2x-1\right). \tag{1.75}$$

Nous verrons que la fonction d'échelle duale est une masse de Dirac.

Revenons sur la construction de la fonction d'échelle d'interpolation. S.Bertoluzza [BN96] démontre que, pour une fonction d'échelle quelconque φ à support compact

générant une AMR $(V_\ell)_{\ell\in\mathbb{Z}}$, la fonction définie par

$$\widehat{\phi}\,(\omega) = \frac{\widehat{\varphi}\,(\omega)}{\sum_{\imath\in\mathbb{Z}} \widehat{\varphi}\,(\omega + 2\pi\imath)},$$

est une fonction d'échelle interpolante qui génère la même AMR.

La propriété essentielle de cette AMR interpolante se définit comme suit : l'opérateur de projection sur la base des fonctions d'échelle est un opérateur d'interpolation. La proposition suivante précise les propriétés d'approximation de cet opérateur.

Proposition 1.15 ([BN96]) *Soient l'AMR $(V_\ell)_{\ell\in\mathbb{Z}}$ d'ordre p et ϕ la fonction d'échelle interpolante telles que $\phi \in \mathcal{C}^r$, $r \leq p$. Soit $I_\ell : H^1(\mathbb{R}) \to V_\ell$ l'opérateur d'interpolation défini par*

$$I_\ell(u) = \sum_{\imath\in\mathbb{Z}} u\left(2^{-\ell}\imath\right)\phi\left(2^\ell x - \imath\right).$$

Alors, pour toute fonction $u \in H^t(\mathbb{R})$, $1 \leq t \leq p$,

$$\forall 0 \leq s \leq r, \quad \|u - I_\ell(u)\|_{H^s(\mathbb{R})} \leq 2^{-\ell(t-s)}\|u\|_{H^t(\mathbb{R})}.$$

Remarque 1.4 *Dans ce paragraphe, $\phi_{\ell,\imath}(x) = \phi\left(2^\ell x - \imath\right)$. Cette fonction n'est donc plus normalisée en norme $L^2(\mathbb{R})$ mais en norme $L^\infty(\mathbb{R})$.*

La caractérisation (1.75) implique le filtre d'échelle dual associé à l'AMR générée par ϕ. La fonction d'échelle duale vérifie l'équation

$$\widetilde{\phi}(x) = 2\widetilde{\phi}(2x).$$

Cette équation est satisfaite au sens des distributions par la masse de Dirac. Une AMR interpolante vérifie $\widetilde{\phi} = \delta$ et les ondelettes duales sont des combinaisons linéaires de masses de Dirac.

Considérons l'exemple de la base hiérarchique, c.-à-d. la base d'ondelettes engendrée par le couple de générateurs $\left(\varphi^p, \widetilde{\varphi}^{p,0}\right)$. Cette base est une base d'ondelettes d'interpolation. En effet, dans le cas limite de l'AMR biorthogonale (voir [Mas99]) défini par $\widetilde{p} = 0$, la fonction d'échelle duale $\widetilde{\varphi}^{p,0}$ est une distribution de Dirac en 0, correspond à une base d'ondelettes interpolantes.

Pour $p = 2$, nous obtenons les ondelettes de Donoho, la fonction d'échelle est la fonction chapeau. Pour $p = 4$, nous obtenons les ondelettes de Deslauriers et Dubuc [DD89]. Les résultats présentés ci-dessous dans le cadre des ondelettes de Donoho restent en grande partie valables pour des ondelettes interpolantes d'ordre p supérieur [BG04].

La *base nodale* correspond à la base des fonctions d'échelles d'une AMR interpolante générée par la fonction chapeau. La base hiérarchique correspond à la base d'ondelettes associée. Elles sont toutes les deux générées par la translation et la dilatation de la même fonction d'échelle, la fonction chapeau :

$$\phi(x) = \begin{cases} 1 - |x| & \text{si } x \in [-1, 1], \\ 0 & \text{sinon.} \end{cases}$$

Remarque 1.5 *D'après (1.75), nous pouvons exprimer les fonctions de la base d'ondelettes à partir des fonctions d'échelles*

$$\psi_{\ell,\imath}(x) = \psi\left(2^{\ell}x - \imath\right) = \phi\left(2^{\ell+1}x - (2\imath+1)\right) = \phi_{\ell+1,2\imath+1}(x). \tag{1.76}$$

En conséquence, l'ensemble des fonctions $(\psi_{\ell,\imath})_{\ell=0\ldots n-1,i=0\ldots 2^{\ell}-1}$ *est égal à l'ensemble des fonctions d'échelles multi-niveaux* $(\phi_{\ell,\imath})_{\ell=1\ldots n,i=1\ldots 2^{\ell}-1,\, i\ impair}$.

Cette différence de notation ne sera pas toujours explicite dans la suite. La définition des ensembles d'indices sera sous jacente au contexte et aux notations ($\psi_{\ell,\imath}$ *ou* $\phi_{\ell,\imath}$).

Sur $[0,1]$, les techniques d'éléments finis classiques utilisent une grille avec un pas de discrétisation 2^n fixé. Les fonctions de base considérées appartiennent à :

$$\left\{\phi_{n,\imath} \middle| 0 \leq i \leq 2^n\right\}. \tag{1.77}$$

L'interpolation consiste à approcher u par

$$u(x) \approx \sum_{\imath=0}^{2^n} u_\imath \cdot \phi_{n,\imath}(x), \text{ avec } u_\imath = u(x_{n,\imath}) \text{ et } x_{n,\imath} = \imath \cdot 2^{-n}. \tag{1.78}$$

Cette représentation de u sera nommée représentation nodale.

Définition 1.12 (Base hiérarchique) *La base hiérarchique est caractérisée, d'après (1.76), par :*

$$\left\{\phi_{\ell,\imath} \;\middle|\; \ell = 1,\ldots,n \quad \imath \in \left\{0..2^{\ell}\right\}, \quad \forall \imath \ impair \;\right\}. \tag{1.79}$$

Elle correspond à la base d'ondelettes interpolantes

$$\left\{\psi_{\ell,\imath} \;\middle|\; \ell = 0,\ldots n-1, \quad \imath \in \left\{0..2^{\ell}-1\right\}\right\}.$$

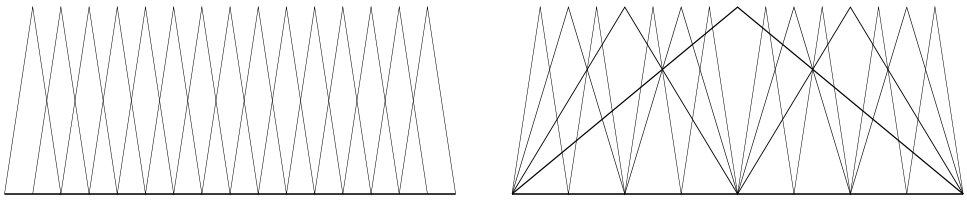

FIG. 1.2 – Fonctions de la base nodale (gauche) et de la base hiérarchique (droite)

Propriétés de ces deux bases Énumérons quelques propriétés qui seront utilisées dans les schémas de différences finies sur les *Sparse Grids*.

Les fonctions de la base nodale vérifient la **relation d'échelle,** illustrée par la figure 1.3,

$$\phi_{\ell,\imath} = \phi_{\ell+1,\imath} + \frac{1}{2}\left(\phi_{\ell+1,\imath-1} + \phi_{\ell+1,\imath+1}\right). \tag{1.80}$$

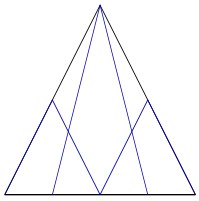

FIG. 1.3 – Relation d'échelle

Les différents filtres se déduisent de (1.80) et (1.75),

$$h(-1,0,1)=\left(\frac{1}{2},1,\frac{1}{2}\right),\quad g(0,1,2)=(0,1,0)\ \text{ et }\ \widetilde{g}(0,1,0)=\left(-\frac{1}{2},1,-\frac{1}{2}\right).$$

Le **filtre d'ondelette duale** permet de définir cette dernière :

$$\widetilde{\psi}(x)=-\frac{1}{2}\delta_0(x)+\delta_{\frac{1}{2}}(x)-\frac{1}{2}\delta_1(x). \tag{1.81}$$

Remarquons que $\widetilde{\psi}(x)=-\psi''(x)$ et que $\widetilde{\psi}_{\ell,\imath}(x)=2^{\ell}\widetilde{\psi}\left(2^{\ell}x-\imath\right)$.

Proposition 1.16 *Soient $u_{\ell,\imath}$, la valeur de la fonction u au point $x_{\ell,\imath}=2^{-\ell}i$ et $\widehat{u}_{\ell,\imath}$ les coefficients de la représentation d'une fonction u sur la base hiérarchique (1.79). Alors les coefficients $\widehat{u}_{\ell,\imath}$ sont liés aux coefficients $u_{\ell,\imath}$ par la* **relation de décomposition** *:*

$$\widehat{u}_{\ell,\imath}=u_{\ell,\imath}-\frac{1}{2}\left(u_{\ell-1,\frac{i-1}{2}}+u_{\ell-1,\frac{i+1}{2}}\right). \tag{1.82}$$

Preuve Rappelons que $\phi_{\ell+1,2\imath+1}(x)=\psi_{\ell,\imath}(x)$,

$$\begin{aligned}\widehat{u}_{\ell+1,2\imath+1}&=\int u(x)\widetilde{\psi}_{\ell,\imath}(x)\,dx=\int u(x)2^{\ell}\widetilde{\psi}\left(2^{\ell}x-\imath\right)dx=\int u\left((\theta+\imath)2^{-\ell}\right)\widetilde{\psi}(\theta)\,d\theta\\&=u\left((\imath+\frac{1}{2})2^{-\ell}\right)-\frac{1}{2}\left(u\left(\imath 2^{-\ell}\right)+u\left((\imath+1)2^{\ell}\right)\right)\\&=u_{\ell+1,2\imath+1}-\frac{1}{2}\left(u_{\ell,\imath}+u_{\ell,\imath+1}\right).\end{aligned} \tag{1.83}$$

■

Nous disposons également de la **relation de reconstruction** :

$$u_{\ell,\imath}=\widehat{u}_{\ell,\imath}+\frac{1}{2}\left(u_{\ell-1,\frac{i-1}{2}}+u_{\ell-1,\frac{i+1}{2}}\right), \tag{1.84}$$

Remarque 1.6 *La numérotation de la grille s'effectue sous la forme d'arbre binaire (voir § 9.1.2), la* **relation de décomposition** *se note :*

$$\widehat{u}=u-\frac{1}{2}\left(u_{P\grave{e}re\ Gauche}+u_{P\grave{e}re\ Droit}\right)=Hu. \tag{1.85}$$

Cette relation est mise en évidence sur la figure 1.4.

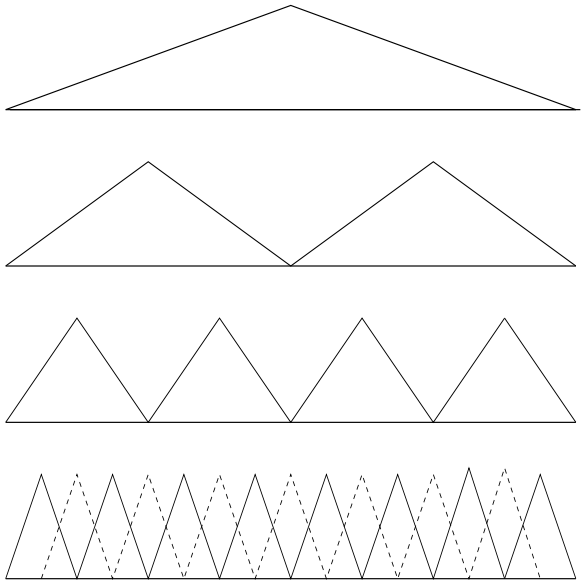

FIG. 1.4 – De la base hiérarchique à la base nodale

Si la fonction u est suffisamment régulière, le coefficient $\widehat{u}_{\ell,\imath}$ est obtenu par la formule intégrale donnée dans la proposition suivante.

Proposition 1.17 *Soient $\psi_{\ell,\imath}$ un élément de la base des ondelettes interpolantes et u appartenant à $H^2(\Omega)$ alors*

$$\widehat{u}_{\ell+1,2\imath+1} = -2^{-\ell} \int_\Omega \frac{\partial^2 u}{\partial x^2}\, \psi_{\ell,\imath}. \tag{1.86}$$

Preuve Cette proposition est démontrée dans [BG04] (p13 lemme 3.2). Nous proposons ici une démonstration basée sur la projection sur la base des $\psi_{\ell,\imath}$.

$$\begin{aligned} \widehat{u}_{\ell+1,2\imath+1} &= \left\langle u, \widetilde{\psi}_{\ell,\imath} \right\rangle = \int_\Omega u\left((\theta+i)2^{-\ell}\right) \widetilde{\psi}(\theta) d\theta = -\int_\Omega u\left((\theta+i)2^{-\ell}\right) \psi''(\theta) d\theta \\ &= -2^{\ell} \int_\Omega u(x)\, \psi''\left(2^\ell x - \imath\right) dx = -2^{-\ell} \int_\Omega u''(x)\, \psi_{\ell,\imath}(x)\, dx. \end{aligned} \tag{1.87}$$

■

Remarque 1.7 *Un résultat analogue (1.86) est obtenu en normalisant $\psi_{\ell,\imath}$ par rapport à la norme $\| \ \|_{L^2(\mathbb{R})}$. En notant $\overline{\psi}_{\ell,\imath}$ la fonction normalisée, $\overline{\psi}_{\ell,\imath} = 2^{\ell/2}\psi\left(2^\ell x - \imath\right) = 2^{\ell/2}\psi_{\ell,\imath}$ et $\overline{\widetilde{\psi}}_{\ell,\imath} = 2^{\ell/2}\widetilde{\psi}\left(2^\ell x - \imath\right) = 2^{-\ell/2}\widetilde{\psi}_{\ell,\imath}$, alors*

$$\left\langle u, \overline{\widetilde{\psi}}_{\ell,\imath} \right\rangle = -2^{-2\ell} \int_\Omega \frac{\partial^2 u}{\partial x^2}(x)\ \overline{\psi}_{\ell,\imath}(x) dx.$$

1.2.3 Analyse multi-résolution sur $L^2\left([0,1]^d\right)$

1.2.3.1 Produit tensoriel

L'approche proposée pour la construction de bases d'ondelettes sur $L^2(\mathbb{R})$ (resp $L^2([0,1])$) se généralise à l'espace $L^2\left(\mathbb{R}^d\right)$ (resp $L^2\left([0,1]^d\right)$ par produit tensoriel. Ce procédé est similaire à celui utilisé dans le cadre des méthodes spectrales [BMR04]. Une extension à des domaines plus généraux peut s'effectuer par des techniques de décomposition de domaine présentée dans [CY89] pour des méthodes d'approximation sur des bases de polynômes et [Mas99] pour des méthodes d'approximation sur des bases d'ondelettes.

Produit tensoriel d'espaces de dimensions finies Soit $\left\{V_{n_k}^k\right\}_{1\leq k\leq d}$ une famille d'espaces de fonctions d'une variable définies sur l'intervalle Ω_k. La dimension de $V_{n_k}^k$ est notée n_k. Chacun des espaces $V_{n_k}^k$ est muni d'une base $\left\{\varphi_i^k\right\}_{1\leq i\leq n_k}$.

Définition 1.13 *Soient $\boldsymbol{\Omega}$ l'hypercube $\Omega_1\times\cdots\times\Omega_d$, et $\boldsymbol{n}$ le multi-indice $(n_1,\ldots,n_d)$, l'espace $\boldsymbol{V_n}(\boldsymbol{\Omega})$ est obtenu par produit tensoriel de la famille $\left\{V_{n_k}^k\right\}_{1\leq k\leq d}$ et noté*

$$\boldsymbol{V_n} = V_{n_1}^1\otimes\cdots\otimes V_{n_d}^d = \bigotimes_{k=1}^{d} V_{n_k}, \tag{1.88}$$

si

$$\boldsymbol{V_n} = span\left\{\varphi_{\boldsymbol{\ell}}\right\}_{\boldsymbol{\ell}\leq\boldsymbol{n}}, \qquad avec \qquad \varphi_{\boldsymbol{\ell}}(\boldsymbol{x}) = \prod_{k=1}^{d}\varphi_{\ell_k}^k(x_k). \tag{1.89}$$

1.2.3.2 Base d'ondelettes sur $L^2\left([0,1]^d\right)$

Deux approches permettent de construire des bases d'ondelettes multi-dimensionnelles. La première consiste en une hiérarchisation isotrope [Dah97, Sch98b]. Les fonctions de base obtenues sont représentées sur la figure 1.5. La seconde approche consiste à utiliser un produit tensoriel anisotrope - figure 1.6.

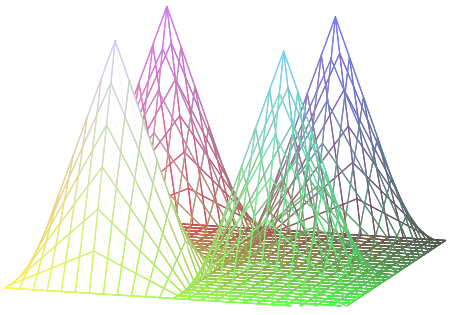

FIG. 1.5 – Fonctions de la base hiérarchique isotrope

Définition 1.14 (Tensorisation anisotrope) *Soit $\boldsymbol{\tau_\ell} = \tau_{\ell_1}\times\ldots\tau_{\ell_d}$, où τ_{ℓ_j} un ensemble d'indices défini pour chaque niveau ℓ. L'*espace de détail *multi-dimensionnel, noté $W_{\boldsymbol{\ell}}$,*

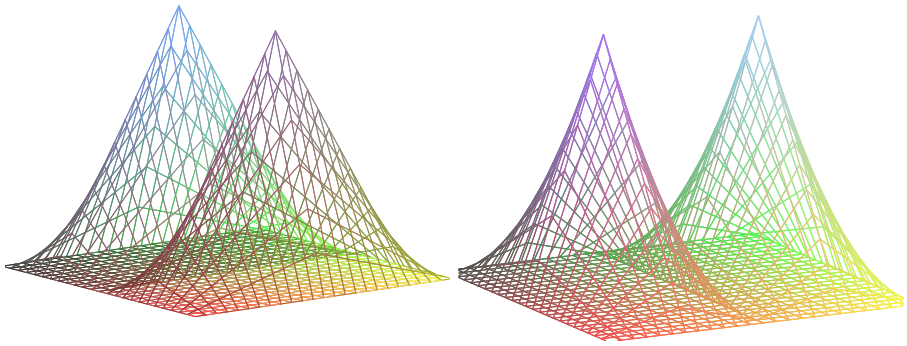

FIG. 1.6 – Fonctions de la base hiérarchique anisotrope

obtenu par tensorisation anisotrope est défini par

$$W_{\boldsymbol{\ell}} \overset{\text{def}}{=} span\left\{\psi_{\boldsymbol{\ell},\boldsymbol{\imath}} | \boldsymbol{\imath} \in \boldsymbol{\tau}_{\boldsymbol{\ell}}\right\} = \bigotimes_{k=1}^{d} W_{\ell_k}. \tag{1.90}$$

Comme dans le cas uni-dimensionnel, l'espace $\boldsymbol{V_n}$ est obtenu comme somme directe des sous-espaces $\boldsymbol{W_\ell}$:

$$\boldsymbol{V_n} = \bigoplus_{\boldsymbol{\ell}=(\mathbf{0},\cdots,\mathbf{0})}^{(\boldsymbol{n}_1,\cdots,\boldsymbol{n}_d)} W_{\boldsymbol{\ell}}. \tag{1.91}$$

Définition 1.15 (Subdivisions dyadiques) *Soient un multi-indice $\boldsymbol{\ell} \in \mathbb{N}^d$ et Ω_ℓ^1 la grille de l'intervalle $(0,1)$ définie par $\Omega_\ell^1 = 2^{-\ell}\left\{1,\ldots,2^\ell - 1\right\}$.*

La grille cartésienne de $\Omega = (0,1)^d$ est définie comme le produit cartésien des grilles en dimension 1,

$$\Omega_{\boldsymbol{\ell}} = \prod_{k=1}^{d} \Omega_1^{\ell_k}. \tag{1.92}$$

Les pas de cette grille sont donnés par :

$$\boldsymbol{h_l} \overset{\text{def}}{=} (h_{l_1},\ldots,h_{l_d}) = \left(2^{-l_1},\ldots,2^{-l_d}\right) = 2^{-\boldsymbol{\ell}}. \tag{1.93}$$

Les points $\boldsymbol{x_{l,i}}$ de la grille $\boldsymbol{\Omega_\ell}$ sont

$$\mathbf{x}_{\ell,\mathbf{i}} \overset{\text{def}}{=} (x_{\imath_1,l_1},\ldots,x_{\imath_d,l_d}) = \left(\imath_1 \cdot 2^{-l_1},\ldots,\imath_d \cdot 2^{l_d}\right) = \boldsymbol{\imath} \cdot \boldsymbol{h_\ell} \quad avec\ \mathbf{0} \leq \imath \leq 2^\ell, \tag{1.94}$$

Le multi-indice $\boldsymbol{\ell}$ indique le niveau du point alors que le multi-indice $\boldsymbol{\imath}$ indique la position du point dans le niveau $\boldsymbol{\ell}$.

Définition 1.16 (Fonction de grille) *Une fonction de grille sur $\Omega_{\boldsymbol{\ell}}$ est une application de $\Omega_{\boldsymbol{\ell}}$ dans $\mathbb{R}$ qui associe à chaque point de la grille une valeur. L'espace des fonctions de grille sur $\Omega_{\boldsymbol{\ell}}$ forme une bijection avec $\prod_{k=1}^{d} \mathbb{R}^{2^{\ell_k}-1}$.*

Proposition 1.18 *L'application qui à* $(u_{\boldsymbol{\imath}})_{\mathbf{1}\leq\boldsymbol{\imath}\leq\mathbf{2^{\ell}-1}} \rightarrow u = \sum_{\mathbf{1}\leq\boldsymbol{\imath}\leq\mathbf{2^{\ell}-1}} u_{\boldsymbol{\imath}}\varphi_{\boldsymbol{\ell},\boldsymbol{\imath}}$ *est un isomorphisme de l'espace des fonctions de grilles sur* $\Omega_{\boldsymbol{\ell}}$ *dans l'espace* $V_{\boldsymbol{\ell}}$.

De plus, la fonction u *peut également s'écrire sur la base d'ondelettes,* $u = \sum_{\boldsymbol{m}\leq\boldsymbol{\ell},\boldsymbol{\imath}\in\tau_{\boldsymbol{m}}} u_{\boldsymbol{m},\boldsymbol{\imath}}\psi_{\boldsymbol{m},\boldsymbol{\imath}}$.

1.2.3.3 Base hiérarchique multi-dimensionnelle

L'espace $\boldsymbol{V_n}$ admet une base de fonctions obtenues par produit tensoriel des bases hiérarchiques des espaces de fonctions d'une variable. Soient $\{\varphi_{l_k,\imath_k}\}_{\imath_k\leq 2^{\ell_k},\,\ell_k\leq n_k}$ les bases de fonctions des espaces V_{n_k}, alors les fonctions

$$\varphi_{\boldsymbol{\ell},\boldsymbol{\imath}}(\boldsymbol{x}) = \prod_{k=1}^{d} \varphi_{l_k,\imath_k}(x_k), \tag{1.95}$$

forment une base de $\boldsymbol{V_n}$:

$$\mathbf{V}_{\boldsymbol{n}} = span\left\{\varphi_{\boldsymbol{\ell},\boldsymbol{\imath}} | \mathbf{0}\leq\boldsymbol{\imath}\leq 2^{\boldsymbol{\ell}},\ \boldsymbol{\ell}\leq\boldsymbol{n}\right\}. \tag{1.96}$$

Les relations de passage uni-dimensionnelles (1.84) et (1.85) deviennent, avec des notations évidentes :

– passage de la base nodale à la base hiérarchique

$$\widehat{u} = \left(H_{x_1}\circ\cdots\circ H_{x_d}\right)u, \tag{1.97}$$

– la relation inverse, passage de la base hiérarchique à la base nodale

$$u = \left(H_{x_1}^{-1}\circ\cdots\circ H_{x_d}^{-1}\right)\widehat{u}. \tag{1.98}$$

Pour la projection sur la base hiérarchique, la proposition 1.17 se généralise ainsi :

Proposition 1.19 *Si* u *appartient à* $X_0^{1,2}(\boldsymbol{\Omega})$*, alors le coefficient* $\widehat{u}_{\boldsymbol{\ell},\boldsymbol{i}}$ *de sa représentation sur la base hiérarchique sparse vérifie*

$$\widehat{u}_{\boldsymbol{\ell}+\mathbf{1},\mathbf{2\imath}+\mathbf{1}} = (-1)^d 2^{-|\boldsymbol{\ell}|_1}\int_{\Omega}\frac{\partial^{2d}u}{\partial x_1^2\ldots\partial x_d^2}\,\psi_{\boldsymbol{\ell},\boldsymbol{\imath}}(x)dx \quad \text{avec } \psi_{\boldsymbol{\ell},\boldsymbol{\imath}}(\boldsymbol{x}) = \prod_{k=1}^{d}\psi_{\ell_k,\imath_k}(x_k), \tag{1.99}$$

et les $\psi_{\ell_k,\imath_k}(x_k)$ *sont données à la définition 1.12.*

Remarque 1.8 *A nouveau nous obtenons, au sens des distributions,* $\widetilde{\psi}_{\boldsymbol{\ell},\boldsymbol{\imath}} = (-1)^d 2^{-2|\boldsymbol{\ell}|_1}\partial_{x_1}^2\ldots\partial_{x_d}^2\psi_{\boldsymbol{\ell},\boldsymbol{\imath}}$.

Remarque 1.9 *Pour la norme* $L^2(\Omega)$*,*

$$\left\langle u, \overline{\widetilde{\psi}}_{\boldsymbol{\ell},\boldsymbol{\imath}}\right\rangle = (-1)^d 2^{-2|\boldsymbol{\ell}|_1}\int_{\Omega}\frac{\partial^{2d}u}{\partial x_1^2\ldots\partial x_d^2}\,\overline{\psi}_{\boldsymbol{\ell},\boldsymbol{\imath}}(x)dx \quad \text{avec } \overline{\psi}_{\boldsymbol{\ell},\boldsymbol{\imath}}(\boldsymbol{x}) = \prod_{k=1}^{d}\overline{\psi}_{\ell_k,\imath_k}(x_k), \tag{1.100}$$

1.3 Produit tensoriel sparse

La construction de grille ou de sous-espace sparse est basée sur la notion de bases multi-échelles. Les sous-espaces sont « creusés » en supprimant certains niveaux, plus précisément, certains multi-indices $\boldsymbol{\ell}$ dans le produit tensoriel (1.91). Les niveaux supprimés contiennent une information « négligeable » lorsque la fonction à approcher vérifie une hypothèse de régularité. Cette construction des espaces sparse est décrite dans [BG04].

1.3.1 Espaces d'approximation sparse

Soit une AMR uni-dimensionnelle munie des espaces d'ondelettes primales W_ℓ, pour laquelle les ondelettes primales (*resp.* duales) admettent $\tilde{p}$ (*resp.* p) moments nuls. Soient un réel s tel que $\min(\tilde{p}, p) \leq s \leq \max(\tilde{p}, p)$ et une fonction $u \in H^s(\Omega)$, qui admet la décomposition $u = \sum_\ell w_\ell$, $w_\ell \in W_\ell$. Nous supposons l'existence de l'équivalence de norme suivante

$$\|u\|^2_{H^s(\Omega)} \approx \sum_\ell \|w_\ell\|^2_{H^s(\Omega)} \approx \sum_\ell 2^{2s\ell} \|w_\ell\|^2_{L^2(\Omega)}. \tag{1.101}$$

Remarque 1.10 *Cette équivalence de norme n'est pas toujours vérifiée. Il est nécessaire de se placer dans le cadre d'application de la proposition 1.13. Dans le cas $s = \max(\tilde{p}, p)$, nous disposons uniquement de la majoration suivante*

$$\|u\|^2_{H^s(\Omega)} \gtrsim 2^{2s\ell} \|w_\ell\|^2_{L^2(\Omega)}. \tag{1.102}$$

Cette hypothèse constitue l'élément principal de démonstration des propriétés des espaces sparse.

La définition de décomposition stable en somme de sous-espaces est extraite de [GO95, GK00], de même que les deux propriétés de ces espaces énoncées dans le paragraphe suivant. Celles-ci fournissent les arguments à la démonstration de la stabilité de la décomposition sur la base d'ondelettes.

1.3.1.1 Décomposition en somme de sous-espaces

Soit $\{V; a\}$ un espace de Hilbert muni du produit scalaire $a(\cdot, \cdot)$. Considérons un ensemble de sous-espaces fermés $V_\ell \subset V$ tel que la somme des V_ℓ soit un sous-espace dense de V. Notons a_ℓ le produit scalaire associé à V_ℓ.

Définition 1.17 *Une décomposition en somme de sous-espaces $\sum_\ell \{V_\ell; a_\ell\}$ est dite stable s'il existe une équivalence de norme entre $\sqrt{a(u,u)}$ et*

$$|||u||| = \left(\inf_{u_\ell \in V_\ell : u = \sum_\ell u_\ell} \left\{ \sum_\ell a_\ell(u_\ell, u_\ell) \right\} \right)^{\frac{1}{2}},$$

ce qui équivaut à l'existence de deux constantes positives λ_{min} et λ_{max} telles que :

$$\lambda_{min} = \inf_{0 \neq u \in V} \frac{a(u,u)}{|||u|||}, \quad \lambda_{max} = \sup_{0 \neq u \in V} \frac{a(u,u)}{|||u|||}. \tag{1.103}$$

la constante κ est le rapport $\frac{\lambda_{max}}{\lambda_{min}}$.

Notons $\boldsymbol{V} = \bigotimes_{j=1}^{d} V_j$, muni du produit scalaire a_j et $\boldsymbol{a}$ le produit scalaire sur $\boldsymbol{V}$ défini par $\boldsymbol{a} = a_1 \otimes \cdots \otimes a_d$.

Proposition 1.20 (Produit tensoriel de décomposition stable) *Si pour tout* $\left\{V^j; a^j\right\}$, $j \in \{1..d\}$, *il existe une décomposition stable en somme de sous-espaces :* $\left\{V^j; a^j\right\} = \sum_{\boldsymbol{\ell}} \left\{V_{\ell_j}; a_{\ell_j}\right\}$ *de constante* κ_j, *alors il existe pour l'espace obtenu par tensorisation des espaces* V^j, *une décomposition stable en somme de sous-espaces :*

$$\{\boldsymbol{V}; \boldsymbol{a}\} = \sum_{\boldsymbol{\ell}} \{V_{\ell_1} \otimes \ldots \mathrm{V}_{\ell_d}; a_{\ell_1} \otimes \cdots \otimes a_{\ell_d}\}, \tag{1.104}$$

avec une constante $\boldsymbol{\kappa} = \prod_j \kappa_j$.

Proposition 1.21 (Intersection de décomposition stable) *Soit une suite de réels positifs* $\{\alpha_{k,\boldsymbol{\ell}}\}_{\boldsymbol{\ell}}$, $k = 1, \ldots, .., n$; *nous supposons que, pour tout* k, *les espaces* $\{\boldsymbol{Z}_k; \boldsymbol{b}_k\}$ *sont munis d'une décomposition stable en somme de sous-espaces suivant la même suite d'espaces* $\boldsymbol{V}_{\boldsymbol{\ell}}$:

$$\{\boldsymbol{Z}_k; \boldsymbol{b}_k\} = \sum_{\boldsymbol{\ell}} \{\boldsymbol{V}_{\boldsymbol{\ell}}; \alpha_{k,\boldsymbol{\ell}} \boldsymbol{a}\}.$$

(La forme bilinéaire $\boldsymbol{a}$ *est la même pour chaque niveau* $\boldsymbol{\ell}$*). Alors, pour tout* $\gamma_k > 0$, $k = 1, \ldots, d$ *la décomposition en somme de sous-espaces*

$$\{\boldsymbol{Z}_1 \cap \cdots \cap \boldsymbol{Z}_n; \gamma_1 \boldsymbol{b}_1 + \cdots + \gamma_n \boldsymbol{b}_n\} = \sum_{\boldsymbol{\ell}} \{V_{\boldsymbol{\ell}}; (\gamma_1 \alpha_{1,\ell} + \cdots + \gamma_n \alpha_{n,\ell})\, \boldsymbol{a}\}, \tag{1.105}$$

est également stable et la constante correspondante vérifie $\boldsymbol{\kappa} \leq \frac{\max\left(\lambda^1_{max}, \ldots, \lambda^d_{max}\right)}{\min\left(\lambda^1_{min}, \ldots, \lambda^d_{min}\right)}$.

Le lecteur trouvera la démonstration de ces deux propriétés dans [GO95].

1.3.1.2 Décomposition sparse en base d'ondelettes

Les deux propositions précédentes permettent de démontrer le résultat suivant :

Théorème 1.22 (Stabilité) *Supposons vérifiée (1.101) pour les AMR primales et duales des fonctions à une variable. Si* u *appartient à* $H^s(\Omega)$, $u = \sum_{\boldsymbol{\ell}} w_{\boldsymbol{\ell}}$, $w_{\boldsymbol{\ell}} \in W_{\boldsymbol{\ell}}$ *alors*

$$\|u\|^2_{H^s(\Omega)} \approx \sum_{\boldsymbol{\ell}} 2^{2s|\boldsymbol{\ell}|_\infty} \|w_{\boldsymbol{\ell}}\|^2_{L^2(\Omega)} \quad \textit{pour } s \in (-\tilde{p}, p). \tag{1.106}$$

Si u *appartient à* $\mathcal{H}^{t,s}(\Omega)$, *alors :*

$$\|u\|^2_{\mathcal{H}^{t,s}(\Omega)} \approx \sum_{\boldsymbol{\ell}} 2^{2t|\boldsymbol{\ell}|_1 + 2s|\boldsymbol{\ell}|_\infty} \|w_{\boldsymbol{\ell}}\|^2_{L^2(\Omega)} \quad \textit{pour } t \geq 0,\ 0 \leq t + s \leq p. \tag{1.107}$$

Remarque 1.11 *En corollaire, le second résultat du théorème appliqué à l'espace*
– $\mathcal{H}^{0,s}(\Omega) = H^s(\Omega)$, *donne le résultat (1.106).*

– $\mathcal{H}^{s,0}(\Omega) = \mathcal{H}^s(\Omega)$, *donne l'équivalence de norme classiquement utilisée dans le cas des* Sparse Grids *:*

$$\|u\|^2_{\mathcal{H}^s(\Omega)} \approx \sum_{\boldsymbol{\ell}} 2^{2s|\boldsymbol{\ell}|_1} \|w_{\boldsymbol{\ell}}\|^2_{L^2(\Omega)}.$$

Preuve La suite des espaces d'ondelettes W_ℓ munie de la forme bilinéaire définie par $a(u,v) = 2^{s\ell}(u,v)$, $u \in W_\ell$, $v \in W_\ell$, est une décomposition stable en somme de sous-espaces de l'espace des fonctions à une variable $H^s(\Omega_1)$. En appliquant la proposition 1.20 à

$$\left\{\mathcal{H}^{(0,\dots,q,\dots,0)}(\Omega); (\cdot,\cdot)_{L^2} \otimes \cdots \otimes (\cdot,\cdot)_{L^2} \otimes a(\cdot,\cdot) \otimes (\cdot,\cdot)_{L^2} \otimes \cdots \otimes (\cdot,\cdot)_{L^2}\right\},$$

(l'exposant q et le produit scalaire a étant en i ème position), nous démontrons que cet espace admet la décomposition stable en somme de sous-espaces

$$\sum_{\boldsymbol{\ell}} \left\{W_{\ell_1} \otimes \cdots \otimes W_{\ell_d}; (\cdot,\cdot)_{L^2} \otimes \cdots \otimes (\cdot,\cdot)_{L^2} \otimes 2^{q\ell_i} (\cdot,\cdot)_{L^2} \otimes (\cdot,\cdot)_{L^2} \otimes \cdots \otimes (\cdot,\cdot)_{L^2}\right\}. \tag{1.108}$$

La caractérisation (1.23) de l'espace $\mathcal{H}^{t,s}(\Omega)$ et la proposition 1.21 permettent de généraliser l'équation (1.108) à $\mathcal{H}^{t,s}(\Omega)$. La stabilité implique l'équivalence entre la norme de l'espace et la norme de la décomposition (1.107). Dans le cas $s > 0$, l'équation (1.106) se démontre en appliquant la remarque 1.11. Le cas $s < 0$ se démontre en appliquant les théorèmes aux ondelettes duales puis en utilisant l'argument de dualité $(H^s)' = H^{-s}$. ■

Définition 1.18 *Soient $n \in \mathbb{N}$ et $\mathcal{I}_n^\tau$, $\tau \in [-1,\infty]$, l'ensemble des multi-indices $\boldsymbol{\ell} \in \mathbb{N}^d$ tel que $|\boldsymbol{\ell}|_1 + \tau|\boldsymbol{\ell}|_\infty < (1+\tau)n$, alors l'espace sparse V_n^τ est défini par*

$$V_n^\tau \stackrel{\text{def}}{=} \bigoplus_{\boldsymbol{\ell} \in \mathcal{I}_n^\tau} W_{\boldsymbol{\ell}}. \tag{1.109}$$

Remarque 1.12 *Ici, les conventions sont différentes pour la définition du multi-indice $\boldsymbol{\ell}$. L'espace est caractérisé avec les fonctions d'ondelettes $\psi_{\ell,i}$ dont le niveau varie entre $0 \leq \ell \leq n-1$. En changeant cette convention, c.-à-d. en décrivant l'espace avec la notation de la base hiérarchique $\phi_{\ell+1,2i+1} = \psi_{\ell,i}$, la définition de l'ensemble $\mathcal{I}_n^\tau$ dépend de $n+d-1+\tau n$, forme du résultat la plus souvent énoncée dans la littérature.*

Remarque 1.13 *V_n^∞ coïncide avec $\bigoplus_{|\boldsymbol{\ell}|_\infty < n} W_{\boldsymbol{\ell}}$, l'espace plein standard. $V_n^0 = \bigoplus_{|\boldsymbol{\ell}|_1 < n} W_{\boldsymbol{\ell}}$, l'espace sparse standard.*

Théorème 1.23 *Soient $s,t \in \mathbb{Z}$ telles que $-\tilde{p} \leq s \leq t \leq p$. En appliquant le théorème 1.22, il existe une constante $C > 0$ telle que pour toute fonction $u \in H^t$:*

$$\inf_{v \in V_n^\infty} \|u - v\|^2_{H^s(\Omega)} \leq C\ 2^{2(s-t)n} \|u\|^2_{H^t(\Omega)}, \tag{1.110}$$

Preuve Utiliser le théorème 1.22. ■

L'ordre d'approximation $(t-s)$ caractérise l'erreur de projection sur l'espace V_n^∞. Étudions à présent l'espace d'approximation V_n^0. La proposition suivante montre que l'ordre diminue significativement, il devient $(s-t)\frac{n}{d}$.

Proposition 1.24 *En reprenant les hypothèses du théorème 1.23, il existe une constante $C > 0$ telle que, pour toute fonction $u \in H^t(\Omega)$,*

$$\inf_{v \in V_n^0} \|u - v\|^2_{H^s(\Omega)} \leq C \ 2^{2(s-t)(-1/d)} 2^{2(s-t)\, n/d} \|u\|^2_{H^t(\Omega)}, \tag{1.111}$$

Supposons à présent que la fonction u est plus régulière, *i.e.* $u \in \mathcal{H}^t(\Omega)$, alors l'ordre d'approximation est à nouveau $(s-t)n$.

Proposition 1.25 *Soient s, t telles que $-\tilde{p} \leq s \leq t \leq p$. Si (1.101) est vérifiée alors il existe une constante $C > 0$ telle que pour toute fonction $u \in \mathcal{H}^t(\Omega)$:*

$$\inf_{v \in V_n^0} \|u - v\|^2_{H^s(\Omega)} \leq C \ 2^{-2(t-s)n} \|u\|^2_{\mathcal{H}^t(\Omega)}. \tag{1.112}$$

En tenant compte de la remarque 1.10, (1.112) devient

$$\inf_{v \in V_n^0} \|u - v\|^2_{H^s(\Omega)} \leq \begin{cases} C \ 2^{-2t\ n} n^{d-1} \|u\|^2_{\mathcal{H}^t(\Omega)}, & \text{pour } s = 0 \text{ et } t = p, \\ C \ 2^{-2(t-s)n} \|u\|^2_{\mathcal{H}^t(\Omega)}, & \text{sinon.} \end{cases} \tag{1.113}$$

Preuve

$$\inf_{v \in V_n^0} \|u - v\|^2_{H^s(\Omega)} \leq \left\| u - \sum_{\boldsymbol{\ell} \in I_n^0} w_{\boldsymbol{\ell}} \right\|^2_{H^s(\Omega)} \leq \sum_{|\boldsymbol{\ell}|_1 \geq n} 2^{2s|\boldsymbol{\ell}|_\infty} \|w_{\boldsymbol{\ell}}\|^2_{L^2(\Omega)}. \tag{1.114}$$

Dans le cas $t < p$, (1.106) permet de montrer que

$$\left\| u - \sum_{\boldsymbol{\ell} \in I_n^0} w_{\boldsymbol{\ell}} \right\|^2_{H^s(\Omega)} \leq \max_{|\boldsymbol{\ell}|_1 \geq n} 2^{2s|\boldsymbol{\ell}|_\infty - 2t|\boldsymbol{\ell}|_1} \sum_{|\boldsymbol{\ell}|_1 \geq n} 2^{2t|\boldsymbol{\ell}|_1} \|w_{\boldsymbol{\ell}}\|^2_{L^2(\Omega)}. \tag{1.115}$$

En utilisant (1.107) ($t = t$ et $s = 0$) et $1/d|\boldsymbol{\ell}|_1 \leq |\boldsymbol{\ell}|_\infty \leq |\boldsymbol{\ell}|_1$, nous obtenons :

$$\inf_{v \in V_n^0} \|u - v\|^2_{H^s(\Omega)} \leq \max_{|\boldsymbol{\ell}|_1 \geq n} 2^{2s|\boldsymbol{\ell}|_\infty - 2t|\boldsymbol{\ell}|_1} \|u\|^2_{\mathcal{H}^t(\Omega)} = \max_{|\boldsymbol{\ell}|_1 \geq n} 2^{2(s-t)|\boldsymbol{\ell}|_1} \|u\|^2_{\mathcal{H}^t(\Omega)}. \tag{1.116}$$

Dans le cas $t = p$ et $s < p$, il faut tenir compte de la remarque 1.10. L'Eq.(1.107) n'est plus applicable, nous avons recours à la majoration $2^{2t|\boldsymbol{\ell}|_1} \|w_{\boldsymbol{\ell}}\|^2_{L^2} \leq C \|u\|^2_{\mathcal{H}^t}$.

$$\begin{aligned} \left\| u - \sum_{\boldsymbol{\ell} \in I_n^0} w_{\boldsymbol{\ell}} \right\|^2_{H^s} &\leq C \sum_{|\boldsymbol{\ell}|_1 \geq n} 2^{2s|\boldsymbol{\ell}|_\infty - 2t|\boldsymbol{\ell}|_1} \|u\|^2_{\mathcal{H}^t} \\ &\leq C 2^{-2(t-s)(n-1)} \sum_{m=n}^{\infty} 2^{-2(t-s)(m-(n-1))} \underbrace{\sum_{|\boldsymbol{\ell}|_1 = m} 2^{2s(|\boldsymbol{\ell}|_\infty - m)}}_{\mathcal{A}_m} \|u\|^2_{\mathcal{H}^t}. \end{aligned} \tag{1.117}$$

Si $s > 0$, alors $\mathcal{A}_m \leq 1$. Dans le cas $s = 0$,

$$\mathcal{A}_m = \sum_{|\boldsymbol{\ell}|_1 = m} 1 \underbrace{=}_{(2.27)} \binom{d-1+m}{m} \leq C m^{d-1}. \tag{1.118}$$

Il reste alors à montrer que

$$\sum_{m=n}^{\infty} 2^{-2tm} m^{d-1} \approx C 2^{-2tn} n^{d-1}. \tag{1.119}$$

Ce dénombrement se calcule avec l'étude de l'intégrale $\int_n^{\infty} 2^{-2tx} x^{d-1} dx$.

$$\begin{aligned} b_{n,k} &= \int_n^{\infty} 2^{-2tx} x^k = \frac{2^{-2tn}}{2t} n^k + \frac{k}{2t} \int_n^{\infty} 2^{-2tx} x^{k-1} \\ &= \frac{2^{-2tn}}{2t} n^k + \frac{k}{2t} b_{n,k-1} = \frac{2^{-2tn}}{2t} n^k \left(\sum_{m=0}^{k} \binom{k}{m} (2tn)^{-m} \right) \\ &= \frac{1}{2t} 2^{-2tn} n^k \left(1 + \frac{1}{2tn} \right)^k. \end{aligned} \tag{1.120}$$

■

Remarque 1.14 *Dans le cas s négatif, une détérioration de l'ordre d'approximation est constatée (voir [GOS99]). Dans l'équation (1.116),* $2s|\boldsymbol{\ell}|_\infty - 2t|\boldsymbol{\ell}|_1 \leq 2\left(s|\boldsymbol{\ell}|_1/d - t|\boldsymbol{\ell}|_1\right) = 2\left(\frac{s}{d} - t\right)|\boldsymbol{\ell}|_1$, *alors l'erreur d'approximation est de la forme* $O\left(2^{-2n\left(t-\frac{s}{d}\right)}\right)$.

Théorème 1.26 (Erreur de projection sur les espaces sparse) *Soient* $-\tilde{p} \leq s \leq t+q \leq p$, $t+q \geq 0$ *et* $0 \leq t < p$. *Si (1.101) est vérifiée alors la fonction* $u \in \mathcal{H}^{t,q}(\Omega)$ *peut être approchée par une fonction* $v \in V_n^\tau$ *et :*

$$\inf_{v \in V_n^\tau} \|u - v\|^2_{H^s(\Omega)} \leq \begin{cases} C\ 2^{2\left(s-q-t-(\tau t+s-q)\frac{d-1}{d+\tau}\right)n} \|u\|^2_{\mathcal{H}^{t,q}(\Omega)}, & \text{pour } \tau \leq \frac{q-s}{t}, \\ C\ 2^{2(s-q-t)n} \|u\|^2_{\mathcal{H}^{t,q}(\Omega)}, & \text{pour } \tau > \frac{q-s}{t}. \end{cases} \tag{1.121}$$

Preuve La majoration

$$\inf_{v \in V_n^\tau} \|u - v\|^2_{H^s} \leq \max_{|\boldsymbol{\ell}|_1 + \tau|\boldsymbol{\ell}|_\infty \geq n} 2^{2(s-q)|\boldsymbol{\ell}|_\infty - 2t|\boldsymbol{\ell}|_1} \|u\|^2_{\mathcal{H}^t}, \tag{1.122}$$

s'obtient avec un raisonnement analogue à (1.115).

Il reste alors à majorer $|\boldsymbol{\ell}|_1$ et $|\boldsymbol{\ell}|_\infty$.

$$|\boldsymbol{\ell}|_1 + \tau|\boldsymbol{\ell}|_\infty \leq n(1+\tau) - 1, \text{ alors } |\boldsymbol{\ell}|_1 \leq n\frac{d(1+\tau)}{d+\tau} - \frac{d}{d+\tau} \text{ et } |\boldsymbol{\ell}|_\infty \leq n\frac{1+\tau}{d+\tau} - \frac{1}{d+\tau}. \tag{1.123}$$

L'équation (1.121) se déduit alors de (1.122). ■

1.3.2 L'espace sparse V_n^0

La complexité des méthodes numériques appliquées sur l'espace discret V_n^0 dépend du nombre de fonctions de base de cet espace. Ce paragraphe propose d'évaluer la dimension de cet espace dans différentes configurations.

En dimension 2, l'espace V_n^0 est obtenu en prenant la somme directe des espaces $W_{(\ell_1,\ell_2)}$ pour des multi-niveaux $\boldsymbol{\ell} = (\ell_1, \ell_2)$ au dessus de la diagonale sur la figure 1.7.

Dans ce qui suit, nous ferons l'amalgame entre points de la grille et fonctions de base de l'espace discret.

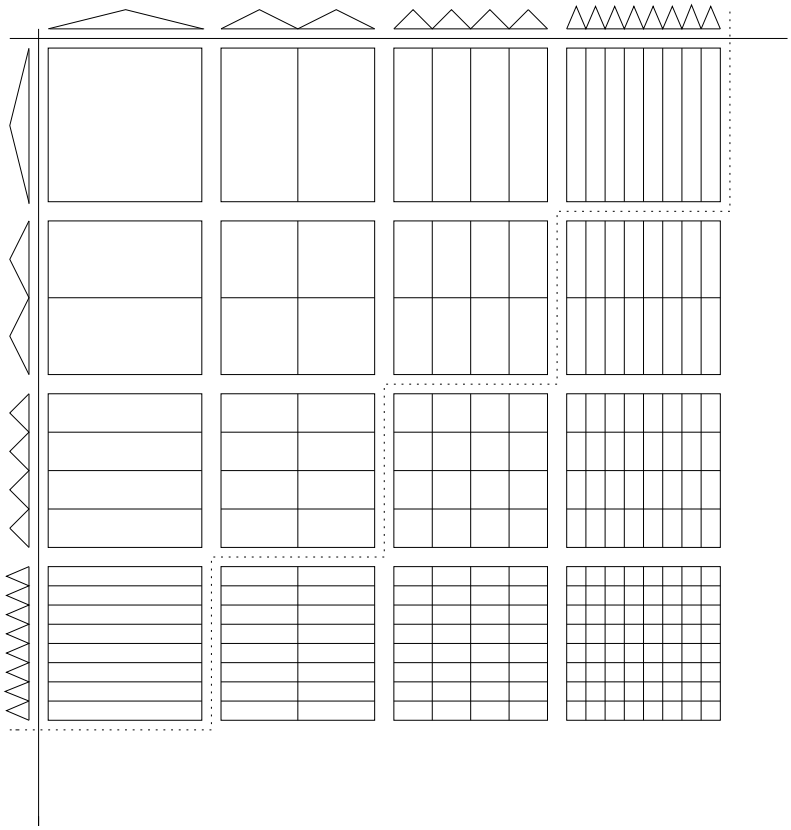

FIG. 1.7 – Représentation de la grille Sparse V_n^0

Nombre de points de grille L'étude est proposée à partir de la numérotation sur les fonctions chapeaux multi-échelle, ceci justifie la borne apparaissant dans le symbole de sommation :$|\boldsymbol{\ell}|_1 \leq n + d - 1$.

Dans le cas de grilles sans point de bord, le nombre de points, noté $a_{n,d}$ où n représente le niveau et d la dimension, vérifie la relation de récurrence :

$$a_{n,d} = a_{n,d-1} + 2a_{n-1,d}. \tag{1.124}$$

Preuve

$$a_{n,d} \overset{\text{def}}{=} \left|V_n^0\right| = \left|\bigoplus_{\boldsymbol{\ell}\in I_n^0} W_{\boldsymbol{\ell}}\right| = \sum_{\boldsymbol{\ell}\in I_n^0} 2^{|\boldsymbol{\ell}-\mathbf{1}|_1} = \sum_{i=1}^{n-1} 2^i \sum_{|\boldsymbol{\ell}-\mathbf{1}|_1=i} 1 = \sum_{i=d}^{n+d-1} 2^{i-d} \sum_{|\boldsymbol{\ell}|_1=i} 1. \tag{1.125}$$

Soit $\alpha_i^d = \sum_{|\boldsymbol{\ell}|_1=i} 1$, alors $\alpha_i^{d-1} + \alpha_i^d = \alpha_{i+1}^d$ et

$$\alpha_i^d = \binom{i-1}{d-1}. \tag{1.126}$$

$$\begin{aligned} a_{n,d} &= \sum_{i=d}^{n+d-1} 2^{i-d}\left(\alpha_{i-1}^{d-1} + \alpha_{i-1}^{d}\right) \\ &= \sum_{i=d}^{n+d-1} 2^{i-d}\alpha_{i-1}^{d-1} + \sum_{i=d}^{n+d-1} 2^{i-d}\alpha_{i-1}^{d} \\ &= \sum_{i=d-1}^{n+(d-1)-1} 2^{i-(d-1)}\alpha_{i}^{d-1} + 2\sum_{i=d-1}^{(n-1)+d-1} 2^{i-d}\alpha_{i}^{d} \\ &= a_{n,d-1} + 2a_{n-1,d}. \end{aligned} \tag{1.127}$$

■

Le comportement asymptotique de cette récurrence donne une approximation du nombre de points de grille :

$$a_{n,d} \approx 2^n \left(\frac{n^{d-1}}{(d-1)!}\right). \tag{1.128}$$

Dans le cas de grilles avec les points de bord, le nombre de points de grille, noté $b_{n,d}$, vérifie la relation :

$$b_{n,d} = \sum_{k=0}^{d} \binom{d}{k} 2^{d-k} a_{n,d-k}, \text{ avec } a_{n,0} = 1 \tag{1.129}$$

En effet, la grille $b_{n,d}$ est obtenue en ajoutant aux points intérieurs (la grille $a_{n,d}$) une grille de dimension $(d-k)$ sur chacune des faces de dimension k de l'hypercube, soit $\binom{d}{k}2^{d-k}$ faces.

Nous déduisons de (1.129) le comportement asymptotique du nombre de points de grille :

$$b_{n,d} \gtrsim 2^d \left(1 + \frac{1}{2n}\right)^d a_{n,d}. \tag{1.130}$$

Dans le cas de *Sparse Grid* de niveau n, le nombre de points de grille est au moins multiplié par 2^d en ajoutant les points de bord. En conséquence, une condition de Dirichlet ou de Neumann non-homogène implique une augmentation importante de la complexité du problème. Nous chercherons donc à formaliser le problème de manière à imposer des conditions aux bords de Dirichlet homogènes.

Le nombre $a_{n,d}$ (*resp.* $b_{n,d}$) de points de grille obtenu pour les niveaux $n \in \{7, \ldots, 10\}$ et les dimensions $d \in \{3, \ldots, 7\}$ est donné au tableau 1.1 (*resp.* 1.2). Ces informations sont synthétisées au tableau 1.3, nous y illustrons l'influence des conditions aux bords sur la complexité du problème en donnant le rapport : « nombre de points intérieurs / nombre de points de grille » pour les valeurs de n et d précédemment mentionnées.

Remarque 1.15 *Sous certaines hypothèses de régularité supplémentaires sur la fonction u, il est possible d'obtenir une approximation en norme $H^1(\Omega)$ de u d'ordre 2^n à l'aide de grilles plus creuses que V_n^0. En particulier, il est possible de trouver des grilles dont la dimension est indépendante de n^{d-1}. Leur dimension ne dépend de d que de manière algébrique. Le lecteur trouvera dans [Tod03] un exemple d'utilisation de ces grilles.*

TAB. 1.1 – Nombre de points intérieurs

$a_{n,d}$	7	8	9	10
3	2 816	7 484	18 944	47 104
4	7 938	23 298	6 5538	178 178
5	18 944	61 184	187 904	553 984
6	40 194	141 570	471 042	1 496 066
7	78 080	297 728	1 066 496	3 629 055
10	397825	1 862 145	8 085 505	32 978 945

TAB. 1.2 – Nombre de points pour une grille avec ses bords

$b_{n,d}$	7	8	9	10
3	8 962	21 250	49 666	114 690
4	59 145	148 223	364 553	882 697
5	369 695	975 125	$2.5 \cdot 10^6$	$6.3 \cdot 10^6$
6	$2.5 \cdot 10^6$	$5.9 \cdot 10^6$	$1.6 \cdot 10^7$	$4.3 \cdot 10^7$
7	$1.18 \cdot 10^7$	$3.43 \cdot 10^7$	$9.67 \cdot 10^7$	$2.65 \cdot 10^8$
10	$1.41 \cdot 10^9$	$4.65 \cdot 10^9$	$1.47 \cdot 10^{10}$	$4.51 \cdot 10^{10}$

1.3.3 Espaces d'approximation sparse sur une AMR interpolante

Les résultats présentés au § 1.3.1 nécessitent l'hypothèse (1.101) qui n'est pas vérifiée dans le cas de l'AMR interpolante $\left(\phi^2, \widetilde{\phi}^{2,0}\right)$, c.-à-d. la base hiérarchique présentée au § 1.2.2.2. Nous précisons ici le résultat de convergence obtenu pour l'opérateur d'interpolation.

Pour les raisons précédemment évoquées, le choix de l'AMR implique que l'opérateur de projection sur la base d'ondelettes $\{\psi_{\boldsymbol{\ell},\boldsymbol{i}}\}_{\boldsymbol{\ell}\in I_n}$ de l'espace discret V_n se confond avec l'opérateur d'interpolation.

Proposition 1.27 (Interpolation sur une *Sparse Grid*) *Soient l'AMR interpolante* $\left(\phi^2, \widetilde{\phi}^{2,0}\right)$, *et une fonction u appartenant à* $\mathcal{C}^2_{mix}(\bar{\Omega})$ *alors l'opérateur de projection* P_n *sur la base d'ondelette* $\{\psi_{\boldsymbol{\ell},\boldsymbol{i}}\}_{\boldsymbol{\ell}\in I_n^0}$ *de* V_n *vérifie*

$$\|u - P_n(u)\|_{L^2(\Omega)} \leq C 2^{-2n} n^{\frac{d-1}{2}} \|u\|_{\mathcal{C}^2_{mix}}, \tag{1.131}$$

et

$$\|u - P_n(u)\|_{L^\infty(\Omega)} \leq C 2^{-n} n^{d-1} \|u\|_{\mathcal{C}^2_{mix}}. \tag{1.132}$$

Preuve Pour simplifier les calculs, nous choisissons de normaliser les ondelettes de telle façon que $\|\psi_{\boldsymbol{\ell},\boldsymbol{i}}\|_{L^2(\Omega)} = \|\psi\|_{L^2(\Omega)}$.

Les propriétés de l'analyse-multi-résolution permettent de décomposer u sur les espaces de détails $W_{\boldsymbol{\ell}}$, nous déduisons de l'estimation inverse (lemme 1.12) que

$$\|u - P_n(u)\|^2_{L^2(\Omega)} \leq \sum_{|\boldsymbol{\ell}|_1 \geq n} \|w_{\boldsymbol{\ell}}\|^2_{L^2(\Omega)} = \sum_{|\boldsymbol{\ell}|_1 \geq n} \sum_{\boldsymbol{i}\in\tau_{\boldsymbol{\ell}}} \left\langle \widetilde{\psi}_{\boldsymbol{\ell},\boldsymbol{i}}, u \right\rangle^2 \|\psi_{\boldsymbol{\ell},\boldsymbol{i}}\|_{L^2(\Omega)}. \tag{1.133}$$

La démonstration consiste à majorer chacun des termes de cette double somme.

TAB. 1.3 – Rapport : nombre de points intérieurs / nombre de points de grille

$\frac{a_{n,d}}{b_{n,d}}$	7	8	9	10
3	31%	35%	38%	41%
4	13%	16%	18%	20%
5	5%	6%	7.5%	9%
6	1.8%	2.4%	2.9%	3.5%
7	0.7%	0.9%	1.1%	1.4%
10	0.03%	0.04%	0.05%	0.07%

Lemme 1.28 *Si* $u \in \mathcal{C}^2_{mix}(\bar{\Omega})$ *alors*

$$\left|\left\langle \widetilde{\psi}_{\boldsymbol{\ell},\boldsymbol{\imath}}, u \right\rangle\right|^2 \leq C\|u\|^2_{\mathcal{C}^2_{mix}(\bar{\Omega})} 2^{-5|\boldsymbol{\ell}|_1}. \tag{1.134}$$

Admettons provisoirement ce lemme.

Il nous reste à dénombrer les termes de la somme de (1.131), sachant que $\sum_{\boldsymbol{\imath}\in\tau_{\boldsymbol{\ell}}} 1 = 2^{|\boldsymbol{\ell}|_1}$. Nous obtenons :

$$\|u - P_n(u)\|^2_{L^2(\Omega)} \leq C\|u\|^2_{\mathcal{C}^2_{mix}} \sum_{|\boldsymbol{\ell}|_1 \geq n} 2^{-4|\boldsymbol{\ell}|_1}. \tag{1.135}$$

La conclusion se déduit de (1.118) et (1.119) avec $t = 2$.

Dans le cas de la norme $L^\infty(\Omega)$, l'équation (1.133) doit être remplacée par,

$$\begin{aligned} \|u - P_n(u)\|_{L^\infty(\Omega)} &\leq \sum_{|\boldsymbol{\ell}|_1 \geq n} \|w_{\boldsymbol{\ell}}\|_{L^\infty(\Omega)} = \sum_{|\boldsymbol{\ell}|_1 \geq n} \sum_{\boldsymbol{\imath}\in\tau_{\boldsymbol{\ell}}} \left\langle \widetilde{\psi}_{\boldsymbol{\ell},\boldsymbol{\imath}}, u \right\rangle \|\psi_{\boldsymbol{\ell},\boldsymbol{\imath}}\|_{L^\infty(\Omega)} \\ &\leq C\|u\|_{\mathcal{C}^2_{mix}} \sum_{|\boldsymbol{\ell}|_1 \geq n} \sum_{\boldsymbol{\imath}\in\tau_{\boldsymbol{\ell}}} 2^{-5|\boldsymbol{\ell}|_1/2}\ 2^{|\boldsymbol{\ell}|_1/2} \end{aligned} \tag{1.136}$$

La conclusion se déduit de (1.119) avec $t = 1$. ■

Preuve du lemme Ce lemme se déduit de (1.100),

$$\left|\left\langle \widetilde{\psi}_{\boldsymbol{\ell},\boldsymbol{\imath}}, u \right\rangle\right|^2 = 2^{-4|\boldsymbol{\ell}|_1} \left(\int_\Omega \frac{\partial^{2d} u}{\partial x_1^2 \dots \partial x_d^2} \psi_{\boldsymbol{\ell},\boldsymbol{\imath}}(\boldsymbol{x})\, dx \right)^2 \leq \|u\|^2_{\mathcal{C}^2_{mix}(\Omega)} 2^{-4|\boldsymbol{\ell}|_1} \left\| \widetilde{\psi}_{\ell,\imath} \right\|^2_{L^1(\Omega)}. \tag{1.137}$$

La norme 1 de l'ondelette est donnée par le produit des normes 1 dans chacune des dimensions.

$$\int_{\mathbb{R}} \psi_{\ell,\imath}(x)dx = \int_{\mathbb{R}} 2^{\ell/2}\psi\left(2^\ell x - \imath\right) dx = 2^{-\ell/2}\|\psi\|_{L^1(\Omega)}.$$

■

Chapitre 2

Méthode de résolution numérique d'EDP sur une *Sparse Grid*

L'objectif de ce chapitre est de décrire les différentes méthodes de résolution numérique *d'équations aux dérivées partielles* et *d'équations intégro-différentielles* sur une *Sparse Grid*. Les méthodes présentées ici prennent appui sur les notions d'analyse multi-résolution et de produit tensoriel sparse. Ces méthodes s'appliquent aux équations posées sur des *domaines tensoriels*.

La première partie présente les techniques de quadrature introduites par Smolyak [Smo63].

Les deuxième, troisième et quatrième parties développent différentes méthodes de résolution *d'équations aux dérivées partielles*.

Dans la deuxième partie, la méthode de *technique combinatoire*, introduite par Griebel & al [GSZ92], est définie. Celle-ci permet la résolution d'équations linéaires en dimension $2, 3$, et 4 grâce à un algorithme hautement parallélisable.

La troisième partie est consacrée à la méthode des différences finies sur une *Sparse Grid*. Les opérateurs de différences finies classiques sur une *Sparse Grid* n'étant pas consistants, l'utilisation d'opérateurs de différences finies adaptés aux grilles multi-niveaux est nécessaire. Les résultats de consistance pour ces opérateurs, démontrés par Schiekofer [Sch98a] puis par Koster [Kos00], sont ici obtenus à l'aide d'une méthode de collocation.

La quatrième partie reprend les résultats de Schwab relatifs à la résolution d'une équation elliptique par la méthode de Galerkin sur une base d'ondelette Sparse. Des résultats expérimentaux obtenus en résolvant l'équation de Poisson sont également donnés.

La cinquième partie traite de l'approximation numérique de la solution d'équations intégrales. Les propriétés de compression des bases d'ondelettes pour ce type d'opérateurs sont rappelées. La méthode de collocation, introduite dans la troisième partie, permet de justifier la consistance d'un opérateur discret pour un opérateur intégral particulier.

La dernière partie de ce chapitre aborde la résolution de problèmes paraboliques à l'aide de méthodes d'ondelettes. Le recours à des schémas en temps aux propriétés « dissipatives » s'avère être judicieux. Certains de ces schémas, fondés sur des méthodes de Galerkin discontinues, sont présentés.

2.1 Formules de quadrature de Smolyak

Dans cette partie, nous étudions les formules de quadrature de Smolyak [Smo63] (ou formules de quadrature Sparse). Divers résultats numériques sont présentés, ceux-ci visent à valider une méthode d'évaluation des opérateurs intégraux de la forme $\int_\Omega (u(x+z)-u(x))\,k(z)dz$. Ces opérateurs apparaissent naturellement dans les équations utilisées en finance. Ils interviennent, par exemple, dans des *équations intégro différentielles* pour l'évaluation d'options européennes obtenues dans le cadre d'un modèle à saut de type Poisson.

2.1.1 Description

Pour approcher l'intégrale d'une fonction sur un intervalle, il existe de nombreuses formules de quadrature. Citons quelques-unes de ces formules parmi les plus utilisées,

- les formules de Newton-Côtes (trapèze, Simpson).
- les formules de Gauss (Legendre, Hermite, Lobatto,...).
- la formule de Clenshaw-Curtis,

$$\int_{-1}^{1} f(x)\,dx \approx a_0 + \sum_{k=1}^{n-1} \frac{2a_k}{1-k^2}, \tag{2.1}$$
$$\text{avec } a_k = \frac{f(1)}{n} + \frac{f(-1)}{n}(-1)^k + \frac{2}{n}\sum_{i=1}^{n-1} f(\cos[\frac{i\pi}{n}])\cos(k\frac{i\pi}{n}).$$

- les formules de Gauss-Kronrod [Kro65] (ou Formule de Patterson dans le cas des polynômes de Legendre [Pat68]).

Remarque 2.1 *Une remarque sur l'utilisation de ces formules et sur le choix de l'une d'entre elles : hormis le domaine d'intégration et l'utilisation d'une fonction de poids dans le calcul de l'intégrale, le choix de la formule repose également sur l'utilisation ou non de techniques d'adaptation pour le calcul de l'intégrale.*

Afin de déterminer un critère local d'adaptation pour une formule numérique de calcul d'intégrale, il est nécessaire de faire deux évaluations de cette formule :

- *une première évaluation grossière de la formule telle que les points ζ_1, ζ_2 appartiennent à l'ensemble des points de quadrature.*
- *un second calcul en considérant une formule plus fine pour laquelle un point $\zeta_{12} \in]\zeta_1, \zeta_2[$ est ajouté à l'ensemble des points de quadrature.*

Lorsque l'écart entre les deux calculs dépasse un seuil de tolérance, le point ζ_{12} est effectivement ajouté à l'ensemble des points de quadrature.

Une formule de Newton-Côtes de niveau n (formule pour laquelle les points de quadrature sont équirépartis) utilise les points de quadrature du niveau $n/2$.
Cette propriété n'est plus vérifiée par les formules de Gauss. (Les racines d'un polynôme de degré N ne sont pas, en général, racines d'un polynôme de degré $N+K$.) En conclusion, ces formules se prêtent assez mal aux techniques adaptatives.
Les formules de Gauss-Kronrod vérifient à la fois la propriété d'invariance des points de quadrature entre les niveaux de raffinement et la plupart des propriétés des formules de Gauss. La formule de Clenshaw-Curtis se prête également à l'adaptation.

L'écriture générique pour ces formules de quadrature est la suivante :

$$Q_\ell(u) \stackrel{\text{def}}{=} \sum_{i=1}^{n_\ell} \omega_i u(\zeta_i), \tag{2.2}$$

où les $(\omega_i)_{1\leq i\leq 2^{\ell-1}+1}$ sont les poids de quadrature (supposés positifs) et les $(\zeta_i)_{1\leq i\leq 2^{\ell-1}+1}$ les points de quadrature. Considérons des formules de quadrature dont le nombre de points n_ℓ est de la forme $n_\ell = 2^{\ell-1}+1$.

A cette formule de quadrature est associée une erreur de quadrature $E_{n_\ell} \stackrel{\text{def}}{=} \left\| \int_\Omega u - Q_\ell u \right\|_\infty$. Nous supposerons que l'erreur de quadrature est de la forme :

$$E_{n_\ell}(u) = O\left(n_\ell^{-m}\right), \quad \text{pour toute fonction } u \in \mathcal{C}^m(\Omega). \tag{2.3}$$

Des formules de quadrature sur un domaine rectangulaire de $\mathbb{R}^d, d>1$, s'obtiennent par tensorisation des formules de quadrature sur l'intervalle. Ces formules nécessitent $(n_\ell)^d$ points de quadrature (n_ℓ est le nombre de points utilisés pour l'intégrale sur l'intervalle). Ces méthodes deviennent rapidement trop coûteuses en temps et en mémoire si d est supérieure à 3.

Une alternative est proposée par Smolyak [Smo63] dans le cas de fonctions suffisamment régulières.

Définition 2.1 *[Quadrature de Smolyak] Considérons un domaine de la forme $\mathbf{\Omega} = \Omega_1 \times \cdots \times \Omega_d$, et les formules de quadrature sur l'intervalle définies par (2.2), notées $Q^i_{\ell_i}$ pour $1 \leq i \leq d$.*

La méthode de quadrature de Smolyak introduit la différence entre deux niveaux de quadrature (sur l'intervalle) :

$$\Delta^i_{\ell_i}(u) \stackrel{\text{def}}{=} \left(Q^i_{\ell_i} - Q^i_{\ell_i-1}\right)(u), \quad \textit{avec } Q^i_0(u) = 0. \tag{2.4}$$

La formule est alors obtenue en sommant les différences de quadrature $\Delta^i_{\ell_i}$ pour $|\boldsymbol{\ell}|_1 \leq n+d-1$:

$$\boldsymbol{Q}^{Sparse}_n(u) \stackrel{\text{def}}{=} \sum_{\boldsymbol{\ell}:|\boldsymbol{\ell}|_1\leq n+d-1} \left(\Delta_{\ell_1} \otimes \cdots \otimes \Delta_{\ell_d}\right)(u). \tag{2.5}$$

Cette formule peut s'écrire en faisant apparaître les Q_{ℓ_i},

$$\boldsymbol{Q}^{Sparse}_n(u) = \sum_{\boldsymbol{\ell}:n\leq|\boldsymbol{\ell}|_1\leq n+d-1} (-1)^{n+d-|\boldsymbol{\ell}|_1-1} \begin{pmatrix} d-1 \\ |\boldsymbol{\ell}|_1 - n \end{pmatrix} \left(Q_{\ell_1} \otimes \cdots \otimes Q_{\ell_d}\right)(u). \tag{2.6}$$

Ce dernier résultat se déduit assez naturellement de la preuve du théorème 2.2).

Lorsque $\mathbf{\Omega}$ est un domaine de $\mathbb{R}^d$ obtenu par produit tensoriel d'intervalles de $\mathbb{R}$, le résultat (2.3) se généralise aux formules de quadrature obtenues par tensorisation de formules de type (2.2). Si les formules en dimension un vérifient (2.3), alors

$$E_I(u) = O\left(n_\ell^{-m}\right), \quad \text{pour toute fonction } u \in \mathcal{C}^m(\mathbf{\Omega}). \tag{2.7}$$

Dans le cas des formules de quadrature de Smolyak, la fonction u doit respecter une hypothèse de régularité supplémentaire.

Théorème 2.1 *Considérons la formule de quadrature de Smolyak définie par (2.5) et l'erreur de quadrature* $E_n^{Sparse} = \left\| \int_{\mathbf{\Omega}} u - \boldsymbol{Q}_n^{Sparse} u \right\|_\infty$. *Nous disposons du résultat d'approximation suivant :*

$$E_n^{Sparse}(u) = O\left(n_\ell^{-m} |\log n_\ell|^{(d-1)(m+1)}\right), \quad \text{pour toute fonction } u \in X^{m,\infty}(\mathbf{\Omega}). \tag{2.8}$$

Le lecteur trouvera une définition de l'espace $X^{m,\infty}(\mathbf{\Omega})$ au § 1.1.3.

Remarque 2.2 *Le résultat de convergence dépend peu de la dimension d du problème, mais il dépend fortement de la régularité m de la fonction.*

Preuve Le lecteur trouvera dans [WW95] une démonstration de ce résultat. ■

Remarque 2.3 *Nous donnerons dans § 2.2 les idées de démonstration de (2.5).*

Remarque 2.4 *Citons la généralisation de ce résultat pour une intégration sur des espaces à poids [PW01].*

2.1.2 Résultats numériques

Notre objectif est de proposer une méthode permettant de prendre en compte l'effet des sauts dans le cas de l'évaluation de prix d'options européennes pour un modèle à volatilité stochastique auquel des sauts de Poisson sur le spot et la volatilité sont ajoutés. Nous décrirons dans une partie ultérieure ce modèle. Néanmoins, le problème de la prise en compte de l'effet des sauts revient à évaluer l'intégrale du produit d'une fonction régulière par une gaussienne. Supposons une approximation polynomiale de cette fonction, le problème se ramène alors à évaluer les différents moments d'une gaussienne. Ces hypothèses nous ont conduits aux tests suivants.

Dans le cas particulier de la formule de Newton Cotes $\left(\omega_i = 2^{-(\ell-1)},\ x_{\ell,\imath} = (\imath - 1)\, 2^{-(\ell-1)}, \imath \in \{1 \dots d\}\right)$, les points de quadrature sont ceux de la grille sparse standard. La formule de quadrature (2.6) devient :

$$\boldsymbol{Q}_n^{\text{sparse}}(u) = \sum_{\boldsymbol{\ell} \in \mathcal{I}_n^0} (-1)^{n+d-|\boldsymbol{\ell}|_1 - 1} \begin{pmatrix} d-1 \\ |\boldsymbol{\ell}|_1 - n \end{pmatrix} 2^{-|\boldsymbol{\ell}|_1 + d} \sum_{\boldsymbol{i}}^{2^{\boldsymbol{\ell}-\mathbf{1}}+\mathbf{1}} u(\boldsymbol{x}_{\boldsymbol{\ell},\boldsymbol{i}}), \tag{2.9}$$

où $\boldsymbol{x}_{\boldsymbol{\ell},\boldsymbol{i}}$ sont les points de la grille dyadique.

Ces formules sont implémentées sur le domaine $[0,1]^d$. Les résultats numériques obtenus sont présentés pour deux fonctions f et pour des dimensions d comprises entre 2 et 6, voir les tables 2.1, 2.2 dans le cas :

- d'une parabole centrée en $\left(\frac{1}{2}, \dots, \frac{1}{2}\right)$: $f(\boldsymbol{x}) := \prod_{i=1}^d x_i (1 - x_i)$.
- d'une gaussienne multi-dimensionnelle de moyenne $\mu = (1.0, 0.5, 0.6)$ et de variance $\sigma = (0.2, 0.15, 0.30)$. (Seules les premières valeurs sont utilisées lorsque la dimension est inférieure à la taille de μ). Le changement de variable $x = \tan\left(\left(z - \frac{1}{2}\right)\pi\right)$ permet de se ramener au domaine $[0,1]^d$.
  Les résultats sont donnés en fonction de la dimension et des moments de la distribution. Le moment d'ordre 2 est obtenu par la formule $\prod_i \sigma_i^2 + \mu_i^2$.

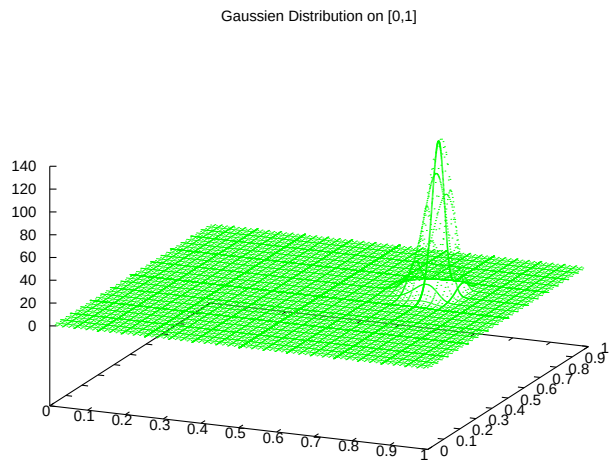

FIG. 2.1 – Distribution gaussienne $\mu = (1.0, 0.5, 0, 0)$, $\sigma = (0.2, 0.15, 0, 0)$

– Considérons à présent le domaine $\Omega = [0,1]^4$. Nous étudions l'erreur commise en utilisant la formule de quadrature dans le cas d'une gaussienne dégénérée. Le vecteur moyenne est $\mu = (1.0, 0.5, 0.6, 0.7)$. La variance est $\sigma = (0.2, 0.15, 0, 0)$ (voir la figure 2.1). En supprimant la variance sur les deux dernières composantes, l'intégration est une intégration suivant un plan. Nous cherchons à évaluer les performances d'une formule de quadrature sparse dans ce cas. Les résultats sont donnés pour un niveau de raffinement $n = 9$. La formule de quadrature de Smolyak permet d'obtenir :
  – l'erreur relative sur le moment d'ordre 0 : environ 0.01%,,
  – l'erreur sur le moment d'ordre 1 : environ 0.8%,
  – l'erreur sur le moment d'ordre 2 : environ 3%.

Analyse des tests numériques En dimension 2, les erreurs sont acceptables à partir du niveau 9. Pour des dimensions supérieures, le niveau d'erreur souhaité demande un niveau de raffinement incompatible avec les contraintes de temps de calcul.

TAB. 2.1 – Erreur d'intégration obtenue par la formule de quadrature de Newton Cotes Sparse dans le cas d'une parabole

Niveau	Valeur	Erreur Relative $1e-4$
dim =2	0.44444444	
6	0.443359	24.42
7	0.44409	7.97
8	0.4443359	2.44
9	0.444412231	0.72
10	0.444435	0.21
11	0.44444180	0.06
12	0.44444370	0.02
dim =3	0.296296296	
6	0.29590	13.4425
7	0.29638672	3.0518
8	0.29638672	3.0518
9	0.29634942	1.7931
10	0.29631424	0.6056
11	0.29630280	0.2194
12	0.29629850	0.0745
dim =4	0.197530864	
6	0.19921875	85.45
7	0.19824219	36.01
8	0.19775391	11.29
9	0.19758606	2.79
10	0.19754028	0.48
11	0.19753075	0.01
12	0.19752979	0.05
dim =5	0.131687243	
6	0.13476563	233.76
7	0.13232419	48.37
8	0.13171387	2.02
9	0.13163757	3.77
10	0.13165665	2.32
11	0.13167477	0.95
12	0.13168311	0.31
dim =6	0.087791495	
8	0.08740234	44.3
9	0.08763123	18.3
10	0.08774567	5.2

TAB. 2.2 – Erreur d'intégration obtenue par la formule de quadrature de Newton Cotes Sparse dans le cas d'une Gaussienne

Niveau	Valeur	Erreur Relative $1e-4$
dim = 2	moment 0	
6	1.13164587	1316.5
7	1.64918470	6491.8
8	0.96005392	399.5
9	0.99989445	1.06
10	1.00000868	0.09
11	1.00000001	0.00
12	1.00000000	0.00
dim = 3	moment 0	
7	6.88041649	58804.2
8	1.87271747	8727.2
9	0.44732202	5526.8
10	1.01320485	132.05
11	1.00505194	50.52
12	0.99991819	0.82
dim = 2	moment 1 0.50	
7	0.79640068	5928.0
8	0.53255642	651.1
9	0.49599020	80.2
10	0.50008204	1.64
11	0.50000026	0.01
12	0.50000000	0.00
dim = 3	moment 1 0.30	
9	−0.20697318	16899.1
10	0.27101223	966.26
11	0.32150374	716.79
12	0.30050942	16.98
dim = 2	moment 2 0.2834	
7	0.41502404	4644.5
8	0.34330648	2113.8
9	0.27529595	286.0
10	0.28358708	6.60
11	0.28340076	0.03
12	0.28340000	0.00
dim = 3	moment 2 0.12573	
10	−0.00165390	10131.5
11	0.16264540	2936.09
12	0.12958337	306.48

2.1.3 Compléments & perspectives

Dans cette partie, nous présentons les différentes applications financières de cette méthode rencontrées dans la littérature. Nous nous plaçons dans le cadre de l'évaluation d'une classe de produits dérivés sans clause d'exercice anticipée, la valeur de ce produit pouvant éventuellement dépendre de l'historique du sous-jacent. Dans ce cas, l'option est dite « *path dependant* », par exemple les options *loopback* ou les options *asian*.

Ceci nous place dans le cadre d'application de la théorie des martingales et du théorème de Feynam-Kac, qui permettent de montrer que le problème d'évaluation de l'option se ramène à un calcul intégral.

$$u(x,t) = \int_{\mathbb{R}} u_0(x - z(t)) \exp\left(\int_0^t V(x - z(s))\, ds\right) d\omega(z). \tag{2.10}$$

La résolution de (2.10) par la formule de quadrature de Smolyak est proposée dans [PWW00]. Griebel & Al montrent que l'utilisation d'une technique d'adaptation [GG98] permet de proposer une méthode compétitive par rapport aux méthodes de Monte Carlo (pour un certain nombre de problèmes).

Les points importants de l'algorithme et quelques remarques concernant cette méthode sont exposés ci-dessous.

- Le premier chapitre et le théorème 2.1 ont montré l'importance de l'hypothèse de régularité sur la vitesse de convergence des méthodes sparse. Lorsque la fonction à intégrer n'est pas suffisamment régulière, par exemple un put (voir l'exemple 1.1), Griebel propose d'utiliser une transformation pour ne considérer que le domaine sur lequel la fonction est suffisamment régulière.
  Prenons l'exemple d'un call sur un panier de strike K, le domaine d'intégration $\mathbb{R}^{+d}$ devient $\mathbb{R}^{+d} \cap \{S_1 + \cdots + S_d \geq K\}$.
  L'hypersurface sur laquelle la fonction est non-régulière est détectée par une méthode de recherche de zéro sur un problème en dimension 1 : méthode de *coordinate wise transformation* pour ramener le domaine sur un carré.
- La stratégie adaptative détecte les « vraies sources d'aléas » du problème, c.-à-d. les directions principales d'intégration.
  Dans le cadre de l'évaluation d'une option sur indice modélisée par un panier d'actions ($N = 40, 200, \ldots$), l'analyse en composante principale montre qu'un nombre d'aléas M raisonnable ($M << N$) permet de modéliser l'indice. (Cette idée est également utilisée par Reisinger [RW07] pour réduire la dimension des problèmes d'évaluation d'options panier par des méthodes d'EDP.) Lorsque M est très inférieur à N, l'algorithme adaptatif proposé par Griebel converge plus rapidement qu'une méthode de Monte-Carlo. Les résultats sont moins concluants lorsque cette propriété n'est plus vérifiée [GH08].

2.1.4 Formules de quadrature pour le calcul du second membre

L'objectif de ce paragraphe est de proposer une formule d'intégration numérique contenant peu de points (≤ 20) afin d'approcher l'intégrale multi-dimensionnelle :

$$\mathcal{I} = \int_{\Omega} f(x)dx, \text{ avec } \Omega = [-1,1]^d. \tag{2.11}$$

La formule de quadrature proposée tiendra compte de la propriété suivante sur la fonction f :

$$f_{|\partial\Omega} = 0. \tag{2.12}$$

D'après les résultats obtenus par Griebel [GG98], nous disposons de deux types de formules de quadrature qui correspondent à nos deux critères :

1. la formule de Clenshaw-Curtis adaptée au grille sparse.
2. la formule de Gauss-Patterson où la famille de polynômes est choisie en fonction de la propriété (2.12). Nous avons essayé d'adapter la méthode de Patterson en considérant les formules de Gauss-Lobatto. Le lecteur trouvera dans [BMR04] une présentation de cette méthode de quadrature dans le cas de la dimension 1. Cette formule diffère de celle de Gauss-Legendre dans la mesure où les points de bord font partie de l'ensemble des points de quadrature.

Dans le cas de la dimension 1, les formules de Gauss sont plus précises pour un nombre de points fixé que la formule des trapèzes ou Clenshaw-Curtis. Ce résultat reste valable dans le cas des formules « adaptatives ». La formule de Gauss-Patterson-Lobatto fournit une meilleure approximation que celle de Clenshaw-Curtis. Cependant, dans le cadre de la construction de Smolyak (2.6), le nombre de points obtenu pour la formule de Gauss-Patterson-Lobatto peut être très supérieur à celui obtenu pour la formule de Clenshaw-Curtis.

Nous présentons dans ce qui suit l'algorithme de construction de Gauss-Patterson appliqué aux formules de Gauss-Lobatto. Nous donnons les points obtenus en dimension 1 et les grilles sparse obtenues par application de (2.6). Une comparaison avec les grilles obtenues par la méthode de Clenshaw-Curtis nous permet de conclure que cette dernière méthode s'avère être mieux adaptée à notre problème.

Méthode de Gauss-Kronrod Lorsqu'une méthode de Gauss de raffinement n est appliquée, les points de quadrature, c.-à-d. les zéros de la dérivée du polynôme de Legendre n (noté L'_n), ne coïncident jamais avec les points de quadratures d'un raffinement plus élevé. Kronrod montre qu'il est possible de choisir un polynôme de degré $n-1+p$ dont $n-1$ racines sont les racines d'un polynôme L'_n. Notons $G(n-1+p,x)$ le polynôme de degré $n-1+p$ dont les zéros sont les points de quadrature de la formule d'intégration. Un polynôme f de degré $n+2p-1$ peut alors s'exprimer sous la forme :

$$f(x) = G(n-1+p,x)h(x) + g(x), \quad g(x) = \sum_{i=0}^{n+p-2} a_i x^i \text{ et } h(x) = \sum_{i=0}^{p-1} b_i x^i. \tag{2.13}$$

Si la formule de quadrature est exacte pour tout polynôme g de degré $n+p-2$ alors

$$\int_{-1}^{1} G(n-1+p,x)h(x)dx = 0. \tag{2.14}$$

Le polynôme h de degré $p-1$ peut s'exprimer comme une combinaison linéaire des L'_n. Ainsi, G vérifie

$$\int_{-1}^{1} G(n-1+p,x)L'_k(x)dx = 0, \quad k = 1,\ldots,p. \tag{2.15}$$

Dans nos applications, seuls les cas n pair sont étudiés. En effet, $n-1$ est donné par la relation $n-1=2^\ell-1$. Afin de ne pas introduire de points supplémentaires en augmentant le niveau ℓ, nous choisissons p tel que $n-1+p=2^{\ell+1}-1$ soit $p=2^\ell=n$. Soit le polynôme K tel que $G(n-1+p,x)=K(n,x)L'_n(x)$. La condition (2.12) est alors donnée par

$$\int_{-1}^{1} K(n,x)L'_n(x)L'_k(x)dx=0, \quad k=1,\ldots,n+1. \tag{2.16}$$

n étant pair, $K(n,x)$ est une fonction paire. Ce polynôme se décompose donc sur la base des L'_n suivant la formule

$$K(n,x)=\sum_{i=0}^{n/2} a_i L'_{2i+1}(x). \tag{2.17}$$

Les coefficients a_i sont alors calculés en résolvant le système

$$\sum_{i=0}^{n/2} a_i \int_{-1}^{1} L'_{2i+1}(x)L'_n(x)L'_k(x)dx=0, \quad k=1,\ldots,n+1. \tag{2.18}$$

Pour k impair, les équations (2.18) sont automatiquement satisfaites ($L'_{2i+1}\ L'_k$ est un polynôme pair alors que L'_n est un polynôme impair). Les coefficients a_i sont obtenus en résolvant le système de $(n+1)/2$ inconnues correspondantes aux valeurs impaires de k. Cette méthode fut introduite par Patterson [Pat89].

Il est donc possible en théorie de construire une base de polynômes $G(n-1,x)$ telle que les $n-1$ zéros de ce polynôme annulent $G(2n-1)$. Cependant, l'initialisation de la procédure n'est pas toujours possible. La méthode décrite précédemment et appliquée au cas $n=2$ donne les points de quadrature du tableau 2.3. Ces points sont calculés à l'aide de la librairie OPQ [Gau05] en suivant la méthode proposée dans [Lau01].

Tab. 2.3 – Points de quadrature Gauss-Patterson-Lobatto

niveau 2	niveau 3	niveau 4
0.000	0.000	0.000
0.654	0.654	0.677
	0.890	0.899
	0.340	0.363
		0.969
		0.800
		0.530
		0.183

La construction dyadique de la grille sparse est donc incompatible avec la construction proposée par Patterson. Pour cette raison, nous utilisons les formules de Clenshaw-Curtis dont les grilles sont données à la figure 2.2. Les grilles obtenues pour la méthode de Gauss-Patterson-Lobatto et Gauss-Legendre sont données à la figure 2.3.

Le nombre de points de quadrature en dimension 4 pour les formules de Clenshaw-Curtis est respectivement $(10,50,210)$ pour les niveaux de raffinement $(2,3,4)$. En conclusion, le calcul du produit scalaire de f avec toutes les fonctions de base $(\psi_{\ell,\imath})_{\boldsymbol{\ell}\in I_n,\boldsymbol{\imath}\in\tau_{\boldsymbol{\ell}}}$ de la grille est trop coûteux si nous utilisons cette méthode.

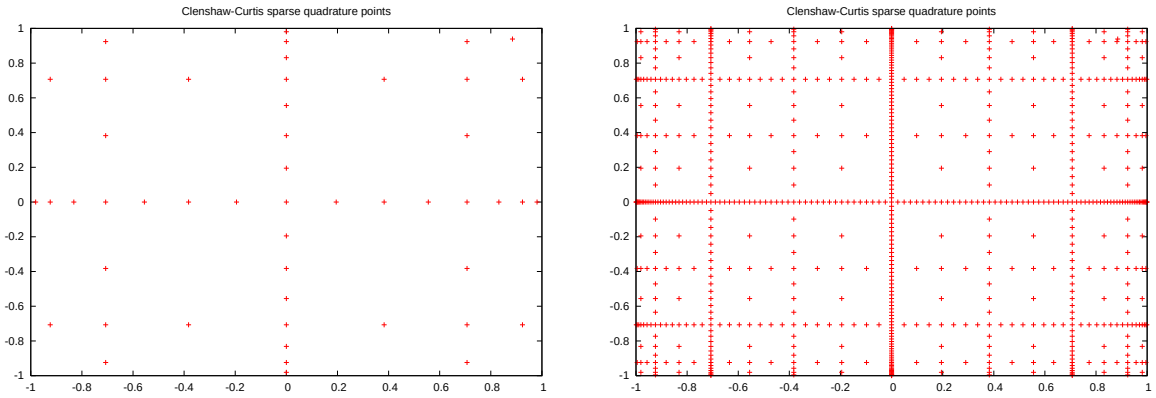

FIG. 2.2 – Grille de points de quadrature, Clenshaw-Curtis $d = 2$, $l = 4$ (gauche), $l = 7$ (droite)

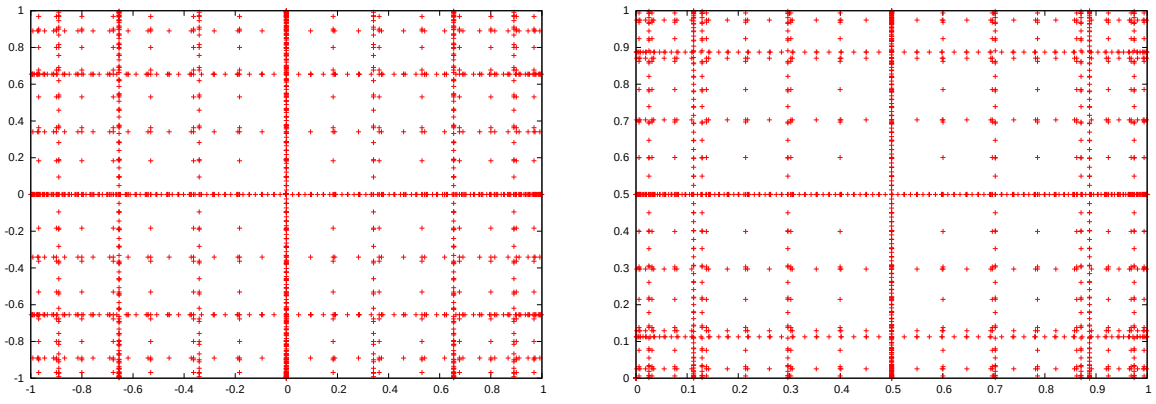

FIG. 2.3 – Grille de points de quadrature, $d = 2$ Gauss-Patterson Lobatto $l = 7$ (gauche), Gauss-Legendre $l = 7$(droite) 1813 pts

2.2 Technique combinatoire

La méthode de *technique combinatoire* consiste à décomposer la solution u d'une équation aux dérivées partielles comme une somme pondérée. Les termes de cette somme sont les solutions obtenues en résolvant le problème initial sur des grilles conventionnelles mais anisotropes de petite taille. A la différence des autres méthodes de résolution d'EDP (différences finies et Galerkin), cette résolution multiple est complètement parallélisable. Cette méthode est utilisée avec succès en mécanique quantique [Gar05] et en finance [Rei04, Bla04]. Avant de décrire cette méthode, nous donnons un résultat de décomposition d'une fonction u sur les espaces sparse décrits au chapitre précédent. Nous revenons ensuite sur le résultat de convergence de la méthode de *technique combinatoire*.

2.2.1 Décomposition sur des grilles anisotropes

Ce résultat permet de démontrer un théorème de convergence pour la méthode de *technique combinatoire*. Il s'applique également à la formule de quadrature de Smolyak (2.6) et intervient dans la démonstration de la consistance des opérateurs de différences finies.

Théorème 2.2 *Soit u_n la projection de u sur V_n^0 (voir la définition 1.18). Elle se décompose de la manière suivante :*

$$u_n = \sum_{j=0}^{d-1} (-1)^{d-1-j} \binom{d-1}{j} \sum_{|\boldsymbol{\ell}|_1 = n+j} u_{\boldsymbol{\ell}}, \tag{2.19}$$

où $u_{\boldsymbol{\ell}}$ est la projection de u sur l'espace tensoriel $V_{\boldsymbol{\ell}}$ (voir la définition 1.13).

Dans le cas de la méthode de *technique combinatoire*, les fonctions $u_{\boldsymbol{\ell}}$ sont les solutions de problèmes approchés sur des grilles pleines anisotropes, c'est-à-dire sur des maillages cartésiens de pas de raffinement différents suivant chacune des dimensions.

Preuve La démonstration dans le cas $d = 2$ présentée ici est extraite de [Gar05]. Notons v_n le membre de droite de (2.19) et u_n la projection de la fonction u sur l'espace Sparse

classique V_n^0. L'utilisation des sommes télescopiques permet de montrer que $u_n = v_n$.

$$\begin{aligned}
v_n &= \sum_{\ell_1+\ell_2=n+1} u_{\ell_1,\ell_2} - \sum_{\ell_1+\ell_2=n} u_{\ell_1,\ell_2} \\
&= \sum_{\ell_1+\ell_2\leq n+1} \hat{u}_{\ell_1,\ell_2} - \sum_{\ell_1+\ell_2\leq n} \hat{u}_{\ell_1,\ell_2} \\
&= \sum_{\ell_1\leq n+1} \sum_{k_1\leq \ell_1} \sum_{k_2\leq n+1-\ell_1} \hat{u}_{k_1,k_2} - \sum_{\ell_1\leq n} \sum_{k_1\leq \ell_1} \sum_{k_2\leq n-\ell_1} \hat{u}_{k_1,k_2} \\
&= \sum_{k_1\leq n+1} \sum_{k_2=0} \hat{u}_{k_1,k_2} + \sum_{\ell_1\leq n} \sum_{k_1\leq \ell_1} \left(\sum_{k_2\leq n+1-\ell_1} \hat{u}_{k_1,k_2} - \sum_{k_2\leq n-\ell_1} \hat{u}_{k_1,k_2} \right) \\
&= \sum_{k_1\leq n+1} \hat{u}_{k_1,0} + \sum_{\ell_1\leq n} \sum_{k_1\leq \ell_1} \sum_{k_2=n+1-\ell_1} \hat{u}_{k_1,k_2} \\
&= \sum_{k_1\leq n+1} \hat{u}_{k_1,0} + \sum_{n+1-k_2\leq n} \sum_{k_1\leq n+1-k_2} \hat{u}_{k_1,k_2} \\
&= \sum_{k_1\leq n+1} \hat{u}_{k_1,0} + \sum_{1\leq k_2} \sum_{k_1+k_2\leq n+1} \hat{u}_{k_1,k_2} \\
&= \sum_{k_1+k_2\leq n+1} \hat{u}_{k_1,k_2} \quad = u_n.
\end{aligned} \tag{2.20}$$

Considérons à présent la démonstration dans le cas général. A nouveau, exprimons la relation (2.19) sur la base hiérarchique, puis utilisons les relations sur les sommes télescopiques pour montrer l'équivalence entre la représentation (2.19) et la projection de u sur l'espace V_n^0.

L'idée, proposée par Koster [Kos00], consiste à introduire une fonction de pondération f telle que :

$$v_n = \sum_{\boldsymbol{\ell}:|\boldsymbol{\ell}|_1\leq n+d-1} f(\boldsymbol{\ell})\, u_{\boldsymbol{\ell}}, \quad \text{avec } \sum_{\boldsymbol{k}\geq\boldsymbol{\ell}} f(\boldsymbol{k}) = 1 \quad \text{et } f(\boldsymbol{\ell}) = 0 \text{ si } |\boldsymbol{\ell}|_1 < n. \tag{2.21}$$

Supposons qu'une telle fonction f existe. En introduisant la relation de passage de la base nodale à la base hiérarchique dans l'équation (2.21), nous obtenons l'égalité entre la combinaison linéaire v_n et la projection u_n de u sur l'espace V_n^0 :

$$\begin{aligned}
v_n &= \sum_{|\boldsymbol{\ell}|_1=n}^{n+d-1} f(\boldsymbol{\ell}) \sum_{\boldsymbol{k}\leq\boldsymbol{\ell}} \hat{u}_{\boldsymbol{k}} \\
&= \sum_{\boldsymbol{k}\leq n+d-1} \hat{u}_{\boldsymbol{k}} \sum_{\boldsymbol{\ell}\geq\boldsymbol{k}} f(\boldsymbol{\ell}) \\
&= \sum_{\boldsymbol{k}\leq n+d-1} \hat{u}_{\boldsymbol{k}} \quad = u_n.
\end{aligned} \tag{2.22}$$

A présent, nous identifions la fonction de pondération f qui apparaît dans l'équation (2.19) et nous montrons qu'elle vérifie la propriété $\sum_{\boldsymbol{k}\geq\boldsymbol{\ell}} f(\boldsymbol{k}) = 1$.

$$\begin{aligned} v_n &= \sum_{j=0}^{d-1} (-1)^{d-1-j} \binom{d-1}{j} \sum_{|\boldsymbol{\ell}|_1=n+j} u_{\boldsymbol{\ell}} \\ &= \sum_{|\boldsymbol{\ell}|_1=n}^{n+d-1} u_{\boldsymbol{\ell}} (-1)^{n+d-1-|\boldsymbol{\ell}|_1} \binom{d-1}{|\boldsymbol{\ell}|_1 - n}. \end{aligned} \tag{2.23}$$

Proposition 2.3 *La fonction de pondération f est donnée par*

$$f(\boldsymbol{\ell}) = (-1)^{n+d-1-|\boldsymbol{\ell}|_1} \binom{d-1}{|\boldsymbol{\ell}|_1 - n} = f(|\boldsymbol{\ell}|_1) \quad \textit{avec } f(\boldsymbol{\ell}) = 0 \textit{ si } |\boldsymbol{\ell}|_1 \notin [n, n+d-1].$$

Elle vérifie

$$\sum_{\boldsymbol{k}\geq\boldsymbol{\ell}} f(\boldsymbol{k}) = 1, \quad \forall \boldsymbol{\ell},\ |\boldsymbol{\ell}|_1 \leq n+d-1. \tag{2.24}$$

Preuve Nous aurons besoin des deux lemmes suivants :

Lemme 2.4 *Le nombre de multi-indices $\boldsymbol{k}$ de taille d tels que $\boldsymbol{k} \geq \boldsymbol{l}$ ($|\boldsymbol{k}|_1 \leq n+d-1$) est donné par $\binom{p+d-1}{p}$ avec $p = |\boldsymbol{k}|_1 - |\boldsymbol{\ell}|_1$.*

Lemme 2.5 *Soit $\mathcal{B}_{m,r}$ définie par*

$$\mathcal{B}_{m,r} = \sum_{p=0}^{m} \binom{m}{p} \binom{r+m-p}{r-p} (-1)^p, \tag{2.25}$$

alors

$$\mathcal{B}_{m,r} = 1, \quad \textit{pour tout } (m,r) \in \mathbb{N}^2.$$

Soit $\ell = |\boldsymbol{\ell}|_1$,

$$\begin{aligned} \sum_{\boldsymbol{k}\geq\boldsymbol{\ell}} f(\boldsymbol{k}) &= \sum_{m=n}^{n+d-1} f(m) \sum_{|\boldsymbol{k}|_1=m} 1_{\boldsymbol{k}\geq\boldsymbol{\ell}} \\ &= \sum_{m=n}^{n+d-1} (-1)^{n+d-1-m} \binom{d-1}{m-n} \binom{m+d-1-\ell}{m-\ell} \\ &= \sum_{p=0}^{d-1} (-1)^{d-1-p} \binom{d-1}{p} \binom{n+d-1-\ell+p}{n-\ell+p} \\ . &= \mathcal{B}_{d-1,n+d-1-\ell} \\ &= 1. \end{aligned} \tag{2.26}$$

Ceci conclut la démonstration de la proposition 2.3. ■

Démonstration du lemme 2.4 Nous aurons besoin du résultat intermédiaire suivant :

$$\sum_{\boldsymbol{\ell}:|\boldsymbol{\ell}|_1=j} 1 = \binom{d-1+j}{j}. \tag{2.27}$$

La démonstration de (2.27) s'obtient par une récurrence sur la dimension à partir des deux arguments :

1. Le passage à la dimension d sur le terme de gauche de (2.27) :

$$\sum_{\boldsymbol{\ell}:|\boldsymbol{\ell}|_1=j} 1 = \sum_{m=0}^{j} \sum_{\boldsymbol{\ell}:|\boldsymbol{\ell}|_1^{d-1}=m} 1, \quad \text{où } |\boldsymbol{\ell}|_1^{d-1} = \sum_{k=1}^{d-1} \ell_k. \tag{2.28}$$

2. Le passage à la dimension d sur le terme de droite de (2.27) :

$$\binom{d-1+j}{j} = \sum_{m=0}^{j} \binom{d-2+m}{m}. \tag{2.29}$$

La relation (2.29) est obtenue par application successive du triangle de Pascal :

$$\begin{aligned}
\binom{d-1+j}{j} &= \binom{d-2+j}{j} + \binom{d-2+j}{j-1} \qquad (2.30)\\
\binom{d-1+j}{j} &= \binom{d-2+j}{j} + \binom{d-3+j}{j-1} + \binom{d-3+j}{j-2}\\
\binom{d-1+j}{j} &= \binom{d-2+j}{j} + \binom{d-2+j-1}{j-1} + \binom{d-2+j-2}{j-2} + \binom{d-4+j}{j-3} \cdots.
\end{aligned}$$

Le lemme 2.4 se déduit de (2.27) :

$$\sum_{\substack{\boldsymbol{k} \geq \boldsymbol{\ell} \\ |\boldsymbol{k}|_1 = j}} 1 = \sum_{\substack{\boldsymbol{k} - \boldsymbol{\ell} \geq \mathbf{0} \\ |\boldsymbol{k} - \boldsymbol{\ell}|_1 = p}} 1 = \sum_{\boldsymbol{p}:|\boldsymbol{p}|_1=p} 1 = \binom{d-1+p}{p}. \tag{2.31}$$

■

Démonstration du lemme 2.5 Remarquons que $\mathcal{B}_{0,r} = 1$ pour tout $r \geq 0$ et $\mathcal{B}_{m,0} = 1$ pour tout $m \geq 0$. Le résultat est une conséquence triviale de la relation de récurrence suivante

$$\mathcal{B}_{m,r} = \mathcal{B}_{m,r-1} + \mathcal{B}_{m-1,r} - \mathcal{B}_{m-1,r-1}. \tag{2.32}$$

Cette relation est obtenue par applications successives de la formule de Pascal : une première fois à $\binom{r-p}{r+m-p}$ puis à $\binom{m}{p}$.

$$\begin{aligned}
\mathcal{B}_{m,r} &= \sum_{p=0}^{m} \binom{m}{p} \binom{r+m-p}{r-p} (-1)^p \\
&= \sum_{p=0}^{m} \binom{m}{p} \binom{r-1+m-p}{r-1-p} (-1)^p + \sum_{p=0}^{m} \binom{m}{p} \binom{r-1+m-p}{r-p} (-1)^p \\
&= \mathcal{B}_{m,r-1} + \sum_{p=0}^{m} \binom{m-1}{p-1} \binom{r-1+m-p}{r-p} (-1)^p + \sum_{p=0}^{m} \binom{m-1}{p} \binom{r-1+m-p}{r-p} (-1)^p \\
&= \mathcal{B}_{m,r-1} - \sum_{p=0}^{m} \binom{m-1}{p-1} \binom{r-1+m-1-(p-1)}{r-1-(p-1)} (-1)^{p-1} + \mathcal{B}_{m-1,r} \\
&= \mathcal{B}_{m,r-1} - \sum_{p=0}^{m-1} \binom{m-1}{p} \binom{r-1+m-1-p}{r-1-p} (-1)^p + \mathcal{B}_{m-1,r} \\
&= \mathcal{B}_{m,r-1} - \mathcal{B}_{m-1,r-1} + \mathcal{B}_{m-1,r} \qquad (2.33)
\end{aligned}$$

■

■

2.2.2 Résolution de problèmes aux limites

La méthode est introduite par Griebel et al [GSZ92]. Elle consiste à décomposer la solution en une somme pondérée de contributions issues de grilles cartésiennes (voir la figure 2.4). Des approximations $u_{\boldsymbol{\ell}}$ de la solution sont calculées sur ces grilles de taille relativement petite. L'approximation sparse u_n est obtenue en recombinant les approximations suivant la relation (2.19). La résolution sur chacune des grilles $\Omega_{\boldsymbol{\ell}}$ s'effectue en parallèle : elle peut être obtenue par un schéma de discrétisation quelconque.

A partir du Théorème 2.2 et de la transformée de Fourrier, Griebel [GSZ92] propose un résultat de convergence dans le cas de l'équation de Poisson en dimension 2. Reisinger [Rei04, RW07] généralise le résultat à des schémas d'ordres supérieurs et à des problèmes plus généraux (équation de Poisson et équation d'advection). Cette méthode est basée sur un « développement de l'erreur en puissance du pas de maillage » obtenu sur les différentes grilles. La convergence de la méthode repose sur l'hypothèse suivante :

Hypothèse 2.1 *Nous supposons que, sur les points de la grille $\Omega_{\boldsymbol{\ell}}$ (de pas $\left(2^{-\ell_1}, \ldots, 2^{-\ell_d}\right)$), l'erreur commise sur l'approximation numérique par une méthode d'ordre p suit le développement :*

$$u\left(\boldsymbol{x}_{\boldsymbol{\ell}}\right) - u_{\boldsymbol{\ell}} = \sum_{m=1}^{d} \sum_{\substack{\{k_1, \ldots, k_m\} \\ \subset \{1, \ldots, d\} \\ k_i \neq k_j}} \gamma_{k_1,\ldots,k_m} h_{\ell_{k_1}}^{p} \ldots h_{\ell_{k_m}}^{p}, \qquad (2.34)$$

où $\gamma_{k_1,\ldots,k_m}$ dépend de u.

Remarque 2.5 *Reisinger [RW07] démontre ce résultat (2.34) pour l'équation de Poisson et pour une équation d'advection. Le résultat est établi pour un schéma de différences finies d'ordre p sur des grilles $\Omega_{\boldsymbol{\ell}}$.*

Théorème 2.6 (Convergence de la méthode de technique combinatoire) *Soient $u_{\boldsymbol{\ell}}$ les solutions approchées sur les grilles $\Omega_{\boldsymbol{\ell}}$. Nous supposons vérifiée l'hypothèse 2.1, où les constantes $\gamma_{k_1,\ldots,k_m}$ de l'équation (2.34) vérifient :*

$$|\gamma_{k_1,\ldots,k_m}| \leq K, \quad \forall 1 \leq m \leq d, \quad \forall \{k_1,\ldots,k_m\} \subset \{1,\ldots,d\}, \quad k_i \neq k_j. \tag{2.35}$$

Alors l'approximation u_n obtenue à partir de la relation (2.19) converge vers la solution u et il existe $K > 0$ qui dépend de u,

$$|u - u_n| \leq \frac{2K}{(d-1)!} \left(\frac{2^p+1}{2^{p-1}}\right)^{d-1} (n+2(d-1))^{d-1}\, 2^{-pn}. \tag{2.36}$$

Remarque 2.6 *Si $d = 1$, nous obtenons le résultat classique $|u - u_n| \leq 2K2^{-pn}$.*

Preuve Nous ne donnons que les points essentiels de la démonstration, voir [RW07] pour la preuve explicite.

La démonstration repose sur le lemme 2.5 et plus généralement sur l'équation (2.24),

$$u - u_n = \sum_{|\boldsymbol{\ell}|_1 = n}^{n+d-1} f(\boldsymbol{\ell})\,(u - u_{\boldsymbol{\ell}}), \tag{2.37}$$

D'après (2.3), $f(\boldsymbol{\ell})$ ne dépend que de $|\boldsymbol{\ell}|_1$, en notant $q = |\boldsymbol{\ell}|_1$. Le membre de gauche de l'équation (2.37) s'écrit comme une somme de termes S_q définie par $S_q = \sum_{|\boldsymbol{\ell}|_1 = q} u - u_{\boldsymbol{\ell}}$.

A partir de l'hypothèse 2.1, cette somme s'écrit :

$$
\begin{aligned}
S_q &= \sum_{|\boldsymbol{\ell}|_1=q} \sum_{m=1}^{d} \sum_{\substack{\{k_1,\dots,k_m\}\\ \subset \{1,\dots,d\}\\ k_i \neq k_j}} \gamma_{k_1,\dots,k_m} h^p_{\ell_{k_1}} \dots h^p_{\ell_{k_m}} \\
&= \sum_{|\boldsymbol{\ell}|_1=q} \sum_{m=1}^{d} \sum_{\substack{\{k_1,\dots,k_m\}\\ \subset \{1,\dots,d\}\\ k_i \neq k_j}} \gamma_{k_1,\dots,k_m} 2^{-p\sum_{k=1}^{m}\ell_k} \\
&\quad \text{fixons } \textstyle\sum_{k=1}^{m}\ell_k \text{ à } j, \text{puis sommons sur les combinaisons restantes} \\
&= \sum_{j=0}^{q} \sum_{m=1}^{d} \sum_{\substack{\{k_1,\dots,k_m\}\\ \subset \{1,\dots,d\}\\ k_i \neq k_j}} \gamma_{k_1,\dots,k_m} 2^{-pj} \sum_{\sum_{k=m+1}^{d}\ell_k = q-j} 1 \qquad (2.38)\\
&\quad \text{par application du lemme 2.4} \\
&= \sum_{j=0}^{q} \sum_{m=1}^{d} \sum_{\substack{\{k_1,\dots,k_m\}\\ \subset \{1,\dots,d\}\\ k_i \neq k_j}} \gamma_{k_1,\dots,k_m} 2^{-pj} \binom{q-j+d-m-1}{d-m-1} \\
&= \sum_{m=1}^{d} \sum_{\substack{\{k_1,\dots,k_m\}\\ \subset \{1,\dots,d\}\\ k_i \neq k_j}} \gamma_{k_1,\dots,k_m} \sum_{j=0}^{q} 2^{-pj} \binom{q-j+d-m-1}{d-m-1}
\end{aligned}
$$

En introduisant la fonction de pondération f puis en sommant les S_q, nous obtenons

$$
u - u_n = \sum_{m=1}^{d} \sum_{\substack{\{k_1,\dots,k_m\}\\ \subset \{1,\dots,d\}\\ k_i \neq k_j}} \gamma_{k_1,\dots,k_m} \sum_{|\boldsymbol{\ell}|_1=n}^{n+d-1} f(\boldsymbol{\ell})\, 2^{-p|\boldsymbol{\ell}|_1} \binom{n-|\boldsymbol{\ell}|_1+d-m-1}{d-m-1}. \qquad (2.39)
$$

La démonstration repose sur le résultat suivant, dont le lecteur trouvera une justification dans [RW07] : il existe une constante $K > 0$ (K dépendant de u) telle que :

$$
\begin{aligned}
&\gamma_{k_1,\dots,k_m} \sum_{|\boldsymbol{\ell}|_1=n}^{n+d-1} f(\boldsymbol{\ell})\, 2^{-p|\boldsymbol{\ell}|_1} \binom{n-|\boldsymbol{\ell}|_1+d-m-1}{d-m-1} \qquad (2.40)\\
&\qquad \leq 2^{-p(n+d-1)} \left(2^p+1\right)^{m-1} K \binom{n+d-m-2}{m-1} \\
&\qquad \leq 2^{-pn} \left(\frac{2^p+1}{2^p}\right)^{d-1} \frac{K}{(d-1)!} \left(n+2\left(d-1\right)\right)^{d-1}.
\end{aligned}
$$

Le terme de droite de l'équation (2.39) contient $2^d - 1$ termes, ce qui conclut la démonstration du théorème 2.6,

$$
|u - u_n|_1 \leq \left(\frac{2^p+1}{2^{p-1}}\right)^{d-1} \frac{K}{(d-1)!} \left(n+2\left(d-1\right)\right)^{d-1} 2^{-pn}. \qquad (2.41)
$$

■

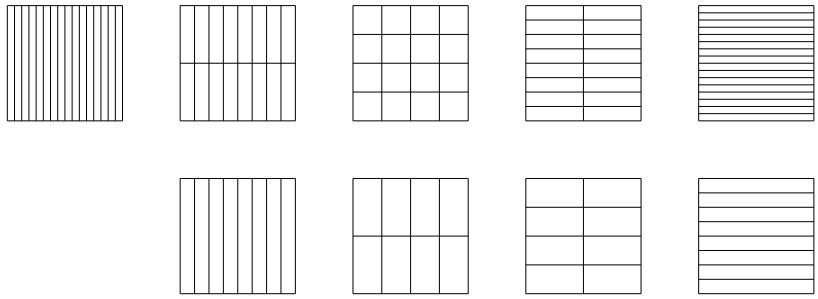

FIG. 2.4 – Combination technique avec $l = 4$ en dimension 2

Analyse de la complexité Ce paragraphe est consacré à l'étude de la complexité de cette méthode. Le nombre d'équations à résoudre sur chaque grille anisotrope, noté N, est obtenu à partir de l'équation (2.19).

$$N = \sum_{j=1}^{d-1} \sum_{|i|_1=n+j} 1 = \sum_{j=0}^{d} \binom{n+j+d-1}{n+j}. \tag{2.42}$$

nl/dim	2	3	4	5	6	7	8
6	15	109	589	2751	11914	49596	202203
7	17	136	791	3906	17640	75804	316767
8	19	166	1035	5396	25416	112848	483879
9	21	199	1325	7281	35757	164109	722601
10	23	235	1665	9626	49259	233717	1057265

TAB. 2.4 – Nombre d'EDP à résoudre dans la méthode de technique combinatoire, en fonction du niveau de raffinement(ligne) et de la dimension(colonne)

Exemple 2.1 *Considérons un problème aux limites de type advection diffusion en dimension $d = 4$. Le niveau de raffinement $n = 8$ est fixé. Le nombre d'équations à résoudre est donné par l'équation (2.42),*

$$N = \sum_{j=1}^{3} \frac{(11+j)!}{(4-(j+1))!\,(8+j)!j!} \approx 2000.$$

Cette deuxième étude propose d'évaluer le coût en terme d'opérations de cette méthode. Celle-ci est basée sur l'hypothèse suivante : la résolution du système linéaire s'effectue par une méthode directe de type LU sur chacune des grilles $\Omega_{\boldsymbol{\ell}}$. Le nombre de calcul pour la résolution du système linéaire est de l'ordre de $N\,L^2$, où N est le nombre d'inconnues et L la plus grande largeur de bande.

Dans le cas de la grille sparse caractérisée par l'espace V_n^0, $N = 2^{|\boldsymbol{\ell}|_1}$ et $L = 2^{|\boldsymbol{\ell}|_1 - |\boldsymbol{\ell}|_\infty}$.

Pour chacune des grilles, le coût de la résolution du système linéaire est de l'ordre de $M_{\boldsymbol{\ell}} = 2^{3|\boldsymbol{\ell}|_1 - 2|\boldsymbol{\ell}|_\infty}$.

Si N processeurs, où N est donné par (2.42), sont utilisés, le temps de calcul est de l'ordre de $\max_{\boldsymbol{\ell}:|\boldsymbol{\ell}|_1=n+d-1} M_{\boldsymbol{\ell}}$.

En sommant chacune des résolutions $M_{\boldsymbol{\ell}}$, nous obtenons la complexité, notée C, de cette méthode

$$C = \sum_{j=0}^{d-1} \sum_{\boldsymbol{\ell}:|\boldsymbol{\ell}|_1=n+j} M_{\boldsymbol{\ell}} = \sum_{j=0}^{d} 2^{3(n+j)} \sum_{\boldsymbol{\ell}:|\boldsymbol{\ell}|_1=n+j} 2^{-2|\boldsymbol{\ell}|_\infty}. \tag{2.43}$$

En utilisant le lemme 2.4 et $|\boldsymbol{\ell}|_\infty \geq |\boldsymbol{\ell}|_1/d$,

$$C \leq \sum_{j=0}^{d-1} 2^{3(n+j)} \sum_{\boldsymbol{\ell}:|\boldsymbol{\ell}|_1=n+j} 2^{-2\frac{n+j}{d}} = \sum_{j=0}^{d} 2^{(n+j)\left(3-\frac{2}{d}\right)} \binom{n+j+d-1}{n+j}. \tag{2.44}$$

Conclusion et perspectives La méthode propose une alternative intéressante aux méthodes classiques de parallélisation d'une EDP en dimensions 2 et 3. Pour les dimensions supérieures, le nombre élevé de sous-problèmes est un frein à l'utilisation de ces méthodes. Reisinger ([Rei04] page 90-92) propose d'utiliser des schémas d'ordre plus élevé afin de réduire le niveau n. Cependant, l'existence de tels schémas dépend, bien entendu, du problème aux limites considéré.

2.3 Méthode de différences finies

La méthode des différences finies joue un rôle important en analyse numérique. Elle constitue la méthode la plus simple pour approcher un opérateur différentiel et, par extension, pour résoudre numériquement une EDP. Une première application concerne la résolution approchée des *équations différentielles ordinaires* avec, par exemple, le schéma *d'Euler*. La méthode de différences finies sur une *Sparse Grid* est introduite par Griebel [Gri98].

Les deux premiers paragraphes permettent d'introduire quelques notations. Nous revenons sur la définition d'un problème aux limites elliptique, puis sur sa résolution numérique par la méthode des différences finies.

Nous donnons, ensuite, une description de l'opérateur discret sur une *Sparse Grid*et le résultat de consistance obtenu pour cet opérateur. Ce résultat a été d'abord obtenu par Schiekofer [Sch98a], puis par Koster [Kos00] avec des hypothèses plus générales sur les schémas et les grilles considérés. Nous présentons une seconde démonstration de ce résultat à partir des propriétés des ondelettes interpolantes.

2.3.1 Discrétisation d'opérateurs elliptiques

Considérons un problème aux limites sur un domaine borné de la forme $\prod_{k=1}^{d} (A_k, B_k)$. Afin de simplifier l'exposé, nous nous ramenons à l'hypercube $(0,1)^d$ par un changement de variables affines.

Sur ce domaine, l'opérateur linéaire $\mathcal{L}$ est défini :

$$\mathcal{L} = -\sum_{i=1}^{d}\sum_{j=1}^{d} a_{i,j}(\boldsymbol{x})\frac{\partial}{\partial x_i}\frac{\partial}{\partial x_j} + \sum_{i=1}^{d} b_i(\boldsymbol{x})\frac{\partial}{\partial x_i} + c(\boldsymbol{x}). \tag{2.45}$$

Remarque 2.7 *Les coefficients $a_{i,j}$, b_i et c, $1 \leq i,j \leq d$ de l'opérateur $\mathcal{L}$ seront dits tensoriels si ils admettent la représentation*

$$\alpha\left(x_1,\ldots,x_d\right) = \alpha^1\left(x_1\right)\ldots\alpha^d\left(x_d\right). \tag{2.46}$$

Un cas particulier intéressant consiste à considérer les opérateurs $\mathcal{L}$ tels que

$$a_{i,j}\left(\boldsymbol{x}\right) = a^i\left(x_i\right)a^j\left(x_j\right), \quad b_i\left(\boldsymbol{x}\right) = b^i\left(x_i\right). \tag{2.47}$$

2.3.1.1 Opérateur de différences finies sur $(0,1)$

Nous considérons des schémas linéaires aux différences finies, pour lesquels les opérateurs discrets peuvent être représentés comme un produit matrice vecteur. Ceci inclut les schémas usuels d'ordre M quelconque, mais exclut par exemple les schémas ENO.

Les principaux schémas de différences finies sont :

– Pour l'opérateur $\dfrac{\partial}{\partial x}$:

– Schéma centré d'ordre $M = 2$:

$$\frac{\partial u}{\partial x}(x = x_i) \approx \left(D_\ell^1 u\right)_i = \frac{u\left(x_i + 2^{-\ell}\right) - u\left(x_i - 2^{-\ell}\right)}{2 \cdot 2^{-\ell}}. \tag{2.48}$$

– Schéma décentré à gauche d'ordre $M = 1$:

$$\frac{\partial u}{\partial x}(x = x_i) \approx \left(D_\ell^- u\right)_i = \frac{u(x_i) - u\left(x_i - 2^{-\ell}\right)}{2^{-\ell}}. \tag{2.49}$$

– Schéma décentré à droite d'ordre $M = 1$:

$$\frac{\partial u}{\partial x}(x = x_i) \approx \left(D_\ell^+ u\right)_i = \frac{u\left(x_i + 2^{-\ell}\right) - u(x_i)}{2^{-\ell}}. \tag{2.50}$$

– Opérateur de diffusion, schéma centré d'ordre $M = 2$:

$$\frac{\partial^2 u}{\partial x^2}(x = x_i) \approx \left(D_\ell^2 u\right)_i = \frac{u\left(x_i + 2^{-\ell}\right) - 2u(x_i) + u\left(x_i - 2^{-\ell}\right)}{2^{-2\ell}}. \tag{2.51}$$

Quelques notations sont introduites afin de décrire les opérateurs de différences finies sur une base d'ondelettes et les erreurs de consistance associées.

1. Soit $U^\ell = (u^{\ell,1}, \ldots, u^{\ell,2^\ell-1})^T \in \mathbb{R}^{2^\ell-1}$ le vecteur des coefficients de la projection de la fonction u sur la base des fonctions d'échelle de niveau ℓ. Nous introduisons également $U_\ell = (u_{\ell,2i+1})_{i=1,\ldots,2^{\ell-2}} \in \mathbb{R}^{2^{\ell-1}}$, le vecteur de la projection sur la base d'ondelettes de niveau ℓ.
2. Soit N l'ordre d'interpolation de l'ondelette ($N = 2$ dans le cas de la base obtenue par translation et dilatation de la fonction chapeau).
3. Soit T_ℓ l'opérateur de passage de la base des fonctions d'échelle à la base d'ondelettes. Autrement dit, il s'agit de l'opérateur de passage de la base nodale à la base hiérarchique.
$$\mathbb{R}^{2^{\ell-1}} \to \mathbb{R}^{2^{\ell-1}}, \quad T_\ell U^\ell = \left(U_1^T, \ldots, U_\ell^T\right)^T. \tag{2.52}$$
4. Soient P_ℓ l'opérateur de restriction sur l'espace discret :
$$P_\ell : \mathcal{C}^0([0,1]) \to \mathbb{R}^{2^\ell-1}, \quad P_\ell u = U^\ell, \tag{2.53}$$
et I^ℓ l'opérateur d'interpolation :
$$I_\ell : \mathbb{R}^{2^\ell-1} \to \mathcal{C}^0([0,1]), \quad I_\ell U = \sum_{i=1}^{2^\ell-1} u^{\ell,i} \phi_{\ell,i}. \tag{2.54}$$
L'erreur de projection $E_P(\cdot, \ell) : \mathcal{C}^0([0,1]) \to \mathcal{C}^0([0,1])$ sur les fonctions de base est définie par :
$$E_P(u, \ell) = u - I_\ell \circ P_\ell u. \tag{2.55}$$
5. Soit $D_\ell : \mathbb{R}^{2^\ell-1} \to \mathbb{R}^{2^\ell-1}$, la matrice associée au schéma de différences finies d'ordre M. Pour approcher l'opérateur ∂_{xx}, nous considérons l'opérateur D_ℓ^2 défini en (2.51). Soit $E_D(\cdot, \ell) : \mathcal{C}^2([0,1]) \to \mathbb{R}^{2^\ell-1}$, l'erreur de consistance sur les noeuds de la grille,
$$E_D(u, \ell) = P_\ell(\partial_{xx} u) - D_\ell \circ P_\ell(u). \tag{2.56}$$

Proposition 2.7 *Considérons les deux opérateurs $E_D(\cdot,\ell)$ et $E_P(\cdot,\ell)$.*
- *Ces applications sont linéaires par rapport au premier argument.*
- *Si $u \in \mathcal{C}^\alpha, \alpha > 2$ et si le schéma de différences finies associé à la matrice D_ℓ est consistant à l'ordre M, alors*

$$\|E_D(u,\ell)\|_\infty \leq C 2^{-min(M,\alpha-2)\ell} |u|_{min(M+2,\alpha)}. \tag{2.57}$$

- *Si $u \in \mathcal{C}^\alpha, \alpha > 0$ et si les fonctions de base sont des ondelettes d'ordre N, alors*

$$\|E_P(u,\ell)\|_\infty \leq C 2^{-min(N,\alpha)\ell} |u|_{min(N,\alpha)}. \tag{2.58}$$

Preuve La démonstration de l'équation (2.57) est basée sur des développements de Taylor de u aux points x_{s+i} et sur une propriété de décroissance des coefficients extra-diagonaux de la matrice D_ℓ associée au schéma. Cette propriété est en pratique toujours vérifiée pour des opérateurs différentiels et des opérateurs intégraux dont le noyau est suffisamment décroissant. La démonstration de (2.58) compare le développement de Taylor (reste intégral) de la fonction u évaluée en x_s à celui de la fonction approchée u_ℓ. Le lecteur trouvera dans [Kos00] les détails de cette démonstration. ■

2.3.1.2 Schéma de discrétisation sur une grille sparse

Après avoir défini l'opérateur de différences finies sparse pour la discrétisation de $\dfrac{\partial^2}{\partial x_1^2}$, nous donnerons la discrétisation des autres opérateurs différentiels.

Ordonnancement des coefficients de la projection d'une fonction u sur une *Sparse Grid*. Soit Ω_n la *Sparse Grid* de niveau n,

$$\Omega_n = \bigcup_{\boldsymbol{\ell}:|\boldsymbol{\ell}|_1 \leq n+d-1} \Omega_{\boldsymbol{\ell}}.$$

Une fonction de grille u sur Ω_n est représentée par un vecteur U. D'après la Proposition 1.18, la fonction u peut s'écrire sur la base d'ondelettes,

$$u = \sum_{\overline{\boldsymbol{\ell}} \leq \boldsymbol{\ell}, \boldsymbol{\imath} \in \tau_{\overline{\boldsymbol{\ell}}}} u_{\overline{\boldsymbol{\ell}},\boldsymbol{\imath}} \psi_{\overline{\boldsymbol{\ell}},\boldsymbol{\imath}}.$$

Notons $U_{\boldsymbol{\ell}}$ le vecteur $(u_{\boldsymbol{\ell},\boldsymbol{\imath}})_{\boldsymbol{\imath}\in\boldsymbol{\tau}_{\boldsymbol{\ell}}}$ où $\boldsymbol{\tau}_{\boldsymbol{\ell}}$ est donné à la définition 1.14. L'ordre initial est donné par $U = \left(U_{\boldsymbol{\ell}}^T\right)_{\boldsymbol{\ell}:|\boldsymbol{\ell}|_1 \leq n+d-1}^T$ où les $\boldsymbol{\ell}$ sont rangés dans l'ordre lexicographique.

Nous avons besoin de définir une application $\Xi_{(k)}$ de $\mathbb{R}^{\#(\Omega_n)}$ dans $\mathbb{R}^{\#(\Omega_n)}$ qui permute les coefficients du vecteur U. L'application dans le cas $k = 1$ est, tout d'abord, définie.

Définition 2.2 (La permutation $\Xi_{(1)}$) *Soient $\boldsymbol{\ell_2} = (\ell_2, \ldots, \ldots \ell_d) \in \mathbb{N}^{d-1}$ (*resp.*$\boldsymbol{\imath_2} = (i_2, \ldots, i_d) \in \boldsymbol{\tau}_{\boldsymbol{\ell_2}}$) les $d-1$ dernières composantes d'un multi-indice $\boldsymbol{\ell}$ (*resp.*$\boldsymbol{\imath}$). Nous fixons un couple d'indices $\boldsymbol{\ell_2}, \boldsymbol{\imath_2}$ ce qui signifie que nous considérons les points de la grille Ω_n*

appartenant à la droite parallèle à e_1 *et passant par le point* $\boldsymbol{x}_{\boldsymbol{\ell},\boldsymbol{i}}$. *Soient* $\Omega_{|\boldsymbol{\ell_2},\boldsymbol{i_2}}$ *ce sous-ensemble et* $h_{\boldsymbol{\ell_2},\boldsymbol{i_2}}$ *la distance entre deux points voisins de ce sous-ensemble. Le niveau de raffinement maximum, noté* $\widetilde{\ell}$, *dans la dimension* 1, i.e. $h_{\boldsymbol{\ell_2},\boldsymbol{i_2}} = 2^{-\tilde{\ell}}$ *est donné par*

$$\widetilde{\ell} = n + d - 1 - |\boldsymbol{\ell_2}|_1. \tag{2.59}$$

Soit $U_{\boldsymbol{\ell_2},\boldsymbol{i_2}}$ *le vecteur des valeurs de la fonction de grille restreinte à* $\Omega_{|\boldsymbol{\ell_2},\boldsymbol{i_2}}$. *Nous ordonnons les coefficients de* $U_{\boldsymbol{\ell_2},\boldsymbol{i_2}}$ *suivant le niveau* ℓ_1 *:*

$$\begin{aligned} U_{\boldsymbol{\ell_2},\boldsymbol{i_2}} &= \begin{pmatrix} U_{(1,\boldsymbol{\ell_2},\boldsymbol{i_2})} \\ \vdots \\ U_{(\widetilde{\ell},\boldsymbol{\ell_2},\boldsymbol{i_2})} \end{pmatrix} \\ \text{avec} \quad U_{(h,\boldsymbol{\ell_2},\boldsymbol{i_2})} &= \left(u_{(h,\boldsymbol{\ell_2}),(j,\boldsymbol{i_2})}\right)^T_{\{j \text{ impaire}, 1\le j\le 2^h-1\}}. \end{aligned} \tag{2.60}$$

Alors, la permutation $\Xi_{(1)}$ *est donnée par*

$$\Xi_{(1)}(U) = \left(U_{\boldsymbol{\ell_2},\boldsymbol{i_2}}\right)_{\boldsymbol{\ell_2}:|\boldsymbol{\ell_2}|_1\le n+d-2,\boldsymbol{i_2}\in\tau_{\boldsymbol{\ell_2}}}, \tag{2.61}$$

où les $(\boldsymbol{\ell_2},\boldsymbol{i_2})$ *sont rangés dans l'ordre lexicographique.*

Définition 2.3 (La permutation $\Xi_{(k)}$) *La permutation* $\Xi_{(k)}$ *est obtenue en échangeant les indices correspondant aux dimensions* 1 *et* k, *en appliquant* $\Xi_{(1)}$ *puis en effectuant à nouveau l'échange d'indices.*

Les opérations de changement de base (2.52) et de discrétisation $D_{\tilde{\ell}}$ s'appliquent à chacune des composantes $U_{\boldsymbol{\ell_2},\boldsymbol{i_2}}$ de $\Xi_{(1)}(U)$. Par exemple, la discrétisation de $\frac{\partial^2}{\partial x_1^2}$ sur une *Sparse Grid* est donnée par

$$\begin{aligned} (U_{\boldsymbol{\ell}})_{\mathbf{1}\le\boldsymbol{\ell},|\boldsymbol{\ell}|_1\le n+d-1} &\mapsto (V_{\boldsymbol{\ell}})_{\mathbf{1}\le\boldsymbol{\ell},|\boldsymbol{\ell}|_1\le n+d-1}, \\ \text{où } (V_{\boldsymbol{\ell}})_{1\le|\boldsymbol{\ell}|_1\le n+d-1} \text{ est tel que} & \qquad , \\ V_{\boldsymbol{\ell_2},\boldsymbol{i_2}} &= T_{\tilde{\ell}}\, D^2_{\tilde{\ell}}\, T^{-1}_{\tilde{\ell}}\, U_{\boldsymbol{\ell_2},\boldsymbol{i_2}}, \quad \forall\boldsymbol{\ell}, \quad \forall\boldsymbol{i}\in\boldsymbol{\tau_\ell}, \end{aligned} \tag{2.62}$$

où $D^2_{\tilde{\ell}}$ est donnée par (2.51) et $T_{\tilde{\ell}}$ par (2.52).

Afin de généraliser à d'autres opérateurs, nous aurons besoin des opérateurs

$$\mathbb{T}_{(1)} : \mathbb{R}^{\#(\Omega_n)} \to \mathbb{R}^{\#(\Omega_n)}, \quad \left(U_{\boldsymbol{\ell_2},\boldsymbol{i_2}}\right)_{\boldsymbol{\ell_2}:|\boldsymbol{\ell_2}|_1\le n+d-2,\boldsymbol{i_2}\in\tau_{\boldsymbol{\ell_2}}} \to \left(T_{\tilde{\ell}} U_{\boldsymbol{\ell_2},\boldsymbol{i_2}}\right)_{\boldsymbol{\ell_2}:|\boldsymbol{\ell_2}|_1\le n+d-2,\boldsymbol{i_2}\in\tau_{\boldsymbol{\ell_2}}}. \tag{2.63}$$

$$\mathbb{D}_{(1)} : \mathbb{R}^{\#(\Omega_n)} \to \mathbb{R}^{\#(\Omega_n)}, \quad \left(U_{\boldsymbol{\ell_2},\boldsymbol{i_2}}\right)_{\boldsymbol{\ell_2}:|\boldsymbol{\ell_2}|_1\le n+d-2,\boldsymbol{i_2}\in\tau_{\boldsymbol{\ell_2}}} \to \left(D_{\tilde{\ell}} U_{\boldsymbol{\ell_2},\boldsymbol{i_2}}\right)_{\boldsymbol{\ell_2}:|\boldsymbol{\ell_2}|_1\le n+d-2,\boldsymbol{i_2}\in\tau_{\boldsymbol{\ell_2}}}, \tag{2.64}$$

où $D_{\tilde{\ell}}$ est la matrice associée au schéma aux différences finies (2.48,...,2.51).

Les applications $\mathbb{T}_{(k)}$ et $\mathbb{D}_{(k)}$ se déduisent de $\mathbb{T}_{(1)}$ et $\mathbb{D}_{(1)}$ suivant la méthode appliquée à $\Xi_{(k)}$ dans la définition 2.3.

Différents opérateurs discrets sur une grille sparse Les opérateurs introduits précédemment nous permettent de définir

1. le changement de base de nodale à hiérarchique dans la dimension k,

$$\mathrm{T}_{(k)} : \mathbb{R}^{\#(\Omega_n)} \to \mathbb{R}^{\#(\Omega_n)}, \quad U \to \Xi_{(k)}^{-1}\mathbb{T}_{(k)}\Xi_{(k)}U \tag{2.65}$$

2. le schéma aux différences finies

$$\mathrm{D}_{(k)} : \mathbb{R}^{\#(\Omega_n)} \to \mathbb{R}^{\#(\Omega_n)}, \quad U \to \Xi_{(k)}^{-1}\mathbb{D}_{(k)}\Xi_{(k)}U \tag{2.66}$$

3. $\mathrm{T}_{(k)}^{-1}$ la transformation qui permet de passer de la base hiérarchique à la base nodale dans la dimension k.

Remarque 2.8 *Les opérateurs* $\mathrm{T}_{(i)}$ *et* $\mathrm{T}_{(j)}$ *commutent :*

$$\mathrm{T}_{(i)}\mathrm{T}_{(j)} = \mathrm{T}_{(j)}\mathrm{T}_{(i)}. \tag{2.67}$$

Considérons $\mathrm{D}_{(i)}$ *un opérateur de différences finies dans la direction* i*. Cet opérateur* $\mathrm{D}_{(i)}$ *ne commute pas avec* $\mathrm{T}_{(j)}$*,*

$$\mathrm{T}_{(j)}\mathrm{D}_{(i)} \neq \mathrm{D}_{(i)}\mathrm{T}_{(j)}, \quad \textit{si } i \neq j. \tag{2.68}$$

En reprenant (2.62), l'opérateur de différences finies $\mathrm{D}_{(i)}$ sur une *Sparse Grid* est obtenu par la composition des opérations suivantes :

1. Appliquer la transformation hiérarchique → nodale dans la dimension i.
2. Appliquer le schéma aux différences finies dans la dimension i.
3. Appliquer la transformation nodale → hiérarchique dans la dimension i.

Cette méthode est appliquée aux opérateurs décrits dans § 2.3.1.1. Nous supposons que les coefficients $a_{i,j}$ et b_i de l'opérateur $\mathcal{L}$ vérifient l'équation (2.46). Donnons quelques exemples :

– Opérateurs de diffusion :
  – Opérateur de Laplace :

$$\frac{\partial^2}{\partial x_i^2} \approx \mathrm{T}_{(i)} \circ \mathrm{D}_{(i)}^2 \circ \mathrm{T}_{(i)}^{-1}, \quad \Delta \approx \widehat{\Delta} = \sum_{i=1}^{d} \mathrm{T}_{(i)} \circ \mathrm{D}_{(i)}^2 \circ \mathrm{T}_{(i)}^{-1}, \tag{2.69}$$

  où $\mathrm{D}_{(i)}^2$ est associée à un schéma de différences finies pour l'opérateur $\dfrac{\partial^2}{\partial x^2}$, par exemple (2.51).
  – Dérivées mixtes :

$$\frac{\partial^2}{\partial x_i \partial x_j} \approx \mathrm{T}_{(i)} \circ \mathrm{D}_{(i)} \circ \mathrm{T}_{(i)}^{-1} \circ \mathrm{T}_{(j)} \circ \mathrm{D}_{(j)} \circ \mathrm{T}_{(j)}^{-1}. \tag{2.70}$$

  où $\mathrm{D}_{(i)}$ est associée à un schéma de différences finies pour l'opérateur $\dfrac{\partial}{\partial x}$, par exemple (2.48).

– Opérateur $\dfrac{\partial}{\partial x_j} a_i(x_i) a_j(x_j) \dfrac{\partial}{\partial x_i}$:

$$\frac{\partial}{\partial x_i} a_i(x_i) a_j(x_j) \frac{\partial}{\partial x_j} \approx \mathrm{T}_{(i)} \circ \left(a_i \mathrm{D}_{(i)}\right) \circ \mathrm{T}_{(i)}^{-1} \circ \mathrm{T}_{(j)} \circ \left(a_j \mathrm{D}_{(j)}\right) \circ \mathrm{T}_{(j)}^{-1}, \tag{2.71}$$

où $a_i \mathrm{D}_{(i)}$ est associée à un schéma de différences finies pour l'opérateur $a_i(x) \dfrac{\partial}{\partial x}$, par exemple (2.48) .

– Opérateur de convection :
  – avec une vitesse $b = b_i\,(x_i)\, e_i$,

$$b_i(x_i) \frac{\partial}{\partial x_i} \approx \mathrm{T}_{(i)} \circ \left(b_i\, \mathrm{D}_{(i)}\right) \circ \mathrm{T}_{(i)}^{-1}, \tag{2.72}$$

où $b_i \mathrm{D}_{(i)}$ est associée à un schéma de différences finies pour l'opérateur $b_i(x) \dfrac{\partial}{\partial x}$, d'après (2.49) ou (2.50) en fonction du signe de b_i.
  – avec une vitesse $b = b_j\,(x_j)\, e_i$ (opérateur sur la moyenne pour une option asiatique) :

$$b_j \frac{\partial}{\partial x_i} \approx \mathrm{T}_{(j)} \circ \mathbf{b}_j \circ \mathrm{T}_{(j)}^{-1} \circ \mathrm{T}_{(i)} \circ \mathrm{D}_{(i)} \circ \mathrm{T}_{(i)}^{-1}, \tag{2.73}$$

où $\mathbf{b}_j$ est une matrice diagonale représentant l'opération de multiplication point par point par b_j.

– Multiplication par un coefficient variable de la forme $c\,(\boldsymbol{x}) = c_1\,(x_1) \ldots c_d\,(x_d)$,

$$\prod_{i=1}^{d} c^i \cdot = \mathrm{T}_{(d)} \circ \mathbf{c}_d \circ \mathrm{T}_{(d)}^{-1} \circ \cdots \circ \mathrm{T}_{(1)} \circ \mathbf{c}_1 \circ \mathrm{T}_{(1)}^{-1}. \tag{2.74}$$

Remarque 2.9 *L'opérateur de multiplication par un coefficient quelconque nécessite une multiplication point par point en base nodale. Une telle multiplication est coûteuse en ressource de calcul. En effet, il est nécessaire de revenir sur la base nodale dans chacune des dimensions. Dans le cas où le coefficient est de la forme (2.46), l'opérateur (2.74) est utilisé.*

Justification heuristique du choix des opérateurs sur une grille sparse Notre objectif est de donner au lecteur une intuition sur le choix du schéma de différences finies. Ce paragraphe n'a pas la prétention de fournir une démonstration de la consistance mais simplement de justifier l'intérêt d'un opérateur différent de l'opérateur de différences finies classique. La consistance des différents schémas est étudiée dans le cas simplifié où u est obtenue par produit tensoriel de fonctions d'une variable,

$$u\,(\boldsymbol{x}) := u_1\,(x_1) \ldots u_d\,(x_d)\,. \tag{2.75}$$

Considérons un schéma aux différences finies D linéaire d'ordre $M \geq 2$ associé à un opérateur différentiel sur la dimension 1. Soit $\mathcal{R}\,(u)$ le vecteur des valeurs de u sur les

points de la grille.

$$\begin{aligned}\left\|\mathcal{R}\left(\partial^2_{x_1}u\right)-D\mathcal{R}(u)\right\|_\infty &= \left\|\mathcal{R}(\partial^2_{x_1}u_1)-D\mathcal{R}(u_1)\right\|_\infty\prod_{k=2}^{d}\|\mathcal{R}(u_k)\|_\infty\\ &\approx 2^{-M\ell_g}\|u_1\|_{\mathcal{C}^4([0,1])}\prod_{k=2}^{d}\|u_k\|_{\mathcal{C}^0([0,1])}\\ &\approx 2^{-M\ell_g}\|u\|_{\mathcal{C}^{(4,0,\ldots,0)}\left([0,1]^d\right)},\end{aligned}\tag{2.76}$$

où $2^{-\ell^g}$ est le pas de discrétisation le plus grossier sur la *Sparse Grid*, c.-à-d. $\frac{1}{2}$.

Considérons l'opérateur différentiel discret obtenu par un schéma aux différences finies sur une base sparse. Si $u_1\in\mathcal{C}^4([0,1])$, alors

$$\left\|\partial^2_{x_1}u_1-D_{\ell_1}u_1\right\|_\infty\leq\ C2^{-2\ell_1}\|u_1\|_{\mathcal{C}^4([0,1])}.$$

Notons $\mathcal{P}$ la projection sur la base sparse et $\mathcal{E}$ l'erreur de consistance alors

$$\begin{aligned}\mathcal{E} &= \left\|\mathcal{P}\left(\partial^2_{x_1}u\right)-\mathrm{T}_{(1)}\circ\mathrm{D}^2_{(1)}\circ\mathrm{T}^{-1}_{(1)}\,\mathcal{P}\left(u\right)\right\|_\infty\\ &= \left\|\left(\left\langle\partial^2_{x_1}u,\widetilde{\psi}_{\boldsymbol{\ell},\boldsymbol{\imath}}\right\rangle\right)_{\boldsymbol{\ell}\in\mathcal{I}_n,\boldsymbol{\imath}\in\tau_{\boldsymbol{\ell}}}-\mathrm{T}_{(1)}\circ\mathrm{D}^2_{(1)}\circ\mathrm{T}^{-1}_{(1)}\ \left\langle u,\widetilde{\psi}_{\boldsymbol{\ell},\boldsymbol{\imath}}\right\rangle\right\|_\infty.\end{aligned}$$

En admettant cette égalité, que nous démontrons ultérieurement, nous obtenons par séparation des variables que

$$\begin{aligned}\mathcal{E} &= \max_{\boldsymbol{\ell}\in\mathcal{I},\boldsymbol{\imath}\in\tau_{\boldsymbol{\ell}}}\left\{\left|\left\langle\partial^2_{x_1}u_1,\widetilde{\psi}_{\ell_1,i_1}\right\rangle-\left\langle D^2_{\tilde{l}}u_1,\widetilde{\psi}_{\ell_1,i_1}\right\rangle\right|\prod_{k=2}^{d}\left|\langle u_k,\psi_{\ell_k,\imath_k}\rangle\right|\right\}\\ &= \max_{\boldsymbol{\ell}\in\mathcal{I},\boldsymbol{\imath}\in\tau_{\boldsymbol{\ell}}}\left\{\left|\left\langle\partial^2_{x_1}u_1-D^2_{\tilde{l}}u_1,\widetilde{\psi}_{\ell_1,i_1}\right\rangle\right|\prod_{k=2}^{d}\left|\langle u_k,\psi_{\ell_k,\imath_k}\rangle\right|\right\}\\ &\leq \max_{\boldsymbol{\ell}\in\mathcal{I},\boldsymbol{\imath}\in\tau_{\boldsymbol{\ell}}}\left\{\sup_{x\in supp\widetilde{\psi}_{\ell_1,i_1}}\left\{\left|\partial^2_{x_1}u_1-D^2_{\tilde{\ell}}u_1\right|\right\}\prod_{k=2}^{d}\underbrace{2^{-2\ell_k}\left|\left\langle\partial^2_{x_1}u_2,\psi_{\boldsymbol{\ell},\boldsymbol{\imath}}\right\rangle\right|}_{(1.86)}\right\}\\ &\leq C2^{-2\tilde{\ell}}\|u_1\|_{\mathcal{C}^4([0,1])}\textstyle\prod_{k=2}^{d}2^{-2\ell_k}\|u_k\|_{\mathcal{C}^2([0,1])}.\end{aligned}\tag{2.77}$$

Ce qui permet de conclure que, si $u_k\in\mathcal{C}^2\left([0,1]\right)$, alors

$$\left\|\mathcal{P}\left(\partial^2_{x_1}u\right)-\mathrm{T}_{(1)}\circ\mathrm{D}^2_{(1)}\circ\mathrm{T}^{-1}_{(1)}\mathcal{P}\left(u\right)\right\|_\infty\leq C2^{-n}\|u\|_{\mathcal{C}^{(4,2,\ldots,2)}\left([0,1]^d\right)},\tag{2.78}$$

qui est une estimation bien meilleure que (2.76).

2.3.2 Convergence de la méthode

2.3.2.1 Consistance des opérateurs sur une grille sparse

Les premiers résultats sur la consistance de l'opérateur $\widehat{\Delta}$ défini en (2.69) sont présentés par Schiekofer, [Sch98a]. Ces résultats sont obtenus pour la base de fonctions chapeaux et le schéma classique d'ordre 2 pour le laplacien. Une généralisation a été proposée par

Koster dans [Kos00]. Ces résultats sont valables pour des grilles non uniformes, des schémas d'ordre M quelconques et des ondelettes interpolantes d'ordre N.

Koster démontre un théorème de consistance pour l'opérateur *Sparse Grid* associé à $\dfrac{\partial^2}{\partial x_1^2}$.

Soit $\mathcal{P}$, $u \to u_{|\Omega_n}$ l'opérateur de restriction sur la *Sparse Grid*. Afin de simplifier les notations, l'opérateur de discrétisation de $\partial^2_{x_1}$ sur une *Sparse Grid* est noté $\mathcal{D}$.

$$\mathcal{D} = \mathrm{T}_{(1)} \circ \mathrm{D}^2_{(1)} \circ \mathrm{T}^{-1}_{(1)}.$$

Théorème 2.8 *Soient* $u \in \mathcal{C}^{\boldsymbol{\alpha}}\left([0,1]^d\right), \alpha_1 > 2, \alpha_i > 0, (2 \le i \le d)$, *alors*

$$\begin{aligned} &\left\|\mathcal{P}\left(\partial_{xx} u\right) - \mathcal{D}\left(\mathcal{P}(u)\right)\right\|_\infty \\ &\quad \le C\,|u|_{(\min(\alpha_1, M+2), \min(\alpha_2, N), \ldots, \min(\alpha_d, N))}\, n^{d-1} 2^{-\min(\alpha_1 - 2, \alpha_2 \ldots, \alpha_d, N, M) n} \\ &\quad + \; C \sup_{\substack{0 \le x_i \le 1 \\ 1 \le i \le d}} \left\{ |u(\cdot, x_2, \ldots, x_d)|_{\min(\alpha_1, M+2)}\, 2^{-\min(\alpha_1 - 2, M) n} \right\}. \end{aligned} \tag{2.79}$$

Remarque 2.10 *Supposons* $u \in \mathcal{C}^{\boldsymbol{4}}\left([0,1]^2\right)$. *Pour un schéma de différences finies classique* $(M = 2)$ *avec un pas de discrétisation en* 2^{-n} *(ce qui implique* 4^n *noeuds, l'erreur de consistance est en* $O\left(2^{-2n}\right)$, *l'opérateur sur une* Sparse Grid *permet d'obtenir une erreur en* $O\left(n 2^{-2n}\right)$ *pour seulement* $O\left(n 2^n\right)$ *degrés de liberté.*

Remarque 2.11 *L'estimation (2.79) pour l'opérateur* $\mathcal{D}$ *montre que le choix* $M = N$ *est judicieux. Celui-ci correspond au fait d'adapter l'ordre du schéma aux différences finies à celui des ondelettes interpolantes ou inversement.*

Preuve La démonstration de Koster [Kos00] s'articule en deux étapes. La première consiste à introduire un nouvel opérateur $\widetilde{\mathcal{D}}(u)$ pour la discrétisation de $\partial^2_{x_1} u$ par des techniques combinatoires, voir le théorème 2.2. L'analyse de la consistance est obtenu à la Proposition 2.10.

La deuxième étape consiste à montrer l'égalité entre les deux opérateurs $\mathcal{P}\widetilde{\mathcal{D}}$ et $\mathcal{D}\mathcal{P}$; ce résultat est obtenu à la Proposition 2.13. ■

Dans cette partie, nous notons $\mathcal{I}_n = \mathcal{I}_n^0$ (voir la définition 1.18).

Les différents opérateurs utilisés ensuite sur des grilles anisotropes sont définis :

Définition 2.4 (Opérateurs sur des grilles anisotropiques) *Soit* $\mathbb{P}_{\ell_1} u$ *la fonction de* $[0,1]^{d-1}$ *à valeurs dans* $\mathbb{R}^{2^{\ell_1}-1}$ *définie par*

$$\left[\mathbb{P}_{\ell_1} u\right]_i (x_2, \ldots, x_d) = \left[P_{\ell_1} u\,(\cdot, x_2, \ldots, x_d)\right]\left(\frac{i}{2^{\ell_1} - 1}\right) = u\left(\frac{i}{2^{\ell_1} - 1}, x_2, \ldots, x_d\right). \tag{2.80}$$

Pour toute fonction V *de* $[0,1]^{d-1}$ *à valeurs dans* $\mathbb{R}^{2^{\ell_1}-1}$, $\mathbb{I}_{\ell_1} V$ *est la fonction de* $[0,1]^d$ *à valeurs dans* $\mathbb{R}$ *définie par*

$$\left[\mathbb{I}_{\ell_1} V\right](x_1, \ldots, x_d) = \left(I_{\ell_1} V\right)(x_1, x_2, \ldots, x_d), \tag{2.81}$$

et $\mathbb{D}V$ *la fonction de* $[0,1]^{d-1}$ *à valeurs dans* $\mathbb{R}^{2^{\ell_1}-1}$ *définie par*

$$[\mathbb{D}_{\ell_1} V](x_1,\ldots,x_d) = D_{\ell_1}[V(x_2,\ldots,x_d)]. \tag{2.82}$$

L'opérateur D_{ℓ_1} *correspond au schéma aux différences finies pour* $\partial^2_{x_1}$.

Soit $\mathbb{P}_{\boldsymbol{\ell_2}} u$ *la fonction de* $[0,1]$ *à valeurs dans* $\mathbb{R}^{2^{\ell_2}-1}\times\cdots\times\mathbb{R}^{2^{\ell_d}-1}$ *définie par*

$$[\mathbb{P}_{\boldsymbol{\ell_2}} u]_{\boldsymbol{\imath_2}}(x) = u\left(x, \frac{i_2}{2^{\ell_2}-1},\ldots,\frac{i_d}{2^{\ell_d}-1}\right). \tag{2.83}$$

Pour toute fonction V *de* $[0,1]$ *à valeurs dans* $\mathbb{R}^{2^{\ell_2}-1}\times\cdots\times\mathbb{R}^{2^{\ell_d}-1}$, $\mathbb{I}_{\boldsymbol{\ell_2}} V$ *est la fonction de* $[0,1]^d$ *à valeurs dans* $\mathbb{R}$ *définie par*

$$[\mathbb{I}_{\boldsymbol{\ell_2}} V](x_1,\ldots,x_d) = I_{\ell_2}\otimes\cdots\otimes I_{\ell_d} V(x). \tag{2.84}$$

$I_{\boldsymbol{\ell}}$ *est l'opérateur d'interpolation de* $\Omega_{\boldsymbol{\ell}} \to \mathcal{C}^0\left(\bar{\Omega}\right)$ *défini par* $I_{\boldsymbol{\ell}} = I_{\ell_1}\otimes\cdots\otimes I_{\ell_d}$ *et* $P_{\boldsymbol{\ell}}$ *est l'opérateur de restriction* $u \to u_{|\Omega_{\boldsymbol{\ell}}}$.

Proposition 2.9 *Ces opérateurs sont liés par la relation*

$$I_{\boldsymbol{\ell}} P_{\boldsymbol{\ell}} = \mathbb{I}_{\boldsymbol{\ell_2}} \circ \mathbb{P}_{\boldsymbol{\ell_2}} \circ \mathbb{I}_{\ell_1} \circ \mathbb{P}_{\ell_1}. \tag{2.85}$$

Définition 2.5 (Opérateur $\widetilde{\mathcal{D}}(u)$) *Pour une fonction* $u \in \mathcal{C}^0\left(\bar{\Omega}\right)$ *telle que* $u = 0$ *sur* $\partial\Omega$, *Notons* $\widetilde{\mathcal{D}}(u)$ *une fonction de* V_n *telle que,*

$$\widetilde{\mathcal{D}}(u) = \left(\sum_{\boldsymbol{\ell}\in\mathcal{I}_n} f(\boldsymbol{\ell}) \mathbb{I}_{\ell_1}\mathbb{D}_{\ell_1}\mathbb{P}_{\ell_1}\right) u, \tag{2.86}$$

où $f(\boldsymbol{\ell})$ *est définie par l'équation (2.3).*

Étape 1, consistance de $\widetilde{\mathcal{D}}(u)$

Proposition 2.10 (Consistance de l'opérateur $\widetilde{\mathcal{D}}$) *Soient* $u \in \mathcal{C}^{\boldsymbol{\alpha}}\left([0,1]^d\right), \alpha_1 > 2, \alpha_i > 0, (2 \le i \le d)$, *alors*

$$\begin{aligned}
&\left\|\mathcal{P}\left(\partial^2_{x_1} u\right) - \mathcal{P}\left(\widetilde{\mathcal{D}}(u)\right)\right\|_\infty \\
&\quad \le C\,|u|_{(\min(\alpha_1,M+2),\min(\alpha_2,N),\ldots,\min(\alpha_d,N))}\, n^{d-1} 2^{-\min(\alpha_1-2,\alpha_2\ldots,\alpha_d,N,M)n} \\
&\quad + C \sup_{\substack{0\le x_i\le 1\\ 1\le i\le d}} \left\{|u(\cdot,x_2,\ldots,x_d)|_{\min(\alpha_1,M+2)}\, 2^{-\min(\alpha_1-2,M)n}\right\}.
\end{aligned} \tag{2.87}$$

Afin d'alléger les équations, notons $\boldsymbol{x_2}$, *resp* $\boldsymbol{\alpha_2}$, le $d-1$ uplet $(x_2,\ldots,x_d)$, $(\alpha_2,\ldots,\alpha_d)$. La démonstration de la Proposition 2.10 repose sur la décomposition suivante :

Soit $E_D(u,\ell_1)$ l'erreur de consistance du schéma aux différences finies dans la dimension 1.

$$E_D(u,\ell_1) = \mathbb{I}_{\ell_1}\circ\mathbb{P}_{\ell_1}\left(\partial^2_{x_1} u\right) - \mathbb{I}_{\ell_1}\circ\mathbb{D}_{\ell_1}\circ\mathbb{P}_{\ell_1}(u). \tag{2.88}$$

$E_D(u,\ell_1)$ est une fonction de $[0,1]^{d-1}$ à valeurs dans $\mathbb{R}^{2^{\ell_1}-1}$. Soit $E_P(v,\boldsymbol{\ell_2})$ l'erreur de projection sur la grille $\Omega_{\boldsymbol{\ell_2}}$.

$$E_P(v,\boldsymbol{\ell_2}) = v - \mathbb{I}_{\boldsymbol{\ell_2}}\circ\mathbb{P}_{\boldsymbol{\ell_2}} v. \tag{2.89}$$

$E_P(v,\boldsymbol{\ell_2})$ est une fonction de $[0,1]^d$ à valeurs dans $\mathbb{R}$.

Proposition 2.11

$$\begin{aligned}\mathcal{P}\left(\partial^2_{x_1}u\right)-\mathcal{P}\circ\widetilde{\mathcal{D}}\left(u\right) &= \mathcal{P}\sum_{\boldsymbol{\ell}\in\mathcal{I}_n} f(\boldsymbol{\ell})\,\mathbb{I}_{\widetilde{\ell}}\circ E_D\left(u,\vec{\widetilde{\ell}}\right)\\ &- \mathcal{P}\sum_{\boldsymbol{\ell}\in\mathcal{I}_n} f(\boldsymbol{\ell})\,\mathbb{I}_{\ell_1}\circ E_P\left(E_D(u,\ell_1),\boldsymbol{\ell_2}\right),\end{aligned} \tag{2.90}$$

où $\tilde{\ell}$ est défini par (2.59).

Preuve Par application du Théorème 2.2,

$$\mathcal{P}\,v=\sum_{\boldsymbol{\ell}\in\mathcal{I}_n} f(\boldsymbol{\ell})I_{\boldsymbol{\ell}}\circ P_{\boldsymbol{\ell}}\,v.$$

Nous déduisons de (2.86) que

$$\begin{aligned}&\mathcal{P}\left(\partial^2_{x_1}u\right)-\mathcal{P}\circ\widetilde{\mathcal{D}}\left(u\right)\\ &\quad=\sum_{\boldsymbol{\ell}\in\mathcal{I}_n} f(\boldsymbol{\ell})\mathbb{I}_{\boldsymbol{\ell_2}}\circ\mathbb{P}_{\boldsymbol{\ell_2}}\left[\mathbb{I}_{\ell_1}\circ\mathbb{P}_{\ell_1}\left(\partial^2_{x_1}u\right)-\mathbb{I}_{\ell_1}\circ\mathbb{P}_{\ell_1}\circ\mathbb{I}_{\ell_1}\mathbb{D}_{\ell_1}\circ\mathbb{P}_{\ell_1}\left(u\right)\right]\\ &\quad=\sum_{\boldsymbol{\ell}\in\mathcal{I}_n} f(\boldsymbol{\ell})\mathbb{I}_{\boldsymbol{\ell_2}}\circ\mathbb{P}_{\boldsymbol{\ell_2}}\left[\mathbb{I}_{\ell_1}\circ\mathbb{P}_{\ell_1}\left(\partial^2_{x_1}u\right)-\mathbb{I}_{\ell_1}\mathbb{D}_{\ell_1}\circ\mathbb{P}_{\ell_1}\left(u\right)\right].\end{aligned} \tag{2.91}$$

Alors, en considérant les équations (2.88) et (2.89),

$$\begin{aligned}\mathcal{I} &= \mathbb{I}_{\boldsymbol{\ell_2}}\circ\mathbb{P}_{\boldsymbol{\ell_2}}\left[\mathbb{I}_{\ell_1}\circ\mathbb{P}_{\ell_1}\left(\partial^2_{x_1}u\right)-\mathbb{I}_{\ell_1}\mathbb{D}_{\ell_1}\circ\mathbb{P}_{\ell_1}\left(u\right)\right]\\ &= \left[\mathbb{I}_{\ell_1}\circ\mathbb{P}_{\ell_1}\left(\partial^2_{x_1}u\right)-\mathbb{I}_{\ell_1}\mathbb{D}_{\ell_1}\circ\mathbb{P}_{\ell_1}\left(u\right)\right]-E_P\left(\mathbb{I}_{\ell_1}\circ\mathbb{P}_{\ell_1}\left(\partial^2_{x_1}u\right)-\mathbb{I}_{\ell_1}\mathbb{D}_{\ell_1}\circ\mathbb{P}_{\ell_1}\left(u\right),\boldsymbol{\ell_2}\right)\\ &= \left(E_D\left(u,\ell_1\right)-E_P\left(E_D\left(u,\ell_1\right),\boldsymbol{\ell_2}\right)\right)\\ &= \Gamma_1\left(\ell_1\right)-\Gamma_2\left(\boldsymbol{\ell}\right).\end{aligned} \tag{2.92}$$

Le terme Γ_2 correspond au deuxième terme du second membre de l'équation (2.90). Considérons donc le premier terme Γ_1. Nous aurons besoin du lemme suivant :

Lemme 2.12 *La fonction f définie par l'équation (2.3) vérifie*

$$\sum_{\substack{\boldsymbol{k}\,\geq\,\boldsymbol{\ell}\\ k_1\,=\,\ell_1}} f\left(\boldsymbol{k}\right)=\begin{cases}1, & si\quad \ell_1=\widetilde{\ell},\\ 0, & sinon.\end{cases} \tag{2.93}$$

Preuve D'après (2.24), $\displaystyle\sum_{\boldsymbol{k}\geq\boldsymbol{\ell}} f\left(\boldsymbol{k}\right)=1$. En remarquant que $\boldsymbol{k}\geq\boldsymbol{\ell}$, $|\boldsymbol{k}|_1\leq n+d-1$ et $k_1=\widetilde{\ell}$ impliquent $\boldsymbol{k}=\boldsymbol{\ell}$. Nous concluons dans le cas $\ell_1=\widetilde{\ell}$.
Dans le cas où $\ell_1<\widetilde{\ell}$, une récurrence sur $\widetilde{\ell}-\ell_1$ donne :

$$\begin{aligned}\mathcal{R}\left(\boldsymbol{\ell}\right) &= \sum_{\substack{\boldsymbol{k}\,\geq\,\boldsymbol{\ell}\\ k_1\,\leq\,\ell_1}} f\left(\boldsymbol{k}\right)\\ &= \sum_{\boldsymbol{k}\geq\boldsymbol{\ell}} f\left(\boldsymbol{k}\right)-\sum_{\substack{\boldsymbol{k}\,\geq\,\boldsymbol{\ell}\\ k_1\,=\,\widetilde{\ell}}} f\left(\boldsymbol{k}\right)-\sum_{j=1}^{\widetilde{\ell}-\ell_1-1}\sum_{\substack{\boldsymbol{k}\,\geq\,\boldsymbol{\ell}\\ k_1\,=\,\widetilde{\ell}-j}} f\left(\boldsymbol{k}\right)\\ &= -\sum_{j=1}^{\widetilde{\ell}-\ell_1-1}\sum_{\substack{\boldsymbol{k}\,\geq\,\boldsymbol{\ell}\\ k_1\,=\,\widetilde{\ell}-j}} f\left(\boldsymbol{k}\right).\end{aligned} \tag{2.94}$$

Dans le cas $\ell_1 = \widetilde{\ell} - 1$, trivialement $\mathcal{R}(\boldsymbol{\ell}) = 0$. Une récurrence sur (2.94) permet de montrer que le résultat est vrai pour tout $\ell_1 \in 1, \ldots, \widetilde{\ell} - 1$. Il suffit alors de remarquer que

$$\sum_{\substack{\boldsymbol{k} \geq \boldsymbol{\ell} \\ k_1 = \ell_1}} f(\boldsymbol{k}) = \mathcal{R}(\ell_1) - \mathcal{R}(\ell_1 - 1),$$

pour montrer (2.93). ■

Revenons à la démonstration de (2.90),

$$\begin{aligned} \sum_{\boldsymbol{\ell} \in \mathcal{I}_n} f(\boldsymbol{\ell}) \Gamma_1(\ell_1) &= \sum_{i=0}^{\widetilde{\ell}} \Gamma_1(i) \sum_{\substack{\boldsymbol{\ell} \in \mathcal{I}_n \\ \ell_1 = i}} f(\boldsymbol{\ell}) \\ &= \sum_{i=0}^{\widetilde{\ell}} \Gamma_1(i) \sum_{\substack{\boldsymbol{\ell} \in \mathcal{I}_n \\ \ell_1 = i}} 1_{\{i = \widetilde{\ell}\}} \\ &= \Gamma_1\left(\widetilde{\ell}\right). \end{aligned} \tag{2.95}$$

Ce qui termine la démonstration de la Proposition 2.11. ■

Preuve de la proposition 2.10. La proposition 2.7 permet de montrer les deux majorations :

– Pour $u(\cdot, \boldsymbol{x_2}) \in \mathcal{C}^{\alpha_1}([0,1])$,

$$\|\Gamma_1(\ell_1)\|_\infty \leq C 2^{-\min(M, \alpha_1 - 2)\ell_1} \sup_{\substack{0 \leq x_i \leq 1 \\ 1 \leq i \leq d}} |u(\cdot, \boldsymbol{x_2})|_{\min(\alpha_1, M+2)}. \tag{2.96}$$

– Pour $u \in \mathcal{C}^{\boldsymbol{\alpha}}([0,1]^d)$, $E_D(u, \ell_1) \in \mathcal{C}^{\boldsymbol{\alpha_2}}([0,1]^{d-1})$ et

$$|E_D(u, \ell_1)|_{\boldsymbol{\alpha_2}} \leq C |u|_{\min(\alpha_1, M+2), \alpha_2, \ldots, \alpha_d} 2^{-\min(\alpha_1 - 2, M)\ell_1}. \tag{2.97}$$

L'équation (2.58) permet de conclure

$$\begin{aligned} &\|E_P(E_D(u, \ell_1), \boldsymbol{\ell_2})\|_\infty \\ &\quad \leq C |u|_{\min(\alpha_1, M+2), \min(\alpha_2, N), \ldots, \min(\alpha_d, N)} 2^{-\min(\alpha_1 - 2, M)\ell_1 - \sum_{k=2}^d min(\alpha_k, N)\ell_k}. \end{aligned} \tag{2.98}$$

$$\left| \sum_{\boldsymbol{\ell}} f(\boldsymbol{\ell}) \Gamma_2(\boldsymbol{\ell}) \right| = \left| \sum_{j=n}^{n+d-1} \sum_{\boldsymbol{\ell} : |\boldsymbol{\ell}|_1 = j} f(\boldsymbol{\ell}) \Gamma_2(\boldsymbol{\ell}) \right|. \tag{2.99}$$

La valeur de $f(\boldsymbol{\ell})$ ne dépend que de $|\boldsymbol{\ell}|_1$, nous en déduisons

$$\begin{aligned} \left| \sum_{\boldsymbol{\ell}} f(\boldsymbol{\ell}) \Gamma_2(\boldsymbol{\ell}) \right| &\leq \sum_{j=n}^{n+d-1} \sum_{\boldsymbol{\ell} : |\boldsymbol{\ell}|_1 = n+d-1} |f(\boldsymbol{\ell})| \, \|\Gamma_2(\boldsymbol{\ell})\|_\infty, \quad \text{de (2.3)} \\ &\leq \sum_{j=n}^{n+d-1} \binom{d-1}{j-n} \binom{d-1+j}{j} \|\Gamma_2(\boldsymbol{\ell})\|_\infty, \quad \text{de (2.27)} \\ &\leq \sum_{j=n}^{n+d-1} \frac{(d-1+j)!}{(j-n)!(d-1-j+n)!j!} \max_{n \leq |\boldsymbol{\ell}|_1 \leq n+d-1} \|\Gamma_2(\boldsymbol{\ell})\|_\infty. \end{aligned} \tag{2.100}$$

En remarquant que

$$0 \leq j-n \leq d-1 \quad \text{et} \quad 0 \leq n+d-1-j \leq d-1,$$

$$\begin{aligned} \left| \sum_{\boldsymbol{\ell}} f(\boldsymbol{\ell}) \Gamma_2(\boldsymbol{\ell}) \right| &\leq \sum_{j=n}^{n+d-1} (j+d-1)\dots(j+1) \max_{n \leq |\boldsymbol{\ell}|_1 \leq n+d-1} \|\Gamma_2(\boldsymbol{\ell})\|_\infty \\ &\leq (n+d-1)^{d-1} \max_{n \leq |\boldsymbol{\ell}|_1 \leq n+d-1} \|\Gamma_2(\boldsymbol{\ell})\|_\infty. \end{aligned} \tag{2.101}$$

Les équations (2.98) et (2.101) permettent de conclure.

■

Étape 2, égalité des opérateurs $\mathcal{D}\mathcal{P}$ et $\mathcal{P}\widetilde{\mathcal{D}}$

Proposition 2.13 (Egalité des deux opérateurs) *L'opérateur $\mathcal{P}\widetilde{\mathcal{D}}$ introduit par la définition 2.5 coïncide avec l'opérateur $\mathcal{D}\mathcal{P}$.*

Preuve Selon (2.62), $\mathcal{D}\mathcal{P}(u)$ s'écrit sur la base d'ondelettes

$$\sum_{\boldsymbol{\ell} \in \mathcal{I}_n, \boldsymbol{\imath} \in \boldsymbol{\tau}_{\boldsymbol{\ell}}} \boldsymbol{v}^{\widetilde{\ell}}_{\boldsymbol{\ell},\boldsymbol{\imath}} \psi_{\boldsymbol{\ell},\boldsymbol{\imath}}, \tag{2.102}$$

où les $v_{\boldsymbol{\ell},\boldsymbol{\imath}}$ sont les coefficients du vecteur $\boldsymbol{V}_{\boldsymbol{\ell_2},\boldsymbol{\imath_2}}$ donné par l'équation (2.62),

$$\boldsymbol{v}_{\boldsymbol{\ell},\boldsymbol{\imath}} = \boldsymbol{V}_{\boldsymbol{\ell_2},\boldsymbol{\imath_2}}(\ell_1, \imath_1).$$

L'opérateur $\widetilde{\mathcal{D}}$ appliqué à $\sum_{\boldsymbol{\ell} \in \mathcal{I}_n, \boldsymbol{\imath} \in \boldsymbol{\tau}_{\boldsymbol{\ell}}} \boldsymbol{u}_{\boldsymbol{\ell},i} \psi_{\boldsymbol{\ell},\boldsymbol{\imath}}$ peut s'écrire sous la forme (2.102).

$$\widetilde{\mathcal{D}} \left(\sum_{\boldsymbol{\ell} \in \mathcal{I}_n, \boldsymbol{\imath} \in \tau_{\boldsymbol{\ell}}} \hat{u}_{\boldsymbol{\ell},\boldsymbol{\imath}} \psi_{\boldsymbol{\ell},\boldsymbol{\imath}} \right) = \sum_{\boldsymbol{k} \in \mathcal{I}_n} f(\boldsymbol{k}) \, \mathbb{I}_{k_1} \circ \mathbb{D}_{k_1} \circ P_{k_1} \left(\sum_{\boldsymbol{\ell} \in \mathcal{I}_n, \boldsymbol{\imath} \in \tau_{\boldsymbol{\ell}}} \hat{u}_{\boldsymbol{\ell},\boldsymbol{\imath}} \psi_{\boldsymbol{\ell},\boldsymbol{\imath}} \right). \tag{2.103}$$

En remarquant que $P^{\boldsymbol{k}}(\psi_{\boldsymbol{\ell},\boldsymbol{\imath}}) = 0$ si $\boldsymbol{\ell} > \boldsymbol{k}$, nous déduisons :

$$\begin{aligned} \widetilde{\mathcal{D}} \left(\textstyle\sum_{\boldsymbol{\ell} \in \mathcal{I}_n, \boldsymbol{\imath} \in \tau_{\boldsymbol{\ell}}} \hat{u}_{\boldsymbol{\ell},\boldsymbol{\imath}} \psi_{\boldsymbol{\ell},\boldsymbol{\imath}} \right) &= \sum_{\boldsymbol{k} \in \mathcal{I}_n} f(\boldsymbol{k}) \, \mathbb{I}_{k_1} \circ \mathbb{D}_{k_1} \circ \mathbb{P}_{k_1} \left(\sum_{\boldsymbol{\ell} \leq \boldsymbol{k}, \boldsymbol{\imath} \in \tau_{\boldsymbol{\ell}}} \hat{u}_{\boldsymbol{\ell},\boldsymbol{\imath}} \psi_{\boldsymbol{\ell},\boldsymbol{\imath}} \right) \\ &= \sum_{\boldsymbol{k} \in \mathcal{I}_n} f(\boldsymbol{k}) \left(I_{k_1} \circ D_{k_1} \circ P_{k_1} \right) \left(\sum_{\boldsymbol{\ell} \leq \boldsymbol{k}, \boldsymbol{\imath} \in \tau_{\boldsymbol{\ell}}} \hat{u}_{\boldsymbol{\ell},\boldsymbol{\imath}} \psi_{\boldsymbol{\ell},\boldsymbol{\imath}} \right) \\ &= \sum_{\boldsymbol{k} \in \mathcal{I}_n} f(\boldsymbol{k}) \sum_{\boldsymbol{\ell} \leq \boldsymbol{k}, \boldsymbol{\imath} \in \tau_{\boldsymbol{\ell}}} v^{k_1}_{\boldsymbol{\ell},\boldsymbol{\imath}} \psi_{\boldsymbol{\ell},\boldsymbol{\imath}} \\ &= \sum_{\boldsymbol{\ell} \in \mathcal{I}_n, \boldsymbol{\imath} \in \tau_{\boldsymbol{\ell}}} \sum_{\substack{\boldsymbol{k} \geq \boldsymbol{\ell} \\ \widetilde{\ell}}} f(\boldsymbol{k}) v^{k_1}_{\boldsymbol{\ell},\boldsymbol{\imath}} \psi_{\boldsymbol{\ell},\boldsymbol{\imath}} \\ &= \sum_{\boldsymbol{\ell} \in \mathcal{I}_n, \boldsymbol{\imath} \in \tau_{\boldsymbol{\ell}}} \sum_{j=\ell_1} v^{j}_{\boldsymbol{\ell},\boldsymbol{\imath}} \psi_{\boldsymbol{\ell},\boldsymbol{\imath}} \sum_{\substack{\boldsymbol{k} \geq \boldsymbol{\ell} \\ k_1 = j}} f(\boldsymbol{k}) \\ &= \sum_{\boldsymbol{\ell} \in \mathcal{I}_n, \boldsymbol{\imath} \in \tau_{\boldsymbol{\ell}}} v^{\widetilde{\ell}}_{\boldsymbol{\ell},\boldsymbol{\imath}} \psi_{\boldsymbol{\ell},\boldsymbol{\imath}}. \end{aligned} \tag{2.104}$$

La dernière ligne se déduisant du lemme 2.12. ■

2.3.2.2 Stabilité et convergence des opérateurs sur une grille sparse

Soit L_{Δ^S} la matrice du laplacien discrétisée sur une grille sparse (de niveau n). Schiekofer a étudié le comportement numérique de la norme $\|\ \|_\infty$ de l'inverse de cette matrice qui caractérise la stabilité de la discrétisation par différences finies. Il montre, numériquement, que $\left\|L^{-1}_{\Delta^{SG}}\right\|_\infty / n \approx C$ dans le cas $d = 2$. D'autres tests numériques en dimension 3 permettent de conjecturer que :

$$\left\|L^{-1}_{\Delta^{SG}}\right\|_\infty \leq Cn^{d-1} = C\left|\log h\right|^{d-1} \quad \text{si } d = 3 \text{ ou } d = 2 \text{ et } h = 2^n. \tag{2.105}$$

En admettant (2.105) pour d quelconque (il n'y a pas de preuve à notre connaissance), nous déduisons des résultats de consistance la conjecture suivante :

Conjecture 2.14 *Si la solution du problème de Laplace* $\left(\Delta u = f, \quad u_{|\partial\Omega} = 0\right)$ *est suffisamment régulière* $u \in \mathcal{C}^{\mathbf{4}}_{mix}([0,1]^d)$ *(ce qui impose des conditions de régularité et de compatibilité sur le second membre* f*), alors la discrétisation sur une grille sparse conduit à l'estimation d'erreur :*

$$\|u - \mathcal{P}(u)\| \leq C\ \left|\log h\right|^{2(d-1)}\ h^2 = Ch^2\left|\log h\right|^{2(d-1)}. \tag{2.106}$$

Preuve partielle Il suffit d'utiliser le résultat de consistance donné par le théorème 2.8 et celui de stabilité de l'équation (2.105) dans la relation suivante

$$\begin{aligned} \|u - \mathcal{P}(u)\| &= \left\|L^{-1}_{\Delta^{SG}}\left(\Delta^{SG}\mathcal{P}(u) - \mathcal{P}(\Delta u)\right)\right\|_\infty \\ &\leq \left\|L^{-1}_{\Delta^{SG}}\right\|_\infty \left\|\Delta^{SG}\mathcal{P}(u) - \mathcal{P}(\Delta u)\right\|_\infty \\ &\leq C_1\ \left|\log h\right|^{d-1}\ C_2\, h^2\left|\log h\right|^{d-1} = Ch^2\left|\log h\right|^{2(d-1)}. \end{aligned} \tag{2.107}$$

■

Remarque 2.12 *Dans le cas de la méthode de* technique combinatoire*, l'estimation d'erreur en* $Ch^2\left|\log h\right|^{d-1}$*, voir le théorème 2.6, mais cette erreur n'est valable que sur les points de la grille sparse.*

2.3.3 Différences finies et méthode de collocation

Dans une première partie, nous décrivons la méthode qui permet de retrouver le schéma de discrétisation sur une *Sparse Grid* proposé par Griebel et donné dans les équations (2.69),...,(2.74). Notre méthode présente l'intérêt d'unifier les méthodes de différences finies, de Galerkin et de Petrov Galerkin. Ceci permet de proposer une façon d'écrire le code de différences finies sur une *Sparse Grid*, semblable à celle utilisée pour une méthode de Galerkin. Nous proposons, dans une seconde partie, l'étude de la consistance de ce schéma, qui nous apparaît plus simple que l'approche précédente.

2.3.3.1 Description de l'opérateur

La méthode est décrite en trois étapes. Nous la présentons sur le problème de discrétisation de l'équation elliptique :

Soit $\mathcal{L}$ un opérateur différentiel de $H^2(\Omega) \rightarrow L^2(\Omega)$, et $f \in L^2(\Omega)$, nous cherchons la fonction $u \in H^2(\Omega)$ solution de

$$\begin{array}{rcll} \mathcal{L}u(\boldsymbol{x}) & = & f(\boldsymbol{x}), & \boldsymbol{x} \in \Omega \\ u_{\big|\partial\Omega}(\boldsymbol{x}) & = & 0 & \boldsymbol{x} \in \partial\Omega. \end{array} \tag{2.108}$$

Nous supposons que la solution u est régulière et, plus précisément, que $u \in \mathcal{C}^4_{mix}([0,1]^d)$ (voir la définition 1.5) si l'opérateur $\mathcal{L}$ est d'ordre deux dans chacune des dimensions.

1. En projetant l'équation (2.108) sur l'espace sparse $\widetilde{V}_n^0$, le problème initial prend la forme variationnelle :

   Trouver u telle que

   $$\langle \mathcal{L}u, v_n \rangle = \langle f, v_n \rangle \quad \forall v_n \in \widetilde{V}_n^0. \tag{2.109}$$

   En remplaçant v_n par les fonctions de base $\left(\widetilde{\psi}_{\boldsymbol{\ell},\boldsymbol{i}}\right)_{\boldsymbol{\ell}\in\mathcal{I}_n,\boldsymbol{i}\in\tau_{\boldsymbol{\ell}}}$ de l'espace sparse $\widetilde{V}_n^0$, le membre de gauche de l'équation (2.109) devient combinaison des valeurs de la fonction $\mathcal{L}u$ en différents points.

   En effet, il suffit d'appliquer la relation d'échelle sur les ondelettes duales (1.66) au membre de gauche de l'équation (2.109) après avoir remplacé v_n par $\widetilde{\psi}_{\boldsymbol{\ell},\boldsymbol{i}}$.

   $$\left\langle \mathcal{L}u, \widetilde{\psi}_{\boldsymbol{\ell},\boldsymbol{i}} \right\rangle = \sum_{\tilde{\boldsymbol{n}}} \tilde{g}(\tilde{\boldsymbol{n}}) \left\langle \mathcal{L}u, \widetilde{\varphi}_{\boldsymbol{\ell}+\mathbf{1},2\boldsymbol{i}+\tilde{\boldsymbol{n}}} \right\rangle. \tag{2.110}$$

   De plus, nous avons considéré les ondelettes interpolantes. La fonction d'échelle duale est donc une distribution de Dirac. En conclusion, $\langle \mathcal{L}u, \widetilde{\varphi}_{\boldsymbol{\ell},\boldsymbol{i}} \rangle = (\mathcal{L}u)(x_{\ell,\boldsymbol{i}})$.

2. L'approximation consiste alors à approcher le calcul de $(\mathcal{L}u)(x_{\ell,\boldsymbol{i}})$ par le schéma aux différences finies adapté à l'opérateur $\mathcal{L}$ appliqué au point $\boldsymbol{x}_{\ell,\boldsymbol{i}}$. Afin de rester dans un cadre général, le schéma aux différences finies associé à l'opérateur $\mathcal{L}$ est décrit comme une combinaison linéaire des valeurs aux points d'un voisinage de $x_{\boldsymbol{\ell},\boldsymbol{i}}$. Notons que les $\omega_{\boldsymbol{\ell},\boldsymbol{k}}$ de ce schéma dépendent de la distance entre deux points voisins, donc du niveau $\boldsymbol{\ell}$ du point $\boldsymbol{x}_{\boldsymbol{\ell},\boldsymbol{i}}$. La fonction $\mathcal{L}u$ au point $\boldsymbol{x}_{\boldsymbol{\ell},\boldsymbol{i}}$ est approchée par la combinaison linéaire :

   $$(\mathcal{L}u)(x_{\boldsymbol{\ell},\boldsymbol{i}}) \approx \sum_{\boldsymbol{k}\in\Gamma} \omega_{\boldsymbol{\ell},\boldsymbol{k}} u(x_{\ell,\boldsymbol{i}+\boldsymbol{k}}), \tag{2.111}$$

   en notant Γ le stencil du schéma aux différences finies et M l'ordre de ce schéma. Remarquons que la notation est imprécise, en effet $\boldsymbol{k}$ n'est pas un entier. En dimension une, par exemple x_1, le pas de discrétisation du schéma est donné par $2^{-\tilde{\ell}}$ (voir (2.59) pour la définition de $\tilde{\ell}$). Nous en déduisons

   $$k_1 = j2^{-\tilde{\ell}}, \tag{2.112}$$

   où j varie en fonction du nombre de points du schéma.

   Remarque 2.13 *Dans le cas d'une méthode de Galerkin, nous ne calculons pas exactement les coefficients des matrices de rigidité. Ce calcul est remplacé par une formule de quadrature numérique et* $\langle \mathcal{L}u, \widetilde{\varphi}_{\boldsymbol{\ell}+\mathbf{1},2\boldsymbol{i}+\tilde{\boldsymbol{n}}} \rangle$ *est approché.*

En résumé, le calcul de $\left\langle \mathcal{L}u, \widetilde{\psi}_{\boldsymbol{\ell},\boldsymbol{\imath}} \right\rangle$ est approché par :

$$\begin{aligned} \left\langle \mathcal{L}u, \widetilde{\psi}_{\boldsymbol{\ell},\boldsymbol{\imath}} \right\rangle &\approx \sum_{\tilde{\boldsymbol{n}}} \tilde{g}\left(\tilde{\boldsymbol{n}}\right) \sum_{\boldsymbol{k}\in\Gamma} \omega_{\boldsymbol{k}} u\left(x_{\boldsymbol{\ell}+\mathbf{1},2\boldsymbol{\imath}+\tilde{\boldsymbol{n}}+\boldsymbol{k}}\right) \\ &\approx \sum_{\tilde{\boldsymbol{n}}} \tilde{g}\left(\tilde{\boldsymbol{n}}\right) \sum_{\boldsymbol{k}\in\Gamma} \omega_{\boldsymbol{k}} \left\langle u, \widetilde{\varphi}_{\boldsymbol{\ell}+\mathbf{1},2\boldsymbol{\imath}+\tilde{\boldsymbol{n}}+\boldsymbol{k}} \right\rangle . \end{aligned} \tag{2.113}$$

3. La dernière étape, qui permet d'aboutir à une formulation discrète, est semblable à celle obtenue par une méthode de Petrov Galerkin et consiste à remplacer la fonction u par son approximation sur l'espace V_n^0,

$$u \approx \sum_{\bar{\boldsymbol{\ell}}\in\mathcal{I}_n, \bar{\boldsymbol{\imath}}\in\tau_{\bar{\boldsymbol{\ell}}}} \hat{u}_{\bar{\boldsymbol{\ell}},\bar{\boldsymbol{\imath}}} \psi_{\bar{\boldsymbol{\ell}},\bar{\boldsymbol{\imath}}}. \tag{2.114}$$

Ceci nous donne le schéma de discrétisation

$$\begin{aligned} \left\langle \mathcal{L}u, \widetilde{\psi}_{\boldsymbol{\ell},\boldsymbol{\imath}} \right\rangle &\approx \sum_{\tilde{\boldsymbol{n}}} \tilde{g}\left(\tilde{\boldsymbol{n}}\right) \sum_{\boldsymbol{k}\in\Gamma} \omega_{\boldsymbol{k}} \sum_{\bar{\boldsymbol{\ell}}\in\mathcal{I}_n, \bar{\boldsymbol{\imath}}\in\tau_{\bar{\boldsymbol{\ell}}}} \left\langle \psi_{\bar{\boldsymbol{\ell}},\bar{\boldsymbol{\imath}}}, \widetilde{\varphi}_{\boldsymbol{\ell}+\mathbf{1},2\boldsymbol{\imath}+\tilde{\boldsymbol{n}}+\boldsymbol{k}} \right\rangle \hat{u}_{\bar{\boldsymbol{\ell}},\bar{\boldsymbol{\imath}}} \\ &\approx \sum_{\tilde{\boldsymbol{n}}} \tilde{g}\left(\tilde{\boldsymbol{n}}\right) \sum_{\boldsymbol{k}\in\Gamma} \omega_{\boldsymbol{k}} \sum_{\bar{\boldsymbol{\ell}}\in\mathcal{I}_n, \bar{\boldsymbol{\imath}}\in\tau_{\bar{\boldsymbol{\ell}}}} \sum_{\boldsymbol{n}} g\left(\boldsymbol{n}\right) \left\langle \varphi_{\bar{\boldsymbol{\ell}}+\mathbf{1},2\bar{\boldsymbol{\imath}}+\boldsymbol{n}}, \widetilde{\varphi}_{\boldsymbol{\ell}+\mathbf{1},2\boldsymbol{\imath}+\tilde{\boldsymbol{n}}+\boldsymbol{k}} \right\rangle \hat{u}_{\bar{\boldsymbol{\ell}},\bar{\boldsymbol{\imath}}} \\ &\approx \sum_{\bar{\boldsymbol{\ell}}\in\mathcal{I}_n, \bar{\boldsymbol{\imath}}\in\tau_{\bar{\boldsymbol{\ell}}}} \underbrace{\sum_{\boldsymbol{k}\in\Gamma} \omega_{\boldsymbol{k}}}_{\substack{\text{schéma} \\ \text{aux FD}}} \underbrace{\sum_{\tilde{\boldsymbol{n}},\boldsymbol{n}} \tilde{g}\left(\tilde{\boldsymbol{n}}\right) g\left(\boldsymbol{n}\right) \left\langle \varphi_{\bar{\boldsymbol{\ell}}+\mathbf{1},2\bar{\boldsymbol{\imath}}+\boldsymbol{n}}, \widetilde{\varphi}_{\boldsymbol{\ell}+\mathbf{1},2\boldsymbol{\imath}+\tilde{\boldsymbol{n}}+\boldsymbol{k}} \right\rangle}_{\substack{\text{passage nodale} \\ \rightarrow \text{ hiérarchique}}} \underbrace{\hat{u}_{\bar{\boldsymbol{\ell}},\bar{\boldsymbol{\imath}}}}_{\substack{\& \\ \text{passage hiérarchique} \\ \rightarrow \text{ nodale}}} \end{aligned} \tag{2.115}$$

Remarque 2.14 *Le calcul de* $\sum_{\tilde{\boldsymbol{n}},\boldsymbol{n}} \tilde{g}\left(\tilde{\boldsymbol{n}}\right) g\left(\boldsymbol{n}\right) \left\langle \varphi_{\bar{\boldsymbol{\ell}}+\mathbf{1},2\bar{\boldsymbol{\imath}}+\boldsymbol{n}}, \widetilde{\varphi}_{\boldsymbol{\ell}+\mathbf{1},2\boldsymbol{\imath}+\tilde{\boldsymbol{n}}+\boldsymbol{k}} \right\rangle$ *s'effectue suivant la méthode : appliquer pour chaque dimension* k *le filtre sur la fonction d'échelle primale (*resp.*duale) tant que* $\bar{\ell}_k < \ell_k$, (resp.$\bar{\ell}_k > \ell_k$). *Une fois les fonctions d'échelles primales et duales au même niveau, les bases sont orthogonales, donc* $\left\langle \varphi_{\bar{\ell}_k,\bar{\imath}_k}, \widetilde{\varphi}_{\ell_k,\imath_k} \right\rangle = \delta\left(\bar{\imath}_k - \imath_k\right)$.

Proposition 2.15 *Le schéma de discrétisation (2.115) est identique au schéma proposé par Griebel (2.69),. . . ,(2.74).*

Preuve Le passage de la base hiérarchique à la base nodale s'effectue en appliquant la relation d'ondelette de l'AMR primale, *i.e.* appliquer le filtre $g\left(\bar{n}\right)$ aux coefficients de la représentation sur la base d'ondelettes. La transformation inverse s'effectue en appliquant la relation d'ondelette de l'AMR duale, *i.e.* appliquer le filtre $\tilde{g}(n)$ (voir la Proposition 1.16). ■

Remarque 2.15 *Si l'opérateur* $\mathcal{L}$ *n'agit pas dans la dimension* k, *l'opérateur de discrétisation est l'identité et commute avec le passage de hiérarchique à nodale. Nous pouvons alors montrer qu'en n'appliquant pas la relation d'échelle sur l'ondelette duale dans la dimension* k *à l'étape 1, le même schéma peut être retrouvé. Nous tiendrons compte de cette remarque dans la démonstration de la consistance de l'opérateur discret.*

L'équation (2.115) admet la représentation matricielle suivante :

$$\left(\left\langle \mathcal{L}u, \widetilde{\psi}_{\boldsymbol{\ell},\boldsymbol{\imath}}\right\rangle\right)_{\boldsymbol{\ell}\in\mathcal{I}_n,\boldsymbol{\imath}\in\tau_{\boldsymbol{\ell}}} \approx \mathcal{D}\left(\hat{u}_{\overline{\boldsymbol{\ell}},\overline{\boldsymbol{\imath}}}\right)_{\overline{\boldsymbol{\ell}}\in\mathcal{I}_n,\overline{\boldsymbol{\imath}}\in\tau_{\overline{\boldsymbol{\ell}}}}, \quad \hat{u}_{\overline{\boldsymbol{\ell}},\overline{\boldsymbol{\imath}}} = \left\langle u, \widetilde{\psi}_{\overline{\boldsymbol{\ell}},\overline{\boldsymbol{\imath}}}\right\rangle, \tag{2.116}$$

où

$$\mathcal{D}\left((\overline{\boldsymbol{\ell}},\overline{\boldsymbol{\imath}}),(\boldsymbol{\ell},\boldsymbol{\imath})\right) = \sum_{\tilde{\boldsymbol{n}}} \tilde{g}\left(\tilde{\boldsymbol{n}}\right) \sum_{\boldsymbol{k}\in\Gamma} \omega_{\boldsymbol{k}} \sum_{\boldsymbol{n}} g\left(\boldsymbol{n}\right) \left\langle \varphi_{\overline{\boldsymbol{\ell}}+\mathbf{1},2\overline{\boldsymbol{\imath}}+\boldsymbol{n}}, \widetilde{\varphi}_{\boldsymbol{\ell}+\mathbf{1},2\boldsymbol{\imath}+\tilde{\boldsymbol{n}}+\boldsymbol{k}}\right\rangle. \tag{2.117}$$

Nous étudions, à présent, l'erreur de consistance introduite par ce schéma de différences finies.

2.3.3.2 Démonstration de la consistance

Soit D la matrice définie par

$$\mathcal{D}\left((\overline{\boldsymbol{\ell}},\overline{\boldsymbol{\imath}}),(\boldsymbol{\ell},\boldsymbol{\imath})\right) = \sum_{\tilde{n},k,n} \tilde{g}\left(\tilde{n}\right) \omega_k g\left(n\right) \left\langle \varphi_{\overline{\ell}_1+1,2\overline{\imath}_1+n}, \widetilde{\varphi}_{\ell_1+1,2\imath+\tilde{n}+k}\right\rangle \delta_{\overline{\boldsymbol{\ell}}_{\mathbf{2}}}^{\boldsymbol{\ell}_{\mathbf{2}}} \delta_{\overline{\boldsymbol{\imath}}_{\mathbf{2}}}^{\boldsymbol{\imath}_{\mathbf{2}}}. \tag{2.118}$$

Théorème 2.16 *Soient $u \in \mathcal{C}^{(4,2,\ldots,2)}\left([0,1]^d\right)$ et $(\omega_k)_{k\in\Gamma}$ une discrétisation d'ordre 2 de l'opérateur ∂_x^2 alors*

$$\left\|\mathcal{P}\left(\partial_x^2 u\right) - D\left(\mathcal{P}(u)\right)\right\|_\infty \leq C n^{d-1} 2^{-2n} \|u\|_{\mathcal{C}^{(4,2\ldots,2)}([0,1]^d)}, \tag{2.119}$$

où n est le niveau de discrétisation de la grille sparse.

Preuve L'erreur de consistance du schéma est donnée par

$$\mathcal{E} = \max_{\boldsymbol{\ell}\in\mathcal{I}_n,\boldsymbol{\imath}\in\tau_{\boldsymbol{\ell}}} \left[\left| \left\langle \partial_{x_1}^2 u, \widetilde{\psi}_{\boldsymbol{\ell},\boldsymbol{\imath}}\right\rangle - \sum_{\overline{\boldsymbol{\ell}}\in\mathcal{I}_n,\overline{\boldsymbol{\imath}}\in\tau_{\overline{\boldsymbol{\ell}}}} \mathcal{D}\left((\overline{\boldsymbol{\ell}},\overline{\boldsymbol{\imath}}),(\boldsymbol{\ell},\boldsymbol{\imath})\right) \left\langle u, \widetilde{\psi}_{\overline{\boldsymbol{\ell}},\overline{\boldsymbol{\imath}}}\right\rangle \right|\right]. \tag{2.120}$$

Dans une première étape, nous intégrons deux fois par partie suivant les variables $x_2, \ldots, x_d$ et nous introduisons la fonction v définie par

$$v = \partial_{x_2}^2 \ldots \partial_{x_d}^2 u.$$

Le résultat (1.86) permet de montrer que :

$$\mathcal{E} = \max_{\substack{\boldsymbol{\ell}\in\mathcal{I}_n,\\ \boldsymbol{\imath}\in\tau_{\boldsymbol{\ell}}}} \left[2^{-2|\boldsymbol{\ell}_{\mathbf{2}}|_1} \left| \left\langle \partial_{x_1}^2 v, \widetilde{\psi}_{\ell_1,i_1}\psi_{\boldsymbol{\ell}_{\mathbf{2}},\boldsymbol{\imath}_{\mathbf{2}}}\right\rangle - \sum_{\overline{\boldsymbol{\ell}}\in\mathcal{I}_n,\overline{\boldsymbol{\imath}}\in\tau_{\overline{\boldsymbol{\ell}}}} \mathcal{D}\left((\overline{\boldsymbol{\ell}},\overline{\boldsymbol{\imath}}),(\boldsymbol{\ell},\boldsymbol{\imath})\right) \left\langle v, \widetilde{\psi}_{\ell_1,i_1}\psi_{\overline{\boldsymbol{\ell}}_{\mathbf{2}},\overline{\boldsymbol{\imath}}_{\mathbf{2}}}\right\rangle \right|\right]. \tag{2.121}$$

En reprenant la définition de $\mathcal{D}$:

$$\mathcal{E} = \max_{\substack{\boldsymbol{\ell}\in\mathcal{I}_n,\\ \boldsymbol{\imath}\in\tau_{\boldsymbol{\ell}}}} \left[2^{-2|\boldsymbol{\ell}_{\mathbf{2}}|_1} \left| \left\langle \left\langle \partial_{x_1}^2 v, \widetilde{\psi}_{\ell_1,i_1}\right\rangle - \sum_{\substack{\overline{\ell}\leq\tilde{\ell}\\ \overline{\imath}\in\tau_{\overline{\ell}}\\ \tilde{n},k,n}} \omega_k \chi_{\overline{\ell},\overline{\imath}}^{\ell_1,\imath_1}\left(\tilde{n},n,k\right) \left\langle v, \widetilde{\psi}_{\overline{\boldsymbol{\ell}},\overline{\imath}}\right\rangle, \psi_{\boldsymbol{\ell}_{\mathbf{2}},\boldsymbol{\imath}_{\mathbf{2}}}\right\rangle \right|\right], \tag{2.122}$$

où

$$\chi_{\bar{\ell},\bar{\imath}}^{\ell,\imath}(\tilde{n},n,k)=\tilde{g}(\tilde{n})\,g(n)\left\langle\varphi_{\bar{\ell}+1,2\bar{\imath}+n},\widetilde{\varphi}_{\ell+1,2\imath+\tilde{n}+k}\right\rangle.$$

Soit

$$\mathcal{E}\leq\max_{\substack{\boldsymbol{\ell}\in\mathcal{I}_n,\\ \boldsymbol{\imath}\in\tau_{\boldsymbol{\ell}}}}\left[2^{-2|\boldsymbol{\ell_2}|_1}\left\|\left\langle\partial_{x_1}^2 v,\widetilde{\psi}_{\ell_1,i_1}\right\rangle-\sum_{\substack{\bar{\ell}\leq\tilde{\ell}\\ \bar{\imath}\in\tau_{\bar{\ell}}\\ \tilde{n},k,n}}\omega_k\chi_{\bar{\ell},\bar{\imath}}^{\ell_1,\imath_1}(\tilde{n},n,k)\left\langle v,\widetilde{\psi}_{\bar{\ell},\bar{\imath}}\right\rangle\right\|_{\mathcal{C}^{\mathbf{0}}}\|\psi_{\boldsymbol{\ell_2},\boldsymbol{\imath_2}}\|_{L^1}\right]. \tag{2.123}$$

où la norme $\|\|_{\mathcal{C}^{\mathbf{0}}}$ est la norme des fonctions continues $\mathcal{C}^{\mathbf{0}}([0,1]^{d-1})$ appliquée à la fonction portant sur les variables $x_2,\ldots,x_d$. Nous définissons

$$\mathcal{E}_1=\left\langle\partial_{x_1}^2 v,\widetilde{\psi}_{\ell_1,i_1}\right\rangle-\sum_{\substack{0\leq\bar{\ell}\leq\tilde{\ell}\\ \bar{\imath}\in\tau_{\bar{\ell}}\\ \tilde{n},k,n}}\omega_k\chi_{\bar{\ell},\bar{\imath}}^{\ell_1,\imath_1}(\tilde{n},n,k)\left\langle v,\widetilde{\psi}_{\bar{\ell},\bar{\imath}}\right\rangle. \tag{2.124}$$

La proposition qui suit permet de conclure sur la démonstration du théorème 2.16. ■

Proposition 2.17

$$\|\mathcal{E}_1\|_{\mathcal{C}^{\mathbf{0}}}\leq C2^{-2\tilde{\ell}}\left(1+\log(n)^{d-1}\right)\|v\|_{\mathcal{C}^{(2,0\ldots,0)}}.$$

Preuve Soient les fonctions v_k définies par

$$v_k(x_1,\ldots,x_d)=v\left(x_1+k2^{-\tilde{\ell}},x_2,\ldots,x_d\right). \tag{2.125}$$

Alors

$$\mathcal{E}_1=\tilde{\mathcal{E}}_1+\sum_k\omega_k\mathcal{E}_1^k,$$

avec

$$\tilde{\mathcal{E}}_1=\left\langle\partial_{x_1}^2 v-\sum_k\omega_k v_k,\widetilde{\psi}_{\ell_1,i_1}\right\rangle, \tag{2.126}$$

et

$$\mathcal{E}_1^k=\left\langle v_k,\widetilde{\psi}_{\ell_1,i_1}\right\rangle-\sum_{\substack{0\leq\bar{\ell}\leq\tilde{\ell}\\ \bar{\imath}\in\tau_{\bar{\ell}}\\ \tilde{n},n}}\chi_{\bar{\ell},\bar{\imath}}^{\ell_1,\imath_1}(\tilde{n},n,k)\left\langle v,\widetilde{\psi}_{\bar{\ell},\bar{\imath}}\right\rangle. \tag{2.127}$$

Lemme 2.18 *L'erreur de consistance sur l'opérateur discret associé à $\partial_{x_1}^2$ permet de montrer que*

$$\left\|\tilde{\mathcal{E}}_1\right\|_{\mathcal{C}^{\mathbf{0}}}\leq C2^{-2\tilde{\ell}}\|v\|_{\mathcal{C}^{(2,0,\ldots,0)}}. \tag{2.128}$$

La majoration de $\left|\mathcal{E}_1^k\right|$ donnée par le lemme qui suit nous permet de conclure sur la démonstration de la proposition 2.17. ■

Lemme 2.19

$$\left|\mathcal{E}_1^k\right| \leq C 2^{-2\tilde{\ell}} \log(n)^{d-1} \|v\|_{\mathcal{C}^{(2,0,\ldots,0)}},$$

Preuve

$$\mathcal{E}_1^k = \left\langle v_k, \widetilde{\psi}_{\ell_1,i_1} \right\rangle - \sum_{\substack{0 \leq \bar{\ell} \leq \tilde{\ell} \\ \bar{\imath} \in \tau_{\bar{\ell}} \\ \tilde{n}}} \tilde{g}(\tilde{n}) \left\langle \left\langle v, \widetilde{\psi}_{\bar{\ell},\bar{\imath}} \right\rangle \psi_{\bar{\ell},\bar{\imath}}, \widetilde{\varphi}_{\ell_1+1,2i_1+\tilde{n}+k} \right\rangle. \tag{2.129}$$

$$\mathcal{E}_1^k = \sum_{\tilde{n}} \tilde{g}(\tilde{n}) \left(\langle v_k, \widetilde{\varphi}_{\ell_1+1,2i_1+\tilde{n}} \rangle - \left\langle \sum_{\substack{0 \leq \bar{\ell} \leq \tilde{\ell} \\ \bar{\imath} \in \tau_{\bar{\ell}}}} \left\langle v, \widetilde{\psi}_{\bar{\ell},\bar{\imath}} \right\rangle \psi_{\bar{\ell},\bar{\imath}}, \widetilde{\varphi}_{\ell_1+1,2i_1+\tilde{n}+k} \right\rangle \right). \tag{2.130}$$

Les termes de cette somme sont décomposés en deux quantités :

$$\mathcal{F}_{\tilde{n}}^k = \langle v_k, \widetilde{\varphi}_{\ell_1+1,2i_1+\tilde{n}} \rangle - \langle v, \widetilde{\varphi}_{\ell_1+1,2i_1+\tilde{n}+k} \rangle, \tag{2.131}$$

et

$$\tilde{\mathcal{F}}_{\tilde{n}}^k = \left\langle v - \sum_{\substack{0 \leq \bar{\ell} \leq \tilde{\ell} \\ \bar{\imath} \in \tau_{\bar{\ell}}}} \left\langle v, \widetilde{\psi}_{\bar{\ell},\bar{\imath}} \right\rangle \psi_{\bar{\ell},\bar{\imath}}, \widetilde{\varphi}_{\ell_1+1,2i_1+\tilde{n}+k} \right\rangle \tag{2.132}$$

Le premier terme est nul puisque, d'après la définition de k (2.112),

$$v_k\left(2^{-(\ell_1+1)}(2i+\tilde{n})\right) = v\left(2^{-(\ell_1+1)}(2i+\tilde{n}) + k2^{\tilde{\ell}}\right) = v\left(2^{-(\ell_1+1)}(2i+\tilde{n}+k)\right).$$

La majoration de $\tilde{\mathcal{F}}_{\tilde{n}}^k$ est obtenue en remarquant que $\widetilde{\varphi}$ est une masse de Dirac.

$$2^{-|\boldsymbol{\ell_2}|_1} \left\| \tilde{\mathcal{F}}_{\tilde{n}}^k \right\|_{\mathcal{C}^{\mathbf{0}}} \leq \sup_{\boldsymbol{x} \in \Omega} \left[\sum_{\boldsymbol{\ell} \notin \mathcal{I}_n, \boldsymbol{\imath} \in \tau_{\boldsymbol{\ell}}} \left\langle u, \widetilde{\psi}_{\boldsymbol{\ell},\boldsymbol{\imath}} \right\rangle \psi_{\boldsymbol{\ell},\boldsymbol{\imath}}(\boldsymbol{x}) \right]. \tag{2.133}$$

Nous en déduisons que $2^{-|\boldsymbol{\ell_2}|_1} \left\| \tilde{\mathcal{F}}_{\tilde{n}}^k \right\|_{\mathcal{C}^{\mathbf{0}}}$ est majoré par l'erreur de projection sur la grille sparse :

$$2^{-|\boldsymbol{\ell_2}|_1} \left\| \tilde{\mathcal{F}}_{\tilde{n}}^k \right\|_{\mathcal{C}^{\mathbf{0}}} \leq Cn \log(n)^{d-1} \|u\|_{\mathcal{C}^{(2,2,\ldots,2)}} = Cn \log(n)^{d-1} \|v\|_{\mathcal{C}^{(0,\ldots,0)}}.$$

■

2.4 Méthodes de Galerkin

Dans cette partie, nous présentons une méthode de Galerkin sur une base d'ondelettes pour le calcul des solutions de problèmes aux limites elliptiques. Nous verrons dans un chapitre ultérieur une application de cette méthode à des problèmes paraboliques et ultra-paraboliques. Le principe de cette méthode est directement issu de l'approche variationnelle. L'idée de base consiste, comme dans le cas des éléments finis, à remplacer l'espace de Hilbert sur lequel est posée la formulation variationnelle par un sous-espace de dimension finie. Le problème approché posé sur le sous-espace de dimension finie se ramène à la résolution d'un système linéaire dont la matrice est appelée matrice de rigidité. La méthode de Galerkin sur une base d'ondelettes sparse (*wavelet Galerkin with sparse tensor product*) consiste à choisir comme espace d'approximation l'espace V_n^0 de la définition 1.18. Les résultats donnés dans cette section se généralisent avec quelques modifications dans les ordres et hypothèses de convergence aux espaces sparse V_n^τ (voir Def 1.18).

Cette section comporte trois parties. Dans une première partie, nous rappelons la formulation variationnelle d'un problème elliptique ainsi que la formulation du problème approché associé. Dans une seconde partie, nous donnons les résultats sur la convergence de la méthode établis par Schwab [PS04]. Nous concluons sur l'aspect pratique (assemblage des matrices de rigidité et propriétés de celles-ci). Nous étudions également la compression de ces matrices et leur conditionnement.

2.4.1 Description de la méthode

Soit le domaine $\Omega = (0,1)^d$. Nous considérons le problème aux limites elliptique :

$$\begin{aligned} \mathcal{L}u &= f, \qquad \boldsymbol{x} \in \Omega, \\ u &= 0, \qquad \boldsymbol{x} \in \partial\Omega. \end{aligned} \tag{2.134}$$

L'opérateur $\mathcal{L}$ est défini par

$$\mathcal{L}u = -\nabla \cdot \boldsymbol{a}(\boldsymbol{x}) \nabla u + \boldsymbol{b}(\boldsymbol{x}) \cdot \nabla u + c(\boldsymbol{x}) u, \tag{2.135}$$

où les coefficients de la matrice symétrique semi-définie positive $\boldsymbol{a}$, du vecteur $\boldsymbol{b}$ et c appartiennent à $\mathcal{C}^0\left(\bar{\Omega}\right)$.

Formulation variationnelle Le passage de la *formulation forte* (2.134) à la *formulation variationnelle* sera étudié dans des chapitres ultérieurs, nous admettons le résultat pour l'instant. Dans le cas du problème aux limites (2.134), une *formulation variationnelle* est la suivante :
pour $f \in L^2(\Omega)$, trouver $u \in H_0^1(\Omega)$ telle que

$$a(u,v) = (f,v), \qquad \forall v \in H_0^1(\Omega), \tag{2.136}$$

où a est une forme bilinéaire sur $H^1(\Omega) \times H^1(\Omega)$ définie par

$$a(u,v) = \int_\Omega \left(\sum_{i,j=1}^{d} \boldsymbol{a}_{i,j}(\boldsymbol{x}) \frac{\partial u}{\partial x_j} \frac{\partial v}{\partial x_i} + \sum_{i=1}^{d} \boldsymbol{b}_i(\boldsymbol{x}) \frac{\partial u}{\partial x_i}\ v + c(\boldsymbol{x})\, u\ v \right) d\boldsymbol{x}. \tag{2.137}$$

et où $(\cdot,\cdot)$ est le produit scalaire de $L^2(\Omega)$,

$$(u,v)=\int_\Omega u\ v\ d\boldsymbol{x}.$$

Soit $\mathcal{V}=H^1(\Omega)$, nous supposons vérifiées les hypothèses suivantes :
– Il existe deux constantes $\overline{\gamma},\underline{\gamma}>0$ telles que, $\forall\ \boldsymbol{x}\in\overline{\Omega}$,

$$|a_{i,j}(\boldsymbol{x})|\leq\overline{\gamma}\quad\text{et}\quad\boldsymbol{\zeta}^T\boldsymbol{a}(\boldsymbol{x})\,\boldsymbol{\zeta}\geq\underline{\gamma}\,|\boldsymbol{\zeta}|^2\quad\forall\boldsymbol{\zeta}\in\mathbb{R}^d. \tag{2.138}$$

– Il existe deux constantes $\overline{c},\underline{c}$ telles que, $\forall\ \boldsymbol{x}\in\overline{\Omega}$,

$$\overline{c}\geq c(\boldsymbol{x})\geq\underline{c}>0. \tag{2.139}$$

– Le champ $\boldsymbol{b}$ vérifie

$$\sup_{\boldsymbol{x}\in\Omega}|\boldsymbol{b}(\boldsymbol{x})|<2\sqrt{\underline{\gamma}\underline{c}}. \tag{2.140}$$

Lemme 2.20 *Si les hypothèses (2.138,2.139,2.140) sont vérifiées, alors la forme bilinéaire a définie par (2.137) est :*

1. *continue sur $\mathcal{V}\times\mathcal{V}$,* i.e. *il existe une constante β telle que*

$$a(u,v)\leq\beta\|u\|_{\mathcal{V}}\|v\|_{\mathcal{V}},\quad\forall\ u,v\in\mathcal{V}. \tag{2.141}$$

2. *coercive,* i.e. *il existe une constante $\alpha>0$ telle que,*

$$a(u,u)\geq\alpha\|u\|_{\mathcal{V}}^2,\quad\forall\ u\in\mathcal{V}. \tag{2.142}$$

le théorème de Lax-Milgram permet de conclure à l'existence et à l'unicité de la solution de (2.136).

Formulation variationnelle discrète : Soit V_n^0 l'espace d'approximation sparse (Def.1.18) défini pour une AMR d'ordre p (c.-à-d. des ondelettes de degré $p-1$). L'approximation $\hat{u}_n\in V_n^0$ de la fonction u solution de (2.136) est définie comme la solution du problème discret :
trouver $\hat{u}_n\in V_n^0$ tel que

$$a(\hat{u}_n,\hat{v})=(f,\hat{v}),\quad\forall\hat{v}\in V_n^0. \tag{2.143}$$

Le théorème de Lax-Milgram assure l'existence et l'unicité de $\hat{u}$.

2.4.2 Convergence de la méthode

Afin de caractériser au mieux la convergence de l'approximation numérique, l'espace $\mathcal{X}_{\theta,s}(\Omega)$ est introduit.

Définition 2.6 (L'espace $\mathcal{X}_{\theta,s}(\Omega)$) *Soient $0\leq\theta\leq1$ et $s\geq1$, notons $\mathcal{X}_{\theta,s}(\Omega)$ l'espace de Hilbert obtenu par interpolation des espaces $H_0^1(\Omega)$ et $\mathcal{H}_0^{s+1}(\Omega)$:*

$$\mathcal{X}_{\theta,s}(\Omega)=\left[H_0^1(\Omega),\mathcal{H}_0^{s+1}(\Omega)\right]_\theta, \tag{2.144}$$

muni de la norme $\|\ \|_{\mathcal{X}_{\theta,s}(\Omega)}$. Cette norme vérifie la propriété

$$\|u\|_{\mathcal{X}_{\theta,s}(\Omega)}\leq\|u\|_{H^1(\Omega)}^{1-\theta}\|u\|_{\mathcal{H}^{s+1}(\Omega)}^{\theta}\quad\forall u\in\mathcal{H}_0^{s+1}(\Omega). \tag{2.145}$$

Proposition 2.21 (Erreur d'approximation en norme H^1) *Soit u la solution de (2.136), si $u \in \mathcal{X}_{\theta,s}(\Omega)$ avec $p \geq s+1$ alors*

$$\|u - \hat{u}\|_{H^1(\Omega)} \leq Ch^{\theta s}\|u\|_{\mathcal{X}_{\theta,s}(\Omega)}. \tag{2.146}$$

Preuve La discrétisation (2.143) du problème (2.136) est conforme (voir [BMR04]). Elle vérifie donc les hypothèses du Lemme de Céa. La continuité et la coercivité de a permettent de conclure :

$$\|u - \hat{u}\|_X \leq \frac{\beta}{\alpha} \inf_{\hat{v} \in V_n^0} \|u - \hat{v}\|_{H_0^1(\Omega)}. \tag{2.147}$$

Le résultat de projection sur V_n^0 (proposition 1.25) permet de montrer :

$$\|u - \hat{u}\|_{H^1(\Omega)} \leq C\|u\|_{H^1(\Omega)} \quad \text{et} \quad \|u - \hat{u}\|_{H^1(\Omega)} \leq Ch^s\|u\|_{\mathcal{H}_0^{s+1}(\Omega)} \text{ si } u \in \mathcal{H}_0^{s+1}(\Omega). \tag{2.148}$$

Nous en déduisons par interpolation :

$$\|u - \hat{u}\|_{H^1(\Omega)} \leq Ch^{\theta s}\|u\|_{\mathcal{X}_{\theta,s}(\Omega)}. \tag{2.149}$$

■

Proposition 2.22 (Erreur d'approximation en norme $L^2(\Omega)$) *Soit $u \in \mathcal{X}_{\theta,s}(\Omega)$ avec $p \geq s+1$, alors*

$$\|u - \hat{u}\|_{L^2(\Omega)} \leq Ch^{s(\theta+\delta(s))}\|u\|_{\mathcal{X}_{\theta,s}(\Omega)}, \quad \textit{avec} \quad \delta(s) = \frac{1}{d(s+1)-1}. \tag{2.150}$$

Preuve Rappelons les éléments de la preuve par dualité proposée par Schwab [PS04].

$$\|u - \hat{u}\|_{L^2(\Omega)} = \sup_{v \in L^2(\Omega)} \frac{(u - \hat{u}, v)}{\|v\|_{L^2(\Omega)}} = \sup_{v \in L^2(\Omega)} \frac{a(u - \hat{u}, w - \hat{w})}{\|v\|_{L^2(\Omega)}}, \tag{2.151}$$

où w est définie comme la solution de l'équation

$$\mathcal{L}^\star w = v, \quad w_{|\partial\Omega} = 0, \tag{2.152}$$

avec $\mathcal{L}^\star$ est l'opérateur adjoint de $\mathcal{L}$. Notons $\hat{\omega}$ la solution du problème discret associé à (2.152). Nous aurons besoin des trois points suivant.

- Si $v \in H^1(\Omega)$ alors $w \in H^2(\Omega)$ et

$$\|w\|_{H^2(\Omega)} \leq C\|v\|_{L^2(\Omega)},$$

- La proposition 1.25 permet de montrer

$$\|w - \hat{w}\|_{H^1(\Omega)} \leq Ch^{\theta' s}\|w\|_{\mathcal{X}_{\theta',s}(\Omega)}, \tag{2.153}$$

  où θ' est à choisir dans $[0,1]$, (voir ci-dessous).
- Les propriétés de l'interpolation réel et des inclusions, $H_0^1(\Omega) \subset \mathcal{X}_{0,s}(\Omega)$, $H_0^1(\Omega) \cap H^{d(s+1)}(\Omega) \subset \mathcal{H}_0^{s+1}(\Omega) = \mathcal{X}_{1,s}(\Omega)$ permettent de montrer que :

$$H_0^1(\Omega) \cap H^{1-\theta'+\theta' d(s+1)}(\Omega) \subset \mathcal{X}_{\theta',s}(\Omega), \quad 0 \leq \theta' \leq 1.$$

En conclusion, si $\theta' = \delta(s) = \frac{1}{d(s+1)-1}$, (2.153) devient

$$\|w - \hat{w}\|_{H^1(\Omega)} \leq Ch^{\delta(s)s}\|w\|_{H^2(\Omega)} \leq Ch^{\delta(s)s}\|v\|_{L^2(\Omega)}, \tag{2.154}$$

et (2.151) implique

$$\|u - \hat{u}\|_{L^2(\Omega)} \leq \|u - \hat{u}\|_{H^1(\Omega)} \sup_{v \in L^2(\Omega)} \frac{\|w - \hat{w}\|_{H^1(\Omega)}}{\|v\|_{L^2(\Omega)}} \leq Ch^{\delta(s)s}\|u - \hat{u}\|_{H^1(\Omega)}. \tag{2.155}$$

Le résultat se déduit de (2.149) et (2.155). ■

Remarque 2.16 *L'approximation du problème (2.136) sur l'espace sparse V_n^0 donne le même taux de convergence en norme H^1 que l'approximation sur la grille pleine V_n^∞ c.-à-d. $\|u - \hat{u}\|_{H^1(\Omega)} \leq Ch^s\|u\|_{\mathcal{H}^{s+1}(\Omega)}$, à condition que la fonction u soit suffisamment régulière (*i.e.*$u \in \mathcal{H}_0^{s+1}(\Omega)$). La convergence $L^2(\Omega)$ est ralentie : nous obtenons un facteur $\left(h^{s(1+\delta(s))}\right)$ au lieu d'un facteur $O\left(h^{2s}\right)$.*

2.4.3 Mise en oeuvre

La mise en oeuvre de la méthode de Galerkin sur une base d'ondelettes sparse est décrite dans ce paragraphe.

2.4.3.1 Système linéaire

Rappelons brièvement le cadre général de la méthode de Galerkin. La méthode consiste à discrétiser le problème (2.136) sur la base une base $(\psi_{\boldsymbol{\ell},\boldsymbol{\imath}})_{(\boldsymbol{\ell}\in\mathcal{I}_n,\boldsymbol{\imath}\in\boldsymbol{I}_{\boldsymbol{\ell}})}$ de l'espace V_n^0. Les coefficients de la matrice de rigidité sont donnés par :

$$\boldsymbol{A}_{(\boldsymbol{\ell},\boldsymbol{\imath}),(\bar{\boldsymbol{\ell}},\bar{\boldsymbol{\imath}})} = a\left(\psi_{\bar{\boldsymbol{\ell}},\bar{\boldsymbol{\imath}}}, \psi_{\boldsymbol{\ell},\boldsymbol{\imath}}\right). \tag{2.156}$$

Comme pour les méthodes d'éléments finis, une fois construite cette matrice de rigidité, la solution du problème discrétisé (2.143) est obtenue en résolvant le système linéaire

$$\left(\boldsymbol{A}_{(\boldsymbol{\ell},\boldsymbol{\imath}),(\bar{\boldsymbol{\ell}},\bar{\boldsymbol{\imath}})}\right)_{\substack{(\boldsymbol{\ell},\bar{\boldsymbol{\ell}})\in\mathcal{I}_n\times\mathcal{I}_n,\\(\boldsymbol{\imath},\bar{\boldsymbol{\imath}})\in\tau_{\boldsymbol{\ell}}\times\tau_{\bar{\boldsymbol{\ell}}}}} \left(u_{\bar{\boldsymbol{\ell}},\bar{\boldsymbol{\imath}}}\right)_{\substack{\bar{\boldsymbol{\ell}}\in\mathcal{I}_n\\\bar{\boldsymbol{\imath}}\in\tau_{\bar{\boldsymbol{\ell}}}}} = \left(f_{\boldsymbol{\ell},\boldsymbol{\imath}}\right)_{\substack{\boldsymbol{\ell}\in\mathcal{I}_n\\\boldsymbol{\imath}\in\tau_{\boldsymbol{\ell}}}}, \tag{2.157}$$

où

$$f_{\boldsymbol{\ell},\boldsymbol{\imath}} = (f, \psi_{\boldsymbol{\ell},\boldsymbol{\imath}}), \tag{2.158}$$

et la solution approchée u_n est donnée par :

$$u_n = \sum_{\boldsymbol{\ell}\in\mathcal{I}_n,\boldsymbol{\imath}\in\tau_{\boldsymbol{\ell}}} u_{\boldsymbol{\ell},\boldsymbol{\imath}}\psi_{\boldsymbol{\ell},\boldsymbol{\imath}}. \tag{2.159}$$

Remarque 2.17 *Les bases d'ondelettes ne sont pas des bases locales contrairement à la base nodale des méthodes d'éléments finis. La valeur de u_n en un point dépend d'un nombre important de coefficients $u_{\boldsymbol{\ell},\boldsymbol{\imath}}$. L'évaluation de la fonction u_n au point x_0 est donc beaucoup plus coûteuse que dans le cas d'une méthode d'élément finis.*

A ce stade, la résolution du problème (2.143) nécessite la construction de la matrice $\boldsymbol{A}$ donnée par (2.156) et du second membre (2.158). Ces calculs dépendent du choix des fonctions de base. La construction d'une *Sparse Grid* n'est pas liée au choix de la famille d'ondelettes. Cependant, les ondelettes B-spline biorthogonales constituent une famille adaptée aux méthodes de Galerkin. Elles possèdent les deux propriétés suivantes, qui rendent les calculs de $\boldsymbol{A}$ et du second membre réalisables :

1. leur support est borné,
2. l'expression analytique de la fonction mère est connue.

Remarque 2.18 *Le choix des ondelettes de bord constitue une étape non triviale dans la mise en oeuvre d'une méthode de Galerkin sur une base d'ondelettes. Nous nous référons à [Mas99] pour le choix de ces fonctions. Une fois ces ondelettes choisies, les changements dans les méthodes proposées par la suite sont seulement techniques, c'est pourquoi nous n'abordons pas ces questions dans ce qui suit.*

2.4.3.2 La matrice de rigidité

Nous présentons différentes méthodes pour la construction de la matrice de rigidité $\boldsymbol{A}$. La principale difficulté réside dans l'utilisation de fonctions de base non localisées. L'approximation numérique par des formules de quadrature utilisées pour des méthodes d'éléments finis n'est donc pas applicable.

La méthode présenté ici s'applique à un opérateur dont les coefficients sont à variables séparées (ou dont les coefficients sont des sommes de telles fonctions). Soit $\mathcal{L}$ un de ces opérateurs, prenons par exemple :

$$\mathcal{L} = -\sum_{i=1}^{d}\sum_{j=1}^{d} a_{i,j}(\boldsymbol{x})\frac{\partial}{\partial x_j}\frac{\partial}{\partial x_i} + \sum_{i=1}^{d} b_i(\boldsymbol{x})\frac{\partial}{\partial x_i} + c(\boldsymbol{x}), \tag{2.160}$$

avec

$$a_{i,j}(\boldsymbol{x}) = \prod_{k=1}^{d} a_{i,j}^{k}(x_k) \quad \text{et} \quad b_i(\boldsymbol{x}) = \prod_{k=1}^{d} b_i^{k}(x_k), \tag{2.161}$$

alors l'opérateur se décompose en une somme d'opérateurs $\mathcal{L}_e$

$$\mathcal{L}_e u(\boldsymbol{x}) = \alpha_1(x_1)\ \alpha_2(x_2)\ \alpha_d(x_d)\,\frac{\partial^{|\boldsymbol{m}|}u}{\partial x_1^{m_1}\ldots\partial x_d^{m_d}}(\boldsymbol{x}), \quad \text{avec} \quad 0 \le |m_k| \le 2. \tag{2.162}$$

Remarque 2.19 *Un exemple d'opérateur n'entrant pas dans ce cadre est donné par*

$$|x_2 - x_3|\,\frac{\partial^2 u}{\partial x_1^2}.$$

Nous utiliserons la propriété suivante de la forme bilinéaire a_e (associée à $\mathcal{L}_e$) : dans le cas de deux fonctions u,v de la forme $u(\boldsymbol{x}) = u_1(x_1)\ldots u_d(x_d)$, $a_e(u,v)$ s'écrit

$$a_e(u,v) = \prod_{k=1}^{d} a_e^{k}(u_k, v_k). \tag{2.163}$$

Les fonctions de base $\psi_{\ell,\imath}$ étant, par construction, à variables séparées, les coefficients de la matrice $\boldsymbol{A}$ (associée à a_e) sont obtenus comme produit de coefficients de matrices de rigidité associées à des opérateurs différentiels sur $[0,1]$.

$$\boldsymbol{A}_{(\boldsymbol{\ell},\boldsymbol{\imath}),(\bar{\boldsymbol{\ell}},\bar{\boldsymbol{\imath}})} = \prod_{k=1}^{d} A^k_{(\ell_k,\imath_k)(\bar{\ell}_k,\bar{\imath}_k)}, \quad \boldsymbol{\ell},\bar{\boldsymbol{\ell}} \in \mathcal{I}_n^0,\ \boldsymbol{\imath} \in \tau_{\boldsymbol{\ell}},\ \bar{\boldsymbol{\imath}} \in \tau_{\bar{\boldsymbol{\ell}}}, \tag{2.164}$$

avec

$$A^k_{(\ell,\imath),(\bar{\ell},\bar{\imath})} = a_e^k\left(\psi_{\bar{\ell},\bar{\imath}}, \psi_{\ell,\imath}\right). \tag{2.165}$$

Soit $A^k_{\ell,\bar{\ell}}$ le bloc de la matrice de rigidité associée à a_e^k correspondant aux niveaux $\ell,\ \bar{\ell}$:

$$A^k_{\ell,\bar{\ell}} = \left(A^k_{(\ell,\imath),(\bar{\ell},\bar{\imath})}\right)_{\imath\in\tau_\ell,\ \bar{\imath}\in\tau_{\bar{\ell}}}.$$

La matrice $\boldsymbol{A}$ est constituée des blocs $\boldsymbol{A}_{\boldsymbol{\ell},\bar{\boldsymbol{\ell}}}$ correspondant aux multi-indices $\boldsymbol{\ell},\bar{\boldsymbol{\ell}}$, qui sont eux-même construits par produit tensoriel :

$$\boldsymbol{A}_{\boldsymbol{\ell},\bar{\boldsymbol{\ell}}} = \bigotimes_{k=1}^{d} A^k_{\ell_k,\bar{\ell}_k}. \tag{2.166}$$

Si l'opérateur aux dérivées partielles est somme d'opérateurs élémentaires comme ci-dessus, la matrice de rigidité s'obtient comme somme de matrices dont les blocs sont donnés par (2.166).

Forme bilinéaire sur $[0,1]$ Nous abordons dans ce paragraphe le calcul effectif des coefficients de la matrice de rigidité associée à une forme bilinéaire a définie sur un espace $\mathcal{V}(\Omega)$ où $\Omega \subset [0,1]^d$.

Notons que, dans le cas d'un opérateur différentiel à coefficients constants, il est possible de trouver une formule par un calcul direct. Il est également possible d'utiliser la transformée de Fourier et l'identité de Parseval pour calculer les coefficients de la matrice de rigidité. Nous reviendrons ultérieurement sur cette méthode dans le cas d'un opérateur intégral à coefficients constants.

Soit une forme bilinéaire a définie pour tout $u, v \in \mathcal{V}(\Omega)$ par :

$$a(u,v) = \int_\Omega \alpha(x)\, u^{(m_1)}(x) v^{(m_2)}(x) dx, \tag{2.167}$$

où $u^{(m_1)}$ représente la dérivée d'ordre m_1 de u et $\alpha \in \mathcal{C}^0(\Omega)$. A nouveau, nous supposons que, si la forme bilinéaire n'est pas directement de la forme (2.167), elle peut néanmoins s'écrire comme une somme de telles formes bilinéaires.

Les coefficients de la matrice de rigidité sont obtenus en calculant $a\left(\psi_{\ell,\imath}, \psi_{\bar{\ell},\bar{\imath}}\right)$.

Rappelons que la taille du support des ondelettes $(\psi_{\ell,\imath})_{\ell\in\mathcal{I}_n,\imath\in\tau_\ell}$ ne nous permet pas d'utiliser directement des formules de quadrature pour calculer $a\left(\psi_{\ell,\imath}, \psi_{\bar{\ell},\bar{\imath}}\right)$. Il est toutefois possible d'appliquer des méthodes de quadrature à condition d'adapter le nombre de points à la taille du support de l'ondelette.

Nous proposons deux autres méthodes.

La première consiste à utiliser les relations d'échelle pour se ramener à un cadre connu. L'idée de cette méthode est présentée dans [DM93]. Nous la décrivons ci-dessous :

– Nous appliquons la relation d'échelle sur les ondelettes,

$$a\left(\psi_{\ell,\imath},\psi_{\bar{\ell},\bar{\imath}}\right)=\sum_{n,\bar{n}}g(n)g(\bar{n})a\left(\varphi_{\ell+1,2\imath+n},\varphi_{\bar{\ell}+1,2\bar{\imath}+\bar{n}}\right). \tag{2.168}$$

– Nous considérons le schéma de récurrence qui consiste à appliquer la relation (2.169) tant que $\ell\neq\bar{\ell}$:

$$a\left(\varphi_{\ell,\imath},\varphi_{\bar{\ell},\bar{\imath}}\right)=\begin{cases}\displaystyle\sum_{\bar{n}}h(\bar{n})a\left(\varphi_{\ell,\imath},\varphi_{\bar{\ell}+1,2\bar{\imath}+\bar{n}}\right) & \text{si } \ell>\bar{\ell},\\ \displaystyle\sum_{n}h(n)a\left(\varphi_{\ell+1,2i+n},\varphi_{\bar{\ell},\bar{i}}\right) & \text{sinon .}\end{cases} \tag{2.169}$$

– Nous nous ramenons au calcul des coefficients de la matrice de rigidité sur la base nodale. A la fin de cette étape, soit nous disposons d'une formule analytique pour calculer $a\left(\varphi_{\ell,\imath},\varphi_{\ell,\bar{\imath}}\right)$, soit nous répétons la relation sur les filtres d'échelles pour se ramener à un niveau sur lequel les méthodes de quadrature sont acceptables.

Cette précédente s'applique à tous les opérateurs elliptiques et se généralise à certains opérateurs intégraux.

La seconde méthode est fondée sur l'idée suivante : faire apparaître la dérivée d'ordre p de l'ondelette, en intégrant par partie. Cette relation est intéressante dans la mesure où $\psi^{(p)}_{\ell,\imath}$ est un peigne de Dirac.

Algorithme 2.1 (Calcul analytique des coefficients de la matrice de rigidité) *Supposons que la $p-m_1$ème primitive de $\alpha\,\psi^{(m_2)}_{\bar{\ell},\bar{\imath}}$ existe, elle est notée $D^{-(p-m_1)}\left(\alpha\,\psi^{(m_2)}_{\bar{\ell},\bar{\imath}}\right)$, alors*

$$a\left(\psi_{\ell,\imath},\psi_{\bar{\ell},\bar{\imath}}\right)=(-1)^{p-m_1}\int_{\Omega}\psi^{(p)}_{\ell,\imath}(x)D^{-(p-m_1)}\left(\alpha\,\psi^{(m_2)}_{\bar{\ell},\bar{\imath}}\right)(x)dx. \tag{2.170}$$

En remarquant que $\psi^{(p)}_{\ell,\imath}=\sum_{n}g_n\delta_{x^n_{\ell,\imath}}$, nous obtenons une formule exacte et analytique pour le calcul du coefficient de la matrice de rigidité.

Cette idée n'est pas naturelle dans le cadre des méthodes de Galerkin. Sur le cas simple de l'équation de Poisson, elle consiste à revenir à la définition de la forme bilinéaire $a(u,v)=\langle-\Delta u,v\rangle$ au lieu de considérer la forme symétrique $(\nabla u,\nabla v)$. Cette méthode est beaucoup plus performante en termes de temps de calcul. En particulier, dans le cas d'une architecture du code sans stockage § 9.3.3, cette méthode s'avère être très efficace.

Nous présentons les détails de la méthode appliquée au calcul des coefficients de la matrice de masse avec une base d'ondelettes biorthogonales $(p,\tilde{p})=(2,2)$.

$$\begin{aligned}a\left(\psi_{\ell,\imath},\psi_{\bar{\ell},\bar{\imath}}\right)=\int_{\Omega}\psi_{\ell,\imath}(x)\psi_{\bar{\ell},\bar{\imath}}(x)dx&=2^{(\ell+\bar{\ell})/2}\int_{\Omega}\psi\left(2^{\ell}x-\imath\right)\psi\left(2^{\bar{\ell}}x-\bar{\imath}\right)dx\\&=2^{5/2\ell-3/2\bar{\ell}}\int_{\Omega}\psi^{(2)}\left(2^{\ell}x-\imath\right)\psi^{(-2)}\left(2^{\bar{\ell}}x-\bar{\imath}\right)dx\end{aligned}$$

Soit le changement de variable $z : x \to 2^{\ell} x - \imath$, en posant $q = \ell - \bar{\ell}$ et $k = \bar{\imath} - 2^{-q}\imath$, nous obtenons

$$a\left(\psi_{\ell,\imath}, \psi_{\bar{\ell},\bar{\imath}}\right) = 2^{3q/2} \int_{\Omega} \psi^{(2)}(z)\, \psi^{(-2)}\left(2^{-q} y - k\right) dy = 2^{3q/2} \sum_{n} g_n \psi^{(-2)}\left(2^{-q}\left(x_{\ell,\imath}^{n} - k\right)\right),$$

où les points $x_{\ell,\imath}^{n}$ sont les points de discontinuité de la dérivée de la fonction d'ondelette ψ.

2.4.3.3 Calcul du second membre

Ce paragraphe fait l'objet d'un point critique pour l'utilisation des méthodes de Galerkin sur une base sparse, à savoir la projection d'une fonction $f \in L^2(\Omega)$ sur l'espace Sparse V_n engendré par les fonctions de base $(\psi_{\boldsymbol{\ell},\boldsymbol{\imath}})_{\boldsymbol{\ell} \in \mathcal{I}_n,\ \boldsymbol{\imath} \in \tau_{\boldsymbol{\ell}}}$.

Nous reviendrons ultérieurement sur les propriétés d'approximation de cette projection en fonction de la régularité de f. Nous discutons ici de la mise en oeuvre pratique.

Pour chacune des fonctions de base, nous devons calculer

$$b_{\boldsymbol{\ell},\boldsymbol{\imath}}(f) = \int_{\mathbb{R}^d} f(\boldsymbol{x})\, \psi_{\boldsymbol{\ell},\boldsymbol{\imath}}(\boldsymbol{x})\, d\boldsymbol{x}, \quad \boldsymbol{\ell} \in \mathcal{I}_n,\ \boldsymbol{\imath} \in \tau_{\boldsymbol{\ell}}. \tag{2.171}$$

Dans le cas d'une fonction f à variables séparées, il est possible de conduire les calculs.

Proposition 2.23 *Si la fonction f est à variables séparées, alors $b_{\boldsymbol{\ell},\boldsymbol{\imath}}$ est le produit des*

$$b_{\ell_k,\imath_k}(f) = \int_{\mathbb{R}} f_k(x)\, \psi_{\ell,\imath}(x)\, dx. \tag{2.172}$$

Le facteur $b_{\ell_k,\imath_k}(f)$ se calcule en se ramenant sur la base nodale à l'aide des équations d'échelle, de la même façon que ce qui est fait pour les coefficients de la matrice de rigidité.

Dans le cas d'une fonction f quelconque, la méthode naïve consiste à se ramener par application de la relation d'échelle à un calcul sur la base nodale pleine. Elle est, par conséquent, trop coûteuse. Pour nos applications en mathématiques financières, les fonctions considérées sont de la forme $f_1(\boldsymbol{x}) = \left(K - \sum_{i=1}^{d} x_i\right)^+$ ou $f_2(\boldsymbol{x}) = \delta_{\left(K - \sum_{i=1}^{d} x_i\right)^+}$. La première fonction correspond au payoff d'un put sur panier. Les singularités des fonctions f_1 et f_2 sont locales : des résultats sur l'approximation non-linéaire (voir § 1.2.1.5) s'appliquent.

Remarquons que le calcul de $b_{\boldsymbol{\ell},\boldsymbol{\imath}}(f_2)$ revient à intégrer $\psi_{\boldsymbol{\ell},\boldsymbol{\imath}}$ sur une hypersurface de dimension $d-1$. Trouver une formulation analytique des $\tilde{b}_{\boldsymbol{\ell},\boldsymbol{\imath}}(f_2)$ semble possible mais très difficile.

Nous présentons ci-dessous une méthode permettant de calculer $b_{\boldsymbol{\ell},\boldsymbol{\imath}}(f)$ pour une fonction f a priori quelconque avec une majoration de l'erreur sur le calcul de $b_{\boldsymbol{\ell},\boldsymbol{\imath}}(f)$ en fonction de la régularité de f. Nous adopterons cette méthode dans le chapitre 8 pour le calcul de $b_{\boldsymbol{\ell},\boldsymbol{\imath}}(f_1)$.

Une première remarque permet de simplifier les termes à calculer : il est à nouveau possible d'appliquer la relation d'échelle sur les ondelettes pour se ramener au calcul de

$$\tilde{b}_{\boldsymbol{\ell},\boldsymbol{\imath}}(f) = \int_{\mathbb{R}^d} f(\boldsymbol{x})\, \varphi_{\boldsymbol{\ell},\boldsymbol{\imath}}(\boldsymbol{x})\, d\boldsymbol{x}. \tag{2.173}$$

Notons que, dans le cas des ondelettes B-splines biorthogonales d'ordre 1, la fonction d'échelle $\varphi_{\ell+1,2\imath+1}$ correspond à l'ondelette interpolante $\psi^\star_{\ell,\imath}$.

Une méthode d'approximation des coefficients $\tilde{b}_{\boldsymbol{\ell},\boldsymbol{\imath}}(f)$ est ici proposée dans le cas d'une fonction $f \in \mathcal{H}^1(\Omega)$. Les coefficients approchés sont notés $\tilde{b}^\star_{\boldsymbol{\ell},\boldsymbol{\imath}}(f)$. Ce coefficient est calculé à partir des coefficients obtenus par interpolation de la fonction f. L'approximation $\tilde{b}^\star_{\boldsymbol{\ell},\boldsymbol{\imath}}(f)$ dépend donc de l'ensemble des points d'interpolation, noté $\mathcal{J}$. Nous décrivons la méthode dans le cas d'un ensemble quelconque puis nous donnons un résultat de consistance dans le cas où $\mathcal{J}$ correspond à l'ensemble des points de la grille sparse.

Algorithme 2.2 (Approximation par interpolation de la projection sur V_n^0)

- *Les coefficients $\tilde{b}^\star_{\boldsymbol{\ell},\boldsymbol{\imath}}(f)$ sont calculés pour les multi-indices tels que les i_k sont tous impairs à l'aide de la relation :*

$$\tilde{b}^\star_{\boldsymbol{\ell}+\mathbf{1},2\boldsymbol{\imath}+\mathbf{1}}(f) \approx \sum_{\overline{\boldsymbol{\ell}}\in\mathcal{J},\ \overline{\boldsymbol{\imath}}\in\tau_{\overline{\boldsymbol{\ell}}}} M_{(\boldsymbol{\ell},\boldsymbol{\imath}),(\overline{\boldsymbol{\ell}},\overline{\boldsymbol{\imath}})} \hat{b}_{\overline{\boldsymbol{\ell}},\overline{\boldsymbol{\imath}}}(f), \tag{2.174}$$

  où

$$M_{(\boldsymbol{\ell},\boldsymbol{\imath}),(\overline{\boldsymbol{\ell}},\overline{\boldsymbol{\imath}})} = \left\langle \psi^\star_{\overline{\boldsymbol{\ell}},\overline{\boldsymbol{\imath}}}, \psi^\star_{\boldsymbol{\ell},\boldsymbol{\imath}} \right\rangle \quad \textit{et} \quad \hat{b}_{\overline{\boldsymbol{\ell}},\overline{\boldsymbol{\imath}}}(f) = \left\langle f, \psi^\star_{\overline{\boldsymbol{\ell}},\overline{\boldsymbol{\imath}}} \right\rangle. \tag{2.175}$$

  Le calcul de $\left\langle f, \widetilde{\psi}^\star_{\boldsymbol{\ell},\boldsymbol{\imath}} \right\rangle$ est obtenu par interpolation d'après la proposition 1.19.
- *Dans le cas où les m indices $\imath_{k_1}, \ldots, \imath_{k_m}$ sont pairs, la relation suivante est appliquée dans chacune des dimensions k_j :*

$$\tilde{b}^\star_{\boldsymbol{\ell},\boldsymbol{\imath}}(f) = \tilde{b}^\star_{(\ell_1,\ldots,\ell_k-1,\ldots,\ell_d),(\imath_1,\ldots,\imath_k/2,\ldots\imath_d)}(f) - \frac{1}{2}\left(\tilde{b}^\star_{\boldsymbol{\ell},(\imath_1,\ldots,\imath_k-1,\ldots\imath_d)}(f) + \tilde{b}^\star_{\boldsymbol{\ell},(\imath_1,\ldots,\imath_k+1,\ldots\imath_d)}(f)\right), \tag{2.176}$$

  ce qui nous ramène au cas précédent.

Remarque 2.20 *Si nous appliquons la méthode de Galerkin à une base d'ondelettes B-splines d'ordre supérieur (*i.e. $p > 2, \tilde{p} = 0$*), alors la relation (2.174) est conservée. La relation (2.176) est en revanche plus compliquée puisqu'elle fait intervenir un filtre plus large.*

Preuve La relation algébrique (2.174) est obtenue à partir de la décomposition de f sur la base des ondelettes interpolantes :

$$\begin{aligned} \tilde{b}_{\boldsymbol{\ell},2\boldsymbol{\imath}+1}(f) &= \int_{\mathbb{R}^d} f(\boldsymbol{x})\, \psi^\star_{\boldsymbol{\ell},\boldsymbol{\imath}}(\boldsymbol{x})\, d\boldsymbol{x} \\ &\approx \int_{\mathbb{R}^d} \sum_{\overline{\boldsymbol{\ell}}\in\mathcal{J},\ \overline{\boldsymbol{\imath}}\in\tau_{\overline{\boldsymbol{\ell}}}} \left\langle f, \widetilde{\psi}^\star_{\overline{\boldsymbol{\ell}},\overline{\boldsymbol{\imath}}} \right\rangle \psi^\star_{\overline{\boldsymbol{\ell}},\overline{\boldsymbol{\imath}}} \psi^\star_{\boldsymbol{\ell},\boldsymbol{\imath}}(\boldsymbol{x})\, d\boldsymbol{x} \\ &\approx \sum_{\overline{\boldsymbol{\ell}}\in\mathcal{J},\ \overline{\boldsymbol{\imath}}\in\tau_{\overline{\boldsymbol{\ell}}}} \left\langle f, \widetilde{\psi}^\star_{\overline{\boldsymbol{\ell}},\overline{\boldsymbol{\imath}}} \right\rangle \left\langle \psi^\star_{\overline{\boldsymbol{\ell}},\overline{\boldsymbol{\imath}}} \psi^\star_{\boldsymbol{\ell},\boldsymbol{\imath}} \right\rangle. \end{aligned} \tag{2.177}$$

L'équation (2.176) s'obtient à partir de la relation d'échelle sur les ondelettes interpolantes (1.80). ■

Étudions l'erreur liée à cette approximation. Soit $\mathcal{J} = \mathcal{I}_n$, nous considérons le cas où les $\boldsymbol{\imath}$ sont tous impairs :

$$\begin{aligned}\mathcal{E} &= \sup_{\boldsymbol{\ell}\in\mathcal{I}_n,\boldsymbol{\imath}\in\tau_{\boldsymbol{\ell}}} \left|\tilde{b}_{\boldsymbol{\ell}+\mathbf{1},2\boldsymbol{\imath}+\mathbf{1}}(f) - \tilde{b}^{\star}_{\boldsymbol{\ell}+\mathbf{1},2\boldsymbol{\imath}+\mathbf{1}}(f)\right| \qquad (2.178)\\ &\leq \sup_{\boldsymbol{\ell}\in\mathcal{I}_n,\boldsymbol{\imath}\in\tau_{\boldsymbol{\ell}}} \left|\left\langle f - \sum_{\overline{\boldsymbol{\ell}}\in\mathcal{I}_n,\overline{\boldsymbol{\imath}}\in\tau_{\overline{\boldsymbol{\ell}}}} \left\langle f, \widetilde{\psi}^{\star}_{\overline{\boldsymbol{\ell}},\overline{\boldsymbol{\imath}}}\right\rangle \psi^{\star}_{\overline{\boldsymbol{\ell}},\overline{\boldsymbol{\imath}}}, \psi^{\star}_{\ell,\imath}\right\rangle\right|\\ &\leq \left\| f - \sum_{\overline{\boldsymbol{\ell}}\in\mathcal{I}_n,\overline{\boldsymbol{\imath}}\in\tau_{\overline{\boldsymbol{\ell}}}} \left\langle f, \widetilde{\psi}^{\star}_{\overline{\boldsymbol{\ell}},\overline{\boldsymbol{\imath}}}\right\rangle \psi^{\star}_{\overline{\boldsymbol{\ell}},\overline{\boldsymbol{\imath}}}\right\|_{L^2(\Omega)} \left\|\psi^{\star}_{\boldsymbol{\ell},\boldsymbol{\imath}}\right\|_{L^2(\Omega)} = \left\| f - \sum_{\overline{\boldsymbol{\ell}}\in\mathcal{I}_n,\overline{\boldsymbol{\imath}}\in\tau_{\overline{\boldsymbol{\ell}}}} \left\langle f, \widetilde{\psi}^{\star}_{\overline{\boldsymbol{\ell}},\overline{\boldsymbol{\imath}}}\right\rangle \psi^{\star}_{\overline{\boldsymbol{\ell}},\overline{\boldsymbol{\imath}}}\right\|_{L^2(\Omega)}.\end{aligned}$$

L'erreur commise sur le calcul de $\tilde{b}$ est donc proportionnelle à l'erreur d'interpolation sur la *Sparse Grid*.

Dans le cas où les m indices $\imath_{k_1}, \ldots, \imath_{k_m}$ sont pairs, la relation de filtre (2.176) permet de se ramener au cas impair. Pour un niveau de raffinement maximum n, il existe une constante $C(n)$ telle que

$$\mathcal{E} = \sup_{\boldsymbol{\ell}\in\mathcal{I}_n,\boldsymbol{\imath}\in\tau_{\boldsymbol{\ell}}} \left|\tilde{b}_{\boldsymbol{\ell},\boldsymbol{\imath}}(f) - \tilde{b}^{\star}_{\boldsymbol{\ell},\boldsymbol{\imath}}(f)\right| \leq C \left\| f - \sum_{\overline{\boldsymbol{\ell}}\in\mathcal{I}_n,\overline{\boldsymbol{\imath}}\in\tau_{\overline{\boldsymbol{\ell}}}} \left\langle f, \widetilde{\psi}^{\star}_{\overline{\boldsymbol{\ell}},\overline{\boldsymbol{\imath}}}\right\rangle \psi^{\star}_{\overline{\boldsymbol{\ell}},\overline{\boldsymbol{\imath}}}\right\|_{L^2(\Omega)}. \qquad (2.179)$$

Si la fonction f n'est pas régulière, l'ensemble $\mathcal{J}$ est construit pour tenir compte des singularités de f. En pratique, l'ensemble $\mathcal{J}$ est construit récursivement en ajoutant uniquement les points d'interpolation qui héritent du point $(\boldsymbol{\ell}, \boldsymbol{\imath})$ si $\left|\tilde{b}^{\star}_{\boldsymbol{\ell},\boldsymbol{\imath}}(f)\right| \geq \epsilon$. Les résultats numériques valident cette approximation.

2.4.3.4 Préconditionnement diagonal

Si la forme bilinéaire a associée à l'opérateur $\mathcal{L}$ d'ordre 2 est symétrique, coercive et continue en norme $\mathcal{H}^t(\Omega)$, les propriétés du préconditionnement diagonal de la matrice $\boldsymbol{A}$ sur la base d'ondelettes sont une conséquence de la caractérisation de la norme $H^1(\Omega)$ en fonction des coefficients de la représentation d'une fonction u sur la base d'ondelettes.

L'encadrement suivant se déduit des propriétés de coercivité et de continuité de a en norme H^1

$$\alpha\|u\|^2_{H^1(\Omega)} \underbrace{\leq}_{(2.142)} a(u,u) \underbrace{\leq}_{(2.141)} \beta\|u\|^2_{H^1(\Omega)}. \qquad (2.180)$$

Dans le cas d'une base d'ondelettes obtenue par produit tensoriel isotrope (voir § 1.2.3.2), Cohen [Coh00] montre que le préconditonneur

$$D_{(\boldsymbol{\ell},\boldsymbol{\imath}),(\overline{\boldsymbol{\ell}},\overline{\boldsymbol{\imath}})} = 2^{|\boldsymbol{\ell}|_\infty}\delta^{\overline{\boldsymbol{\ell}}}_{\boldsymbol{\ell}}\delta^{\overline{\boldsymbol{\imath}}}_{\boldsymbol{\imath}},$$

est spectralement équivalent à la matrice du système. Ce résultat est basé sur l'équivalence de norme suivante :

$$c\sum_{\boldsymbol{\ell}} 2^{2|\boldsymbol{\ell}|_\infty}\|w_{\boldsymbol{\ell}}\|^2_{L^2(\Omega)} \leq \|u\|^2_{H^1(\Omega)} \leq C\sum_{\boldsymbol{\ell}} 2^{2|\boldsymbol{\ell}|_\infty}\|w_{\boldsymbol{\ell}}\|^2_{L^2(\Omega)}, \tag{2.181}$$

avec $w_{\boldsymbol{\ell}} = \sum_{\imath\in\tau_{\boldsymbol{\ell}}} \left\langle u, \widetilde{\psi}_{\boldsymbol{\ell},\boldsymbol{\imath}}\right\rangle \psi_{\boldsymbol{\ell},\boldsymbol{\imath}}$ et donc $\|w_{\boldsymbol{\ell}}\|^2_{L^2} \approx \sum_{\imath\in\tau_{\boldsymbol{\ell}}} u^2_{\boldsymbol{\ell},\boldsymbol{\imath}}$.

Si cette relation est bien vérifiée, alors la matrice B a ses valeurs propres comprises entre αc et βC. Le conditionnement de la matrice B est supérieur à $\gamma = \dfrac{\alpha c}{\beta C}$, et est indépendant du niveau n.

Dans le cas du produit tensoriel anisotrope (voir § 1.2.3.2)- donc des bases sparse - nous n'avons pas prouvé que les constantes C et c dans (2.181) sont indépendantes de d et de n, le niveau de discrétisation sur la base sparse, car les ondelettes des premiers niveaux n'ont pas leur premiers moments nuls. La dépendance de n du nombre de conditionnement est au pire algébrique.

On peut aussi remplacer D par la matrice diagonale $\bar{D}$ dont les coefficients sont donnés par

$$\bar{D}_{(\boldsymbol{\ell},\boldsymbol{\imath}),(\overline{\boldsymbol{\ell}},\overline{\boldsymbol{\imath}})} = \left(a\left(\psi_{\overline{\boldsymbol{\ell}},\overline{\boldsymbol{\imath}}}, \psi_{\boldsymbol{\ell},\boldsymbol{\imath}}\right)\right)^{\frac{1}{2}} \delta^{\overline{\boldsymbol{\ell}}}_{\boldsymbol{\ell}}\delta^{\overline{\boldsymbol{\imath}}}_{\boldsymbol{\imath}}. \tag{2.182}$$

En pratique, ce choix est meilleur dans la mesure où il tient compte des variations locales de l'opérateur $\mathcal{L}$.

Remarque 2.21 *Dans le cas d'opérateurs non symétriques et à coefficients variables, il n'est pas clair que le préconditionnement diagonal $\bar{D}$ suffise à obtenir un conditionnement indépendant des coefficients $\boldsymbol{a}_{i,j}$ et $\boldsymbol{b}_i$. D'autres préconditionneurs peuvent être envisagés. Ils sont basés sur le fait que les matrices de rigidité 1D en base d'ondelettes d'opérateurs intégraux ou différentiels ont des inverses qui admettent une structure presque diagonale. Cette propriété, comme nous le verrons dans le paragraphe suivant pour des opérateurs intégraux, est issue de la décroissance du noyau de l'opérateur. Le lecteur trouvera dans [Mas99] les détails sur ces méthodes de préconditionnement dans le cas 1D. Notons cependant que la matrice de rigidité sur la base sparse est beaucoup moins creuse que la matrice de rigidité dans le cas 1D.*

Tests numériques Les résultats présentés dans les paragraphes suivants sont obtenus pour la résolution de l'équation de Poisson

$$-\Delta u = f \quad \text{dans } \Omega, \quad u = 0 \text{ sur } \partial\Omega, \tag{2.183}$$

pour un terme source f égal à 1.

La symétrie de l'opérateur dans (2.183) nous permet de considérer un solveur itératif de type Gradient Conjugué. Comme nous venons de le décrire, la formulation discrète (2.183) consiste à résoudre un système linéaire. Ce système est mal conditionné. La méthode itérative converge en plus de 200 itérations dans le cas de la dimension 2 et du niveau $L = 7$. Comme décrit précédemment, le préconditionneur diagonal est efficace. Les tests numériques Tab.2.4.3.4 ont permis de montrer la détérioration du nombre d'itérations en fonction de la dimension et/ou du niveau de discrétisation.

TAB. 2.5 – Nombre d'itérations du Gradient Conjugué préconditionné par la diagonale pour une méthode de Galerkin sur une base d'ondelettes biorthogonale $(2,2)$ sparse, appliquée à l'opérateur $-\Delta$. L'erreur relative de $1 \cdot 10^{-4}$.

dimension/ Niveau	5	6	7	8	9	10
1	6	6	6	6	7	7
2	11	13	15	16	17	18
3	13	18	24	28	31	
4	17	25	28	32		
5	23	28	30	31		

TAB. 2.6 – Nombre de coefficients non-nuls de la matrice de rigidité en dimension 4

Niveau	Nb pts de Grille	Nb Coef.$\neq 0$ non nul	Nb Coef.$\neq 0$ / sqr(Nb pts)
5	770	554,209	93.5 %
6	2,562	5,740,441	87.4 %
7	7.938	50,266,601	79.7 %

2.4.3.5 Compression de la matrice

Les tests numériques nous permettent de montrer que la matrice de rigidité d'un opérateur différentiel sur une grille sparse est une matrice quasiment pleine. Le tableau 2.6 donne le nombre de coefficients non nuls de la matrice de rigidité associée à l'opérateur Laplacien dans le cas de la dimension 4. Pour un niveau de raffinement $n = 7$, 80% des coefficients sont a priori non nuls.

Cet exemple illustre l'intérêt de la compression des opérateurs en base d'ondelettes. En remarquant que les opérateurs différentiels admettent une représentation plus creuse en base nodale qu'en base d'ondelettes, la question de la compression sur la base d'ondelettes se pose naturellement.

La compression s'articule autour de deux éléments : la compression en espace (c.-à-d. entre les ondelettes d'un même niveau) et la compression en échelle. Dans le cas des opérateurs différentiels, la compression en espace est naturelle. Elle consiste à ne considérer que les paires d'ondelettes de supports non-disjoints.

Nous allons considérer deux approches pour rendre la matrice de rigidité plus creuse. La première méthode consiste à construire la matrice de rigidité et à supprimer les coefficients d'une ligne dont la valeur relative par rapport au coefficient diagonal correspondant est inférieure à une constante ϵ.

$$a_{i,j} \to 0 \quad \text{si} \quad \left| \frac{a_{i,j}}{a_{i,i}} \right| \leq \epsilon. \tag{2.184}$$

La seconde stratégie consiste à travailler d'abord sur les matrices de rigidité 1D. Une règle de compression en fonction des résultats énoncés par Masson [Mas99] appendice 3, sur des bases pleines, est généralisée au cas des bases sparse.

Le taux de compression est défini comme le rapport entre le nombre d'éléments non nuls de la matrice de rigidité de l'opérateur différentiel après compression et celui avant compression.

TAB. 2.7 – Compression par rapport aux coefficients diagonaux de la matrice de rigidité en dimension 1,$\epsilon = 0.01$ et erreur sur la solution de (2.183)

Niveau	Taux de compression	Erreur ℓ_∞
6	61 %	6e-16
7	55 %	4e-16
8	50 %	3e-16
9	45 %	1e-15

TAB. 2.8 – Compression par rapport aux coefficients diagonaux de la matrice de rigidité en dimension 2,$\epsilon = 0.01$ et erreur sur la solution de (2.183)

Niveau	Nb points	Taux de compression	Erreur ℓ_2	Erreur ℓ_∞
6	322	25 %	7.12e-3	1.0e-3
7	770	16.6 %	11.5e-3	1.0e-3
8	1794	10 %	18.0e-3	1.0e-3
9	4098	6.6 %	27.0e-3	1.0e-3

Compression par comparaison avec le coefficient diagonal Les tableaux 2.7 présentent les résultats de compression sur le problème (2.183) en dimension 1 pour différents taux de compression ϵ définis par (2.184). Les tableaux 2.8 et 2.9 (*resp.*, 2.10, 2.11) reprennent le même problème en dimension 2 (*resp.*4). L'erreur indiquée est celle entre la solution calculée sans compression et avec compression. L'erreur notée ℓ_2 correspond à la somme des carrés des erreurs sur chaque point de la *Sparse Grid*. L'erreur absolue ℓ_∞ correspond au maximum des valeurs absolues des erreurs évaluées sur chaque point de la *Sparse Grid*.

Les résultats obtenus sont satisfaisants quant au taux de compression. La taille de la matrice de rigidité est diminuée d'un facteur 10 en dimension 2 pour le niveau 8 (256 points de grilles) voir la figure 2.5. Nous comparons à présent ces résultats à ceux obtenus par une compression sur les matrices de rigidité des opérateurs en dimension 1.

Compression sur la base 1D Nous souhaitons également compresser les matrices d'opérateurs en dimension 1 avant d'effectuer le produit tensoriel sparse. Les résultats énoncés dans [Mas99] sur la compression des opérateurs différentiels sur une base d'ondelettes sont repris.

La compression consiste alors à ne pas calculer les coefficients lorsque la différence de niveau entre deux ondelettes est supérieure à une constante L fixée. Cette technique ne

TAB. 2.9 – Compression par rapport aux coefficients diagonaux de la matrice de rigidité en dimension 2,$\epsilon = 0.001$ et erreur sur la solution de (2.183)

Niveau	Nb points	Taux de compression	Erreur ℓ_2	Erreur ℓ_∞
6	322	44.6 %	1.1e-4	3.1e-5
7	770	30.0 %	1.7e-4	3.6e-5
8	1794	20.4 %	2.6e-4	4.1e-5
9	4098	13.5 %	3.9e-4	4.6e-5

TAB. 2.10 – Compression par rapport aux coefficients diagonaux de la matrice de rigidité en dimension 4,$\epsilon = 0.001$ et erreur sur la solution de (2.183)

Niveau	Nb points	Taux de compression	Erreur ℓ_2	Erreur absolue ℓ_∞
4	210	45 %	5e-3	0.7e-3
5	770	21 %	18e-3	2.3e-3
6	2562	9.68 %	52e-3	3.6e-3

TAB. 2.11 – Compression par rapport aux coefficients diagonaux de la matrice de rigidité en dimension 4,$\epsilon = 1e^{-4}$ et erreur sur la solution de (2.183)

Niveau	Nb points	Taux de compression	Erreur ℓ_2	Erreur ℓ_∞
5	770	50 %	4.8e-3	1.6e-3
6	2562	30 %	7.2e-3	1.2e-3

nous semble pas être adaptée au cas des *Sparse Grid* :

- D'une part, dans le cas de la dimension 1, des phénomènes d'oscillation liés à la compression de la matrice de rigidité sont observés dès que les coefficients sont mis à zéro pour une différence de niveau de 4 ou 5 (voir la figure 2.6). Si la constante est $L = 6$, le taux de compression n'est pas très important, pour des *Sparse Grid* de niveau 7 ou 8. Cette remarque est à moduler par la faible régularité des ondelettes choisies (B-spline d'ordre 2). En augmentant la régularité de ces ondelettes les résultats de compression peuvent être améliorés. Toutefois, d'autres inconvénients apparaîtraient.
- D'autre part, nous n'agissons pas sur la matrice de masse. En effet, les résultats de compression dépendent de l'ordre de l'opérateur et ne peuvent donc pas s'appliquer dans ce cas.

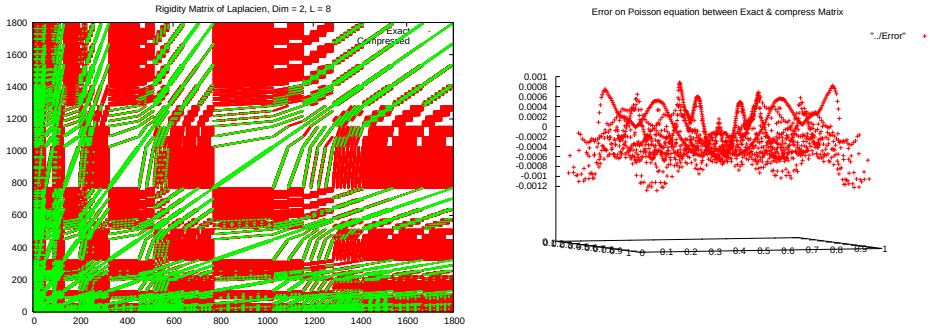

FIG. 2.5 – Coefficients non nuls de la matrice de rigidité d'un opérateur différentiel sur une base d'ondelettes en dimension 2 et coefficients obtenus après compression de cette matrice de rigidité (Gauche) - Erreur entre la solution sans compression de la matrice de rigidité et la solution avec compression de cette matrice Droite (Dim = 2 et L= 8).

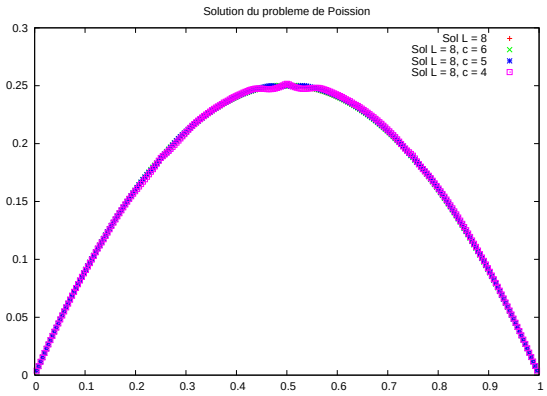

FIG. 2.6 – Solution après compression de la matrice de rigidité suivant la méthode des niveaux pour dim= 1, L= 8. Les résultats sont donnés sur différentes compressions *i.e.* les coefficients sont supprimés si les ondelettes correspondent à des niveaux différents de $4, 5, 6, \infty$.

2.5 Opérateur intégral

Nous étudions ici la discrétisation d'un opérateur intégral $\mathcal{L}$ de la forme

$$\mathcal{L}u\left(\boldsymbol{x}\right) \stackrel{\text{def}}{=} \int K\left(\boldsymbol{x},\boldsymbol{y}\right) u(\boldsymbol{x}+\boldsymbol{y}) d\boldsymbol{y}, \tag{2.185}$$

et les techniques de compression de l'opérateur discret associé.

Nous reviendrons par la suite sur le lien entre ces opérateurs et les équations de valorisation d'options sous une dynamique Markovienne. Les méthodes de Galerkin sur une base d'ondelettes et les algorithmes de compression sont étudiés pour les processus de Lévy en dimension 1 dans [MSW06]. Remarquons cependant que, dans le cas d'un processus de Lévy, le noyau k ne dépend pas de x. Nous verrons que cela ne modifie pas les propriétés de compression de l'opérateur discret associé à $\mathcal{L}$. Cependant, dans le cas général de (2.185), il n'est pas toujours possible de calculer l'opérateur discret associé.

Les méthodes classiques de discrétisation de l'opérateur $\mathcal{L}$ aboutissent à une matrice de rigidité pleine. Concrètement, à partir d'une discrétisation sur une grille uniforme avec $\left(2^{d\,n}\right)$ points (d est la dimension et n le niveau de discrétisation), la matrice de rigidité de l'opérateur $\mathcal{L}$ possède $O\left(2^{2d\,n}\right)$ coefficients non nuls.

Une stratégie standard pour réduire le coût de calcul consiste à exploiter la décroissance du noyau K. Nous caractérisons cette décroissance par la propriété suivante :

Propriété 2.24 (Décroissance du noyau) *L'opérateur intégral de Caldéron-Zygmund $\mathcal{L}$ vérifie la propriété de décroissance du noyau s'il existe une constante $C<\infty$ dépendant de $\boldsymbol{\alpha},\boldsymbol{\beta}$ et un exposant $f\left(\boldsymbol{\alpha},\boldsymbol{\beta}\right)$ tels que*

$$\forall \boldsymbol{x},\boldsymbol{y} \quad \left|\partial_x^{\boldsymbol{\alpha}}\partial_y^{\boldsymbol{\beta}} K\left(\boldsymbol{x},\boldsymbol{x}-\boldsymbol{y}\right)\right| \leq C\left|\boldsymbol{x}-\boldsymbol{y}\right|^{-f(\boldsymbol{\alpha},\boldsymbol{\beta})}, \quad \textit{avec} \quad f\left(\boldsymbol{\alpha},\boldsymbol{\beta}\right)>0. \tag{2.186}$$

Le lecteur trouvera [Mey90b] une description des propriétés des opérateurs de Caldéron-Zygmund. Les techniques de compression de la matrice de rigidité pour des opérateurs intégraux ont fait l'objet de nombreux travaux [DPS93, PS96, Sch98b]. Nous étudions la généralisation de ces résultats sur les bases sparse, tout d'abord, pour les méthodes de Galerkin, puis dans le cas de schémas aux différences finies. Dans ce qui suit, nous supposons que le noyau vérifie la propriété 2.24.

2.5.1 Méthode de Galerkin sur une base d'ondelettes

La méthode de Galerkin appliquée à une famille d'ondelettes biorthogonales (munies d'un nombre suffisant de moments nuls) permet d'obtenir une matrice de rigidité pour laquelle la plupart des coefficients sont proches de zéro. Dans le paragraphe suivant, nous rappelons les résultats d'approximation sur un espace discret caractérisé par l'ensemble $\mathcal{I}$ des paires de multi-indices $\left(\boldsymbol{\ell},\bar{\boldsymbol{\ell}}\right)$ correspondant aux niveaux des ondelettes. Dans un second paragraphe, nous donnons les résultats obtenus après compression de l'opérateur. La compression de l'opérateur $\mathcal{L}$ consiste à remplacer l'opérateur discret $\mathcal{L}_{\mathcal{I}}$, dont la matrice de rigidité est donnée par $\left(\left\langle \mathcal{L}\psi_{\bar{\boldsymbol{\ell}},\bar{\boldsymbol{\imath}}}, \psi_{\boldsymbol{\ell},\boldsymbol{\imath}}\right\rangle\right)_{(\boldsymbol{\ell},\boldsymbol{\imath}),(\bar{\boldsymbol{\ell}},\bar{\boldsymbol{\imath}})}$, par l'opérateur $\tilde{\mathcal{L}}_{\mathcal{I}}$ obtenu en annulant les coefficients proches de zéro de la matrice de rigidité de $\mathcal{L}_{\mathcal{I}}$. Nous montrerons que cette compression permet de conserver l'ordre d'approximation de la méthode numérique.

2.5.1.1 Opérateur intégral sur une base d'ondelettes

Nous rappelons les résultats d'approximation de la représentation sous la forme d'une AMR :
Soit $u = \sum_{\boldsymbol{\ell}\in\mathbb{N}^d} w_{\boldsymbol{\ell}},\ w_{\boldsymbol{\ell}} \in W_{\boldsymbol{\ell}}$ (*resp.*$u = \sum_{\boldsymbol{\ell}\in\mathbb{N}^d} \tilde{w}_{\boldsymbol{\ell}},\ \tilde{w}_{\boldsymbol{\ell}} \in \tilde{W}_{\boldsymbol{\ell}}$) la décomposition de u sur l'AMR primale (*resp.*duale), alors

$$\|u\|^2_{\mathcal{H}^{t,s}(\Omega)} \approx \sum_{\boldsymbol{\ell}\in\mathbb{N}^d} 2^{2t|\boldsymbol{\ell}|_1+2s|\boldsymbol{\ell}|_\infty} \|w_{\boldsymbol{\ell}}\|^2_{L^2(\Omega)} \text{ pour } t \geq 0,\ 0 \leq t+s \leq p, \tag{2.187}$$

$$\|u\|^2_{\mathcal{H}^{t,s}(\Omega)} \approx \sum_{\boldsymbol{\ell}\in\mathbb{N}^d} 2^{2t|\boldsymbol{\ell}|_1+2s|\boldsymbol{\ell}|_\infty} \|\tilde{w}_{\boldsymbol{\ell}}\|^2_{L^2(\Omega)} \text{ pour } t \geq 0,\ 0 \leq t+s \leq \tilde{p}, \tag{2.188}$$

où p (*resp.*$\tilde{p}$) représente le nombre de moments nuls de l'ondelette primale (*resp.*duale).

Dans ce qui suit, nous notons pour simplifier

$$\|K\|_{\mathcal{H}^{t,k}(\Omega)} = \|K\|_{\mathcal{H}^{t,k}(\Omega\times\Omega)}.$$

Théorème 2.25 (Approximation sparse de $\mathcal{L}$,[KS02] théorème 1) *Soit $\mathcal{L}$ l'opérateur intégral défini par (2.185) et $\mathcal{I} \in \left(\mathbb{N}^d\right)^2$. Soient s,q,r,t,k tels que : $-\tilde{p} < s < p$, $q \in [0,p)$, $q+r \in [0,p)$ et $t \in [0,\tilde{p})$, $t+k \in [0,\tilde{p})$, tels que $K \in \mathcal{H}^{t,k}(\Omega\times\Omega)$*

$$\text{et } (\mathcal{L}-\mathcal{L}_{\mathcal{I}}) \text{ pour } u \in H^s(\Omega), \quad u \in \mathcal{H}^{q,r}(\Omega),$$

(La définition de $\mathcal{H}^{q,r}(\Omega)$ est donnée par Def.1.4) alors

$$\|(\mathcal{L}-\mathcal{L}_{\mathcal{I}})u\|_{H^s(\Omega)} \leq C \max_{(\boldsymbol{\ell},\overline{\boldsymbol{\ell}})\notin\mathcal{I}} \left(2^{s|\boldsymbol{\ell}|_\infty - k|\boldsymbol{\ell},\overline{\boldsymbol{\ell}}|_\infty - r|\overline{\boldsymbol{\ell}}|_\infty - t|\boldsymbol{\ell}|_1 - (t+q)|\overline{\boldsymbol{\ell}}|_1}\right) \|K\|_{\mathcal{H}^{t,k}(\Omega)} \|u\|_{\mathcal{H}^{q,r}(\Omega)}. \tag{2.189}$$

Remarque 2.22 *La grille pleine classique correspond à l'ensemble $\mathcal{I} = \left\{(\boldsymbol{\ell},\overline{\boldsymbol{\ell}}) : |\boldsymbol{\ell},\overline{\boldsymbol{\ell}}|_\infty \leq n\right\}$, la sparse grille classique V_n^0 correspond à l'ensemble $\mathcal{I} = \left\{(\boldsymbol{\ell},\overline{\boldsymbol{\ell}}) : \max\left(|\boldsymbol{\ell}|_1, |\overline{\boldsymbol{\ell}}|_1\right) \leq n+d-1\right\}$.*

Remarque 2.23 *La condition $\tilde{p} > 0$ est nécessaire pour prendre en compte l'effet régularisant de K. Si $\tilde{p} = 0$ (cas des ondelettes interpolantes), (2.189) devient*

$$\|(\mathcal{L}-\mathcal{L}_{\mathcal{I}})u\|_{H^s} \leq C \max_{\boldsymbol{\ell},\overline{\boldsymbol{\ell}}\notin\mathcal{I}} \left(2^{s|\boldsymbol{\ell}|_\infty - r|\overline{\boldsymbol{\ell}}|_\infty - q|\overline{\boldsymbol{\ell}}|_1}\right) \|K\|_{L^2(\Omega)} \|u\|_{\mathcal{H}^{q,r}(\Omega)}. \tag{2.190}$$

2.5.1.2 Compression de la matrice de rigidité

Dans ce qui suit, nous nous restreignions à la méthode de Galerkin, *i.e.* nous considérons le même espace pour la décomposition et la projection. Cette méthode est également dénommée méthode d'ondelettes isotropes dans la littérature, par opposition aux méthodes de Petrov Galerkin appelées aussi méthodes d'ondelettes anisotropes. La méthode de compression proposée dans [DPS93, Sch98b] permet de ramener le nombre de coefficients non nuls de la matrice de rigidité à $O\left(n^k 2^{dn}\right)$ avec $k \in \mathbb{R}^+$ au lieu de à $O\left(2^{2dn}\right)$.

En reprenant le cheminement de [DPS93], Knapeck [Kna00] adapte ce résultat au cas des *Sparse Grid*.

Théorème 2.26 ([Kna00] théorème 9 p80) *Nous supposons vérifiées les hypothèses suivantes :*

- *la fonction $\psi_{\boldsymbol{\ell},\boldsymbol{\imath}}$ est obtenue par produit tensoriel d'ondelettes unidimensionnelles à p moments nuls,*
- *le noyau K de l'opérateur $\mathcal{L}$ défini par (2.185) vérifie la propriété 2.24 avec $|\boldsymbol{\alpha},\boldsymbol{\beta}|_\infty \leq p+1$.*

Alors les coefficients de la matrice de rigidité vérifient :

$$\left\langle \mathcal{L}\psi_{\boldsymbol{\ell},\boldsymbol{\imath}}, \psi_{\overline{\boldsymbol{\ell}},\overline{\boldsymbol{\imath}}} \right\rangle \leq C 2^{-|\boldsymbol{\ell},\overline{\boldsymbol{\ell}}|_1}(p+3/2) dist\left(supp\left(\psi_{\boldsymbol{\ell},\boldsymbol{\imath}}\right), supp\left(\psi_{\overline{\boldsymbol{\ell}},\overline{\boldsymbol{\imath}}}\right)\right)^{-f(p+1,p+1)}, \tag{2.191}$$

avec

$$dist\left(supp\left(\psi_{\boldsymbol{\ell},\boldsymbol{\imath}}\right), supp\left(\psi_{\overline{\boldsymbol{\ell}},\overline{\boldsymbol{\imath}}}\right)\right) = \max_{\boldsymbol{x}\in supp(\psi_{\boldsymbol{\ell},\boldsymbol{\imath}}), \boldsymbol{y}\in supp(\psi_{\overline{\boldsymbol{\ell}},\overline{\boldsymbol{\imath}}})} |\boldsymbol{x}-\boldsymbol{y}|. \tag{2.192}$$

La fonction f caractérise la décroissance du noyau d'après la propriété 2.24.

Preuve La démonstration utilise le développement de Taylor du noyau K jusqu'à l'ordre $p+1$, voir [Kna00]. ■

Remarque 2.24 *Une seconde compression est possible sur les niveaux. Cette compression est basée sur la distance entre la singularité de la dérivée de chacune des deux ondelettes. Le lecteur trouvera dans [Rei08] une application de cette compression au cas d'une équation intégro-différentielle obtenue pour l'évaluation dans un modèle à volatilité stochastique avec des processus à saut.*

Soit $B_{\boldsymbol{\ell},\overline{\boldsymbol{\ell}}}(\theta)$ la distance de compression sur les niveaux des ondelettes, pour $(\boldsymbol{\ell},\overline{\boldsymbol{\ell}}) \in \mathcal{I}$,

$$B_{\boldsymbol{\ell},\overline{\boldsymbol{\ell}}}(\theta) = \min\left(1, n^{\frac{d}{f(p+1,p+1)}} 2^{\theta n + s|\boldsymbol{\ell}|_\infty - r|\overline{\boldsymbol{\ell}}|_\infty - q|\overline{\boldsymbol{\ell}}|_1 - (p+1)|\boldsymbol{\ell},\overline{\boldsymbol{\ell}}|_1 (f(p+1,p+1))^{-1}}\right), \tag{2.193}$$

où $n = \max\left(\left|\boldsymbol{\ell},\overline{\boldsymbol{\ell}}\right|_\infty | \left(\boldsymbol{\ell},\overline{\boldsymbol{\ell}}\right) \in \mathcal{I}\right)$. Cette distance $B_{\boldsymbol{\ell},\overline{\boldsymbol{\ell}}}(\theta)$ permet de définir l'ensemble des paires de multi-indices $\gamma_{\boldsymbol{\ell},\overline{\boldsymbol{\ell}}}(\theta)$ tel que

$$\gamma_{(\boldsymbol{\ell},\overline{\boldsymbol{\ell}})}(\theta) = \left\{(\boldsymbol{\imath},\overline{\boldsymbol{\imath}}) \in \tau_{(\boldsymbol{\ell},\overline{\boldsymbol{\ell}})} \middle| dist\left(supp(\psi_{\boldsymbol{\ell},\boldsymbol{\imath}}), supp(\psi_{\overline{\boldsymbol{\ell}},\overline{\boldsymbol{\imath}}})\right) \leq B\left(\boldsymbol{\ell},\overline{\boldsymbol{\ell}}\right)\right\}. \tag{2.194}$$

Théorème 2.27 (Approximation sparse compressée de $\mathcal{L}$, [KS02] théorème 2) *Soit $\widetilde{\mathcal{L}}_{\mathcal{I}}$ l'opérateur différentiel associé à la matrice de rigidité dont les coefficients sont non nuls si et seulement si la paire d'indice appartient à l'ensemble des paires de multi-niveaux $\gamma_{(\boldsymbol{\ell},\overline{\boldsymbol{\ell}})}(\theta)$.*

Si s,q,r vérifient $-\widetilde{p} < s < p$, $0 \leq q < r$, $0 \leq q+r < p$ avec $\left(\mathcal{L}-\widetilde{\mathcal{L}}_{\mathcal{I}}\right)u \in H^s(\Omega)$ et $u \in \mathcal{H}^{q,r}(\Omega)$ alors

$$\left\|\left(\mathcal{L}-\widetilde{\mathcal{L}}_{\mathcal{I}}\right)u\right\|_{H^s(\Omega)} \leq C 2^{-\theta n}\|u\|_{\mathcal{H}^{q,r}(\Omega)}. \tag{2.195}$$

Preuve La démonstration repose sur le lemme

$$\left\|\left(\mathcal{L}-\widetilde{\mathcal{L}}_{\mathcal{I}}\right)u\right\|_{H^s} \leq C \sum_{(\boldsymbol{\ell},\overline{\boldsymbol{\ell}})\in\mathcal{I}} \sum_{\boldsymbol{\imath}\in\tau_{\boldsymbol{\ell}},\ ,\overline{\boldsymbol{\imath}}\in\tau_{,\overline{\boldsymbol{\ell}}}} \left(2^{s|\boldsymbol{\ell}|_\infty - r|\overline{\boldsymbol{\ell}}|_\infty - 2q|\overline{\boldsymbol{\ell}}|_1}\right) \boldsymbol{a}_{(\boldsymbol{\ell},\boldsymbol{\imath}),(\overline{\boldsymbol{\ell}},\overline{\boldsymbol{\imath}})} \|u\|_{\mathcal{H}^{q,r}}, \tag{2.196}$$

où

$$\boldsymbol{a}_{(\boldsymbol{\ell},\boldsymbol{\imath}),(\bar{\boldsymbol{\ell}},\bar{\boldsymbol{\imath}})} = \left\langle \langle K, \psi_{\boldsymbol{\ell},\boldsymbol{\imath}} \rangle_{L^2\left(\mathbb{R}^d\right)}, \psi_{\bar{\boldsymbol{\ell}},\bar{\boldsymbol{\imath}}} \right\rangle_{L^2\left(\mathbb{R}^d\right)}.$$

Ce résultat est obtenu en reprenant les éléments de démonstration du théorème 2.25 pour obtenir (2.196), puis le théorème 2.26 pour la majoration de $\boldsymbol{a}_{(\boldsymbol{\ell},\boldsymbol{\imath}),(\bar{\boldsymbol{\ell}},\bar{\boldsymbol{\imath}})}$. ■

2.5.1.3 Généralisation : méthode de Petrov-Galerkin

Dans le cas d'une méthode anisotrope ou méthode de Petrov Galerkin, le résultat n'est pas démontré mais semble être correct. Cependant, nous verrons, dans un paragraphe suivant, que le calcul des coefficients de la matrice de rigidité d'un opérateur différentiel pour ce type de méthode n'est pas facile. Il en est de même pour les opérateurs intégraux.

2.5.1.4 Mise en oeuvre

Nous avons évoqué dans § 2.4.3.2 les difficultés inhérentes au calcul des coefficients de la matrice de rigidité d'une méthode de Galerkin sur une base d'ondelettes. Dans le cas d'opérateurs intégraux, ces calculs sont abordables lorsque le noyau K de (2.185) vérifie l'une des propriétés suivantes :

- K est à variables séparées, *i.e.*$K(\boldsymbol{x},\boldsymbol{y}) = K_1(x_1,y_1)\dots K_d(x_d,y_d)$.
- K ne dépend pas de $\boldsymbol{x}$.

Dans le premier cas, nous somme ramenés au calcul sur des bases $1D$. Le second cas permet de calculer la matrice de rigidité à l'aide de l'identité de Parseval et de la transformée de Fourier des ondelettes, puisque, dans ce cas, l'opérateur $\mathcal{L}$ est un opérateur de convolution avec K.

Dans l'une de nos applications, l'opérateur intégral $\mathcal{L}$ ne possède aucune de ces deux propriétés. Le calcul analytique ou l'approximation numérique des coefficients de la matrice de rigidité étant trop coûteux en temps CPU et en mémoire, nous avons abandonné la méthode de Galerkin au profit d'une méthode de collocation pour la discrétisation de l'opérateur intégral.

2.5.2 Méthode de collocation

Nous décrivons, dans cette partie, un schéma de discrétisation d'un opérateur intégral pour lequel il n'est pas possible de mettre en oeuvre la méthode de Galerkin. A partir de l'analyse de la méthode de différences finies, voir § 2.3.3, nous obtenons un résultat de consistance sur l'opérateur discret.

2.5.2.1 Définition de différents opérateurs intégraux en dimension 1

Les modèles à saut définissent classiquement l'intensité de saut par rapport à la valeur du sous-jacent. L'équation d'évaluation d'une option européenne est alors une *équation intégro-différentielle* dans laquelle intervient l'opérateur

$$\mathcal{L}u(x) = \int_{\mathbb{R}} \left(u(x+z) - u(x)\right) d\kappa(z). \tag{2.197}$$

Une seconde approche consiste à définir l'intensité du saut en fonction de la volatilité du sous-jacent. Dans ce cas, l'opérateur dans (2.197) devient

$$\int_{\mathbb{R}} \left(u\left(x+\eta\left(x,z\right)\right)-u(x)\right) d\kappa(z) = \int_{\mathbb{R}} \left(u\left(x+y\right)-u(x)\right) d\bar{\kappa}(x,y). \tag{2.198}$$

Nous renvoyons au chapitre 7 pour la justification de ces modèles. Notons simplement que (2.198) ne correspond plus à un produit de convolution.

2.5.2.2 Discrétisation d'un opérateur intégral sur une grille sparse

Cas d'un opérateur de la forme (2.197) Soit u une fonction de $(\mathbb{R}^+)^d$ à valeurs dans $\mathbb{R}$ et κ un noyau d'intégration à une variable, par exemple $\kappa(z) = \exp\left(-\frac{1}{2}\frac{(z-\mu)^2}{\nu^2}\right)$. L'opérateur de l'équation (2.197) devient dans ce cas :

$$\int_{\mathbb{R}} \left(u\left(x_1+z,x_2,\ldots,x_d\right)-u\left(x_1,x_2\ldots,x_d\right)\right) d\kappa\left(z\right). \tag{2.199}$$

Définition 2.7 *Notons Q_{SG} la discrétisation de l'opérateur intégral (2.199) sur une grille sparse, selon le procédé suivant :*

1. *passage hiérarchique → nodal dans la direction x_1,*
2. *approximation par une formule de quadrature (ou une formule composée) de l'intégrale (2.197),*
3. *passage nodal → hiérarchique dans la direction x_1.*

Plus précisément, en notant $Q_M(v)$ ($v:\mathbb{R}\to\mathbb{R}$) une formule de quadrature (ou une formule composée) en dimension 1, $Q_M(v)(x) = \sum_{k=-M}^{M} \omega_k v\left(x+\zeta_k\right)$, où ω_k et ζ_k sont choisis en fonction du noyau intégral κ. En reprenant les notations de la section 2.3, le schéma de discrétisation de l'opérateur (2.199) s'écrit formellement :

$$\begin{aligned} \mathcal{L}u\left(x_{\boldsymbol{\ell},\boldsymbol{i}}\right) &= \int_{\mathbb{R}} \left(u\left(x_{\ell_1,i_1}+z,x_{\ell_2,i_2},\ldots,x_{\ell_d,i_d}\right)-u\left(x_{\boldsymbol{\ell},\boldsymbol{i}}\right)\right) d\kappa\left(z\right) \\ &\approx \left(Q_{SG}\left(u\right)\right)_{\boldsymbol{\ell},\boldsymbol{i}} \stackrel{\text{def}}{=} \left(T_1\circ Q_M\circ T_1^{-1}\hat{u}\right)_{\boldsymbol{\ell},\boldsymbol{i}}. \end{aligned} \tag{2.200}$$

Proposition 2.28 *Supposons que la formule de quadrature adaptée au noyau d'intégration κ vérifie, pour tout $v\in\mathcal{C}^2([0,1])$,*

$$\left|\int_{\mathbb{R}} v(x+z)d\kappa(z) - \sum_{k=-M}^{M} \omega_k v\left(x+\zeta_k\right)\right| \leq C2^{-2n}\left|v\right|_{\mathcal{C}^2([0,1])}, \tag{2.201}$$

alors, pour toute fonction $u\in\boldsymbol{\mathcal{C}^2}\left([0,1]^d\right)$,

$$\left\|\mathcal{P}\left(\mathcal{L}u\right)-\left(Q_{SG}\left(u\right)\right)\right\|_\infty \leq Cn^{d-1}2^{-2n}\left|u\right|_{\boldsymbol{\mathcal{C}^2}\left([0,1]^d\right)}. \tag{2.202}$$

Preuve En reprenant les éléments de démonstration du théorème 2.8, nous remarquons que la formule de quadrature Q_{SG} définit un opérateur continu. L'erreur de consistance $E_D(u,\ell) = P^\ell\left(\mathcal{L}\left(u\right)\right) - Q^\ell\circ P^\ell\left(v\right)$ vérifie la proposition 2.7. ■

Opérateur intégral de la forme (2.198) Le paragraphe suivant décrit la discrétisation d'un opérateur de saut dans le cas d'un modèle à volatilité stochastique décrit au chapitre 5. L'intensité de saut η de l'équation (2.198) est définie comme le produit de la volatilité $\sigma(x_2)$ par l'incrément de saut z. L'opérateur de l'équation (2.198) devient dans ce cas :

$$(\mathcal{L}u)(x_1,x_2) = \int_{\mathbb{R}} (u(x_1+\sigma(x_2)z, x_2) - u(x_1,x_2))\, d\kappa(z). \tag{2.203}$$

Les résultats qui suivent se généralisent sans difficulté au cas de la dimension d, la restriction à $d=2$ est utilisée pour simplifier l'écriture.

Définition 2.8 *Soit Q_{SG} l'opérateur matriciel associé à l'opérateur intégral donné par (2.203). Les coefficients de cette matrice sont donnés par la relation*

$$(Q_{SG})_{(\boldsymbol{\ell},\boldsymbol{\imath}),(\bar{\boldsymbol{\ell}},\bar{\boldsymbol{\imath}})} = \sum_{\tilde{\boldsymbol{n}}} g(\tilde{\boldsymbol{n}}) \left(\mathcal{L}\psi_{\bar{\boldsymbol{\ell}},\bar{\boldsymbol{\imath}}}\right)(x_{\boldsymbol{\ell}+\mathbf{1},\mathbf{2}\boldsymbol{\imath}+\tilde{\boldsymbol{n}}}). \tag{2.204}$$

Ceci donne pour l'opérateur (2.203)

$$(Q_{SG})_{(\boldsymbol{\ell},\boldsymbol{\imath}),(\bar{\boldsymbol{\ell}},\bar{\boldsymbol{\imath}})} = \sum_{\tilde{\boldsymbol{n}}} g(\tilde{\boldsymbol{n}})\, Q_{(\bar{\boldsymbol{\ell}},\bar{\boldsymbol{\imath}})}(x_{\ell_1+1,2i_1+\tilde{n}_1}, x_{\ell_1+2,2i_2+\tilde{n}_2}), \tag{2.205}$$

où la fonction $Q_{(\bar{\boldsymbol{\ell}},\bar{\boldsymbol{\imath}})} : \mathbb{R}^2 \to \mathbb{R}$ est donnée par

$$Q_{(\bar{\boldsymbol{\ell}},\bar{\boldsymbol{\imath}})}(x_1,x_2) = \int_{\mathbb{R}} \left(\psi_{\bar{\ell}_1,\bar{i}_1}(x_1+\sigma(x_2)z) - \psi_{\bar{\ell}_1,\bar{i}_1}(x_1)\right)\psi_{\bar{\ell}_2,\bar{i}_2}(x_2)\, d\kappa(z). \tag{2.206}$$

Le calcul analytique de $\displaystyle\int_{\mathbb{R}} \psi_{\bar{\ell}_1,\bar{i}_1}((x_{\ell_1+1,2i_1+\tilde{n}_1} + \sigma(x_{\ell_2+1,2i_2+n_2}))z)\, d\kappa(z)$ n'est pas toujours possible. Dans ce cas, il faut ajouter une erreur d'intégration numérique dans l'analyse de la consistance de l'opérateur.

Propriété 2.29 *Soit $u \in \mathcal{C}^{\mathbf{2}}\left([0,1]^d\right)$ alors*

$$\|\mathcal{P}(\mathcal{L}u) - Q_{SG}(\mathcal{P}(u))\|_\infty \le Cn^{d-1}2^{-2n}\|u\|_{\mathcal{C}^{(2,2\ldots,2)}}. \tag{2.207}$$

Preuve Reprenons la démonstration du théorème 2.16. L'erreur de consistance du schéma est donnée par

$$E = \max_{\boldsymbol{\ell}\in\mathcal{I}_n,\boldsymbol{\imath}\in\tau_{\boldsymbol{\ell}}} \left[\left|\left\langle \mathcal{L}u, \widetilde{\psi}_{\boldsymbol{\ell},\boldsymbol{\imath}}\right\rangle - \sum_{\bar{\boldsymbol{\ell}}\in\mathcal{I}_n,\bar{\boldsymbol{\imath}}\in\tau_{\bar{\boldsymbol{\ell}}}} Q_{SG}((\bar{\boldsymbol{\ell}},\bar{\boldsymbol{\imath}}),(\boldsymbol{\ell},\boldsymbol{\imath}))\left\langle u, \widetilde{\psi}_{\bar{\boldsymbol{\ell}},\bar{\boldsymbol{\imath}}}\right\rangle\right|\right]. \tag{2.208}$$

Or

$$\begin{aligned}
\sum_{\bar{\boldsymbol{\ell}}\in\mathcal{I}_n,\bar{\boldsymbol{\imath}}\in\tau_{\bar{\boldsymbol{\ell}}}} Q_{SG}((\bar{\boldsymbol{\ell}},\bar{\boldsymbol{\imath}}),(\boldsymbol{\ell},\boldsymbol{\imath}))\left\langle u, \widetilde{\psi}_{\bar{\boldsymbol{\ell}},\bar{\boldsymbol{\imath}}}\right\rangle &= \sum_{\bar{\boldsymbol{\ell}}\in\mathcal{I}_n,\bar{\boldsymbol{\imath}}\in\tau_{\bar{\boldsymbol{\ell}}}}\sum_{\tilde{\boldsymbol{n}}} \left\langle \mathcal{L}\psi_{\bar{\boldsymbol{\ell}},\bar{\boldsymbol{\imath}}}, \widetilde{\varphi}_{\boldsymbol{\ell}+\mathbf{1},\mathbf{2}\boldsymbol{\imath}+\tilde{\boldsymbol{n}}}\right\rangle \hat{u}_{\bar{\boldsymbol{\ell}},\bar{\boldsymbol{\imath}}} \\
&= \sum_{\bar{\boldsymbol{\ell}}\in\mathcal{I}_n,\bar{\boldsymbol{\imath}}\in\tau_{\bar{\boldsymbol{\ell}}}} \left\langle \mathcal{L}\psi_{\bar{\boldsymbol{\ell}},\bar{\boldsymbol{\imath}}}, \widetilde{\psi}_{\boldsymbol{\ell},\boldsymbol{\imath}}\right\rangle \hat{u}_{\bar{\boldsymbol{\ell}},\bar{\boldsymbol{\imath}}} \\
&= \left\langle \sum_{\bar{\boldsymbol{\ell}}\in\mathcal{I}_n,\bar{\boldsymbol{\imath}}\in\tau_{\bar{\boldsymbol{\ell}}}} \hat{u}_{\bar{\boldsymbol{\ell}},\bar{\boldsymbol{\imath}}}\mathcal{L}\psi_{\bar{\boldsymbol{\ell}},\bar{\boldsymbol{\imath}}}, \widetilde{\psi}_{\boldsymbol{\ell},\boldsymbol{\imath}}\right\rangle.
\end{aligned} \tag{2.209}$$

Nous en déduisons

$$E = \max_{\boldsymbol{\ell}\in\mathcal{I}_n,\boldsymbol{\imath}\in\tau_{\boldsymbol{\ell}}} \left[\left\langle \mathcal{L}\left(u - \sum_{\overline{\boldsymbol{\ell}}\in\mathcal{I}_n,\overline{\boldsymbol{\imath}}\in\tau_{\overline{\boldsymbol{\ell}}}} \hat{u}_{\overline{\boldsymbol{\ell}},\overline{\boldsymbol{\imath}}}\psi_{\overline{\boldsymbol{\ell}},\overline{\boldsymbol{\imath}}} \right), \widetilde{\psi}_{\boldsymbol{\ell},\boldsymbol{\imath}} \right\rangle \right] \tag{2.210}$$

Les fonctions $\widetilde{\psi}_{\boldsymbol{\ell},\boldsymbol{\imath}}$ sont une somme de masse de Dirac, donc

$$E \leq C \sup_{\boldsymbol{x}\in\Omega} \left(\mathcal{L}\left(u - \sum_{\overline{\boldsymbol{\ell}}\in\mathcal{I}_n,\overline{\boldsymbol{\imath}}\in\tau_{\overline{\boldsymbol{\ell}}}} \hat{u}_{\overline{\boldsymbol{\ell}},\overline{\boldsymbol{\imath}}}\psi_{\overline{\boldsymbol{\ell}},\overline{\boldsymbol{\imath}}} \right)(\boldsymbol{x}) \right). \tag{2.211}$$

Pour conclure, il suffit de remarquer que l'opérateur $\mathcal{L}$ est un opérateur continu et d'utiliser le théorème de projection sur une grille sparse 1.26 :

$$E \leq C \left\| u - \sum_{\overline{\boldsymbol{\ell}}\in\mathcal{I}_n,\overline{\boldsymbol{\imath}}\in\tau_{\overline{\boldsymbol{\ell}}}} \hat{u}_{\overline{\boldsymbol{\ell}},\overline{\boldsymbol{\imath}}}\psi_{\overline{\boldsymbol{\ell}},\overline{\boldsymbol{\imath}}} \right\|_{\mathcal{C}^{\mathbf{0}}} \leq C 2^{-2n} n^{d-1} \|u\|_{\mathcal{C}^{\mathbf{2}}}. \tag{2.212}$$

■

Le calcul de $Q_{SG}(\mathcal{P}(u))$ reste coûteux, voir le tableau 2.12. L'application de cette méthode est, par conséquent, réduite. En pratique, il n'est pas raisonnable de traiter ce terme intégral de manière implicite dans un problème parabolique intégro-différentiel. Sur ces problèmes, les résultats de convergence sont obtenus pour des modèles de diffusion dans lesquels le terme de saut est un terme « correctif ».

Tab. 2.12 – Temps (en seconde) de calcul pour un opérateur intégral en dimension 4

Niveau	6	7	8	9
Temps de calcul	0.182	1.128	6.5	38.6

2.6 Discrétisation d'un problème parabolique sur une base d'ondelettes

Dans le cas où la condition initiale n'est pas à variables séparées, il est nécessaire d'adapter le schéma de discrétisation en temps du problème parabolique. La première partie illustre la notion de schémas « dissipatifs ». Dans la seconde partie, nous étudions les schémas de discrétisation en temps d'ordres arbitrairement élevés introduits par Schwab & al [WGSS01]. Ces schémas sont basés sur une méthode de Galerkin discontinue.

2.6.1 Schéma dissipatif

L'équation de la chaleur nous permet d'illustrer la propriété « dissipative » de certains schémas,

$$\frac{\partial u}{\partial t} - \Delta u = 0, \quad u(0,x) = u_0(x). \tag{2.213}$$

Appliquons un θ-schéma en temps et analysons le comportement de ce schéma sur les variables de Fourier. Soient $u^n(x) = u(t_n, x)$ et $\widehat{u}^n$ la transformée de Fourier de u^n. Suivant l'espace de Fourier, la semi-discrétisation en temps de (2.213) devient :

$$\frac{\widehat{u}^{n+1} - \widehat{u}^n}{\delta_t} + \xi^2 \left(\theta \widehat{u}^{n+1} + (1-\theta)\widehat{u}^n\right) = 0, \quad \widehat{u}^0(\xi) = \widehat{u}_0(\xi), \tag{2.214}$$

ce qui peut s'écrire sous la forme :

$$\widehat{u}^{n+1}(\xi) = \frac{1 - \delta_t(1-\theta)\xi^2}{1 + \delta_t\ \theta\ \xi^2}\widehat{u}^n(\xi), \quad \widehat{u}^0(\xi) = \widehat{u}_0(\xi). \tag{2.215}$$

Intéressons-nous à présent aux hautes fréquences :

$$\widehat{u}^{n+1}(\xi) \approx \frac{(\theta - 1)}{\theta}\widehat{u}^n(\xi). \tag{2.216}$$

En particulier si $\theta = 1$, le schéma amortit très rapidement les hautes fréquences.

En d'autres termes, le schéma reproduit les propriétés de régularisation de l'équation de la chaleur. Ces propriétés sont essentielles pour pouvoir appliquer une méthode de Sparse Grid. En effet, il est nécessaire que, sur le problème discrétisé en temps, la fonction u^n appartienne à $\mathcal{H}^2(\Omega)$, pour tout $n > 1$. Nous présentons une classe de schémas qui vérifient cette propriété « dissipative ». L'analyse de l'erreur du schéma de discrétisation, en temps par le schéma présenté ci-dessous et en espace par une méthode de Galerkin sur une base d'ondelettes, est proposée par Petersdorff & al [PS04].

2.6.2 Schéma de Galerkin discontinu en temps

Soient $\mathcal{H}$ un espace de fonctions inclus dans $L^2(\Omega)$ et $\mathcal{H}'$ l'espace dual associé. Nous supposons que $\mathcal{H} \subset L^2(\Omega) \subset \mathcal{H}'$ avec injections denses. Nous considérons le problème parabolique sous la forme abstraite suivante

$$u'(t) = F(t, u(t)), \quad t \in (0,T], \quad u(0) = u_0, \tag{2.217}$$

où $F : [0,T] \times \mathcal{H} \to \mathcal{H}'$ vérifie

$$F(t,v) = \mathcal{L}v + f(t), \quad t \in (0,T],\ v \in \mathcal{H}. \tag{2.218}$$

Supposons que $\mathcal{L}$ appartienne à l'ensemble des applications linéaires de $\mathcal{H}$ dans $\mathcal{H}'$.

La forme abstraite (2.217 - 2.218) permet de généraliser la méthode au cas des équations intégro-différentielles, et éventuellement des problèmes non linéaires. Nous admettons l'existence d'une solution faible du problème parabolique dans $u \in L^2(J;\mathcal{H}) \cap H^1\left(J;\mathcal{H}'\right)$.

La discrétisation en temps de (2.217) nécessite d'introduire une partition $\mathcal{M}$ de l'intervalle $J = (0,T)$ en M intervalles de temps $\{I_m\}_{m=1}^M, I_m = (t_{m-1}, t_m),\ 1 \le m \le M$, de taille $\Delta_m = t_m - t_{m-1}$. Les limites de u sur chacun des intervalles I_m sont définies par

$$\begin{aligned} u_m^+ &= \lim_{s \to 0^+} u(t_m + s), \quad 0 \le m \le M-1, \\ u_m^- &= \lim_{s \to 0^+} u(t_m - s), \quad 1 \le m \le M, \\ [u]_m &= u_m^+ - u_m^-, \quad 1 \le m \le M-1. \end{aligned} \tag{2.219}$$

Proposition 2.30 *[SS01] La solution faible $u \in L^2(J;\mathcal{H}) \cap H^1\left(J;\mathcal{H}'\right)$ de (2.217) satisfait*

$$B_{DG}(u,v) = \left(u_0, v_0^+\right) + \sum_{m=1}^{M} \int_{I_m} (g,v)_{\mathcal{H}' \times \mathcal{H}}\, dt, \tag{2.220}$$

pour toute fonction

$$v \in \mathcal{C}(\mathcal{M};\mathcal{H}) = \left\{ u : J \to \mathcal{H} \middle|\ u_{|I_m} \in \mathcal{C}^0\left(\overline{I_m};\mathcal{H}\right),\ I_m \in \mathcal{M} \right\}. \tag{2.221}$$

La forme bilinéaire B_{DG} est donnée par

$$B_{DG}(u,v) = \sum_{m=1}^{M} \int_{I_m} \left\{ \left(u', v\right)_{\mathcal{H}' \times \mathcal{H}} + a(u,v) \right\} dt + \left(u_0^+, v_0^+\right)_{L^2} + \sum_{m=2}^{M} \left([u]_{m-1}, v_{m-1}\right)_{L^2}, \tag{2.222}$$

où a est la forme bilinéaire associée à l'opérateur $\mathcal{L}$.

Soit $\mathcal{S}^r(\mathcal{M})$ l'espace discret des fonctions polynomiales d'ordre au plus r sur chacun des intervalles I_m,

$$\mathcal{S}^r(\mathcal{M};\mathcal{H}) = \left\{ u : J \to \mathcal{H} \middle|\ u_{I_m} \in \mathcal{P}^r(I_m;\mathcal{H}),\ 1 \le m \le M \right\}.$$

Le schéma en temps de Galerkin discontinu consiste à :
trouver $U \in \mathcal{S}^r(\mathcal{M};\mathcal{H})$ telle que

$$B_{DG}(U,W) = \left(u_0, W_0^+\right) + \sum_{m=1}^{M} \int_{I_m} (g,W)_{\mathcal{H}',\mathcal{H}}\, dt, \quad \forall W \in \mathcal{S}^r(\mathcal{M};\mathcal{H}). \tag{2.223}$$

Au pas de temps m, le schéma consiste à résoudre

$$\begin{aligned} &\int_{t_{m-1}}^{t_m} \left[\left(U_m', W\right) + a(U_m, W)\right] dt + \left(U_m(t_{m-1}), W(t_{m-1})\right) \\ &\qquad = \int_{t_{m-1}}^{t_m} (g,W)_{\mathcal{H}' \times \mathcal{H}}\, dt + \left(U_{m-1}(t_{m-1}), W(t_{m-1})\right), \end{aligned} \tag{2.224}$$

où $U_0(t_0)$ est donnée par u_0.

Soit $\{\phi\}_{j=0}^{r_m}$ une base de polynômes de l'espace $\mathcal{P}^{r_m}(-1,1)$. Alors les fonctions de base sont données par $\phi_j \circ F_m^{-1}$ dans l'intervalle I_m où la transformation $F_m : (-1,1) \to I_m$ est donnée par

$$t = F_m(\tau) = \frac{1}{2}(t_{m-1} + t_m) + \frac{\Delta_m}{2}\tau, \quad \Delta_m = t_m - t_{m-1}, \quad \tau \in (0,1).$$

Si nous écrivons $U_m(t)$ et W comme

$$U_m(t) = \sum_{j=0}^{r_m} U_{m,j}\left(\phi_j \circ F_m^{-1}\right)(t), \quad W = \sum_{j=0}^{r_m} W_{m,j}\left(\phi_j \circ F_m^{-1}\right)(t), \tag{2.225}$$

le problème variationnel (2.224) prend la forme suivante :
trouver $(U_{m,j})_{j=0}^{r} \in (\mathcal{H})^{r+1}$ telle que pour tout $(W_{m,j})_{j=0}^{r} \in (\mathcal{H})^{r+1}$,

$$\sum_{i,j=0}^{r}\left\{C_{i,j}(U_{m,j}, W_i) + \frac{\Delta_m}{2}G_{i,j}a(U_{m,j}, W_i)\right\} = \sum_{i=0}^{r}\left\{\frac{\Delta_m}{2}F_{m,i}^1(W_i) + F_{m,i}^2(W_i)\right\}, \tag{2.226}$$

où (voir [WGSS01])

$$F_{m,i}^1(v) = \left(\int_{-1}^{1}(g \circ F_m)\,\phi_i d\tau, v\right)_{L^2}, \quad F_{m,i}^2(v) = \phi_i(-1)(U_{m-1}(t_{m-1}, v)_{L^2} \tag{2.227}$$

$$C_{i,j} = \int_{-1}^{1}\phi_j'\phi_i d\tau + \phi_j(-1)\phi_i(-1), \quad G_{i,j} = \int_{-1}^{1}\phi_j'\phi_i d\tau.$$

Remarque 2.25 *Dans le cas où les coefficients ne dépendent pas de t, nous obtenons pour les fonctions de base $\phi_i(\tau) = \left(i + \frac{1}{2}\right)^{\frac{1}{2}} L_i(\tau)$ où L_i correspond au $i$ème polynôme de Legendre sur $(-1,1)$ normalisé pour que $L_i(1) = 1$, alors $G = I$ dans (2.226) et*

$$C_{i,j} = \left(i + \frac{1}{2}\right)^{\frac{1}{2}}\left(j + \frac{1}{2}\right)^{\frac{1}{2}}\sigma_{i,j}, \quad \sigma_{i,j} = (-1)^{i+j} \text{ si } i > j \text{ et } 1 \text{ sinon.} \tag{2.228}$$

La condition initiale intervient dans le terme $F_{1,i}^2(v)$. La discrétisation en espace, par exemple sur une base d'ondelettes Sparse, du problème (2.226), nécessite de calculer $F_{1,i}^2(W_i)$ pour chaque fonction de base v. Nous appliquerons la méthode de calcul du second membre présentée dans § 2.4.3.3.

Il est également possible de formuler le schéma afin de pouvoir calculer explicitement le terme $F_{1,i}^2(W_i)$ dans le cas d'une base sparse.

Proposition 2.31 *Soit $V_m(t) = U_m(t) - U_{m-1}(t_{m-1})$, alors la discrétisation (2.224) devient*

$$\int_{t_{m-1}}^{t_m}\left[\left(V_m', W\right) + a(V_m, W)\right]dt + (V_m(t_{m-1}), W(t_{m-1})) \tag{2.229}$$

$$= \int_{t_{m-1}}^{t_m}(g, W)_{\mathcal{H}'\times\mathcal{H}} - \sum_{j=1}^{m-1}\int_{t_{m-1}}^{t_m}a(V_j(t_{j-1}), W)\,dt - \int_{t_{m-1}}^{t_m}a(U_0(t_0), W)\,dt.$$

Le problème variationnel (2.229) prend la forme suivante :
trouver $(V_{m,j})_{j=0}^{r} \in (\mathcal{H})^{r+1}$ *telle que pour tout* $(W_{m,j})_{j=0}^{r} \in (\mathcal{H})^{r+1}$,

$$\sum_{i,j=0}^{r} \left\{ C_{i,j}\left(V_{m,j}, W_i\right) + \frac{\Delta_m}{2} G_{i,j} a\left(U_{m,j}, W_i\right) \right\} = \sum_{i=0}^{r} \left\{ \frac{\Delta_m}{2} F_{m,i}^{1}\left(W_i\right) + \sum_{j=0}^{m-1} F_{m,i,j}^{2}\left(W_i\right) \right\}, \tag{2.230}$$

où

$$F_{m,i,j}^{2}(v) = -\int_{-1}^{1} \phi_i d\tau \; a\left(V_j(t_{j-1}), v\right), \quad j > 0, \tag{2.231}$$

$$F_{m,i,0}^{2}(v) = -\int_{-1}^{1} \phi_i d\tau \; a\left(u_0, v\right). \tag{2.232}$$

Preuve

$$\begin{aligned} \int_{t_{m-1}}^{t_m} \left[\left(V_m', W\right) + a\left(V_m, W\right)\right] dt + \int_{t_{m-1}}^{t_m} a\left(U_{m-1}(t_{m-1}), W\right) dt \\ + \left(V_m(t_{m-1}), W(t_{m-1})\right) + \left(U_{m-1}(t_{m-1}), W(t_{m-1})\right) \\ = \int_{t_{m-1}}^{t_m} (g, W)_{\mathcal{H}' \times \mathcal{H}} + \left(U_{m-1}(t_{m-1}), W(t_{m-1})\right), \end{aligned} \tag{2.233}$$

où $V_0(t_0)$ est donnée par 0.

$$\begin{aligned} \int_{t_{m-1}}^{t_m} \left[\left(V_m', W\right) + a\left(V_m, W\right)\right] dt + \left(V_m(t_{m-1}), W(t_{m-1})\right) \\ = \int_{t_{m-1}}^{t_m} (g, W)_{\mathcal{H}' \times \mathcal{H}} - \int_{t_{m-1}}^{t_m} a\left(U_{m-1}(t_{m-1}), W\right) dt. \end{aligned} \tag{2.234}$$

■

Remarque 2.26 *Le calcul de* $F_{m,i,0}^{2}(v)$ *peut dans certains cas s'avérer être plus facile que celui de* $F_{m,i}^{2}(v)$. *Prenons l'exemple de l'équation de la chaleur en dimension* 1,

$$a\left(u_0, v\right) = \left(\Delta u_0, v\right) = \left(u_0, \Delta v\right). \tag{2.235}$$

En choisissant une base d'ondelettes d'ordre primal $p = 2$ *comme base de discrétisation de l'espace des fonctions d'échelles, alors* $\Delta\psi_{\ell,i}$ *est une combinaison linéaire de masse de Dirac et* $(u_0, \Delta\psi_{\ell,i})$ *est calculé par interpolation.*

Deuxième partie

Approximation sparse appliquée à l'évaluation d'options

Chapitre 3

Introduction

L'évaluation d'options constitue un des nombreux problèmes en mathématiques financières. Ce domaine a connu un fort développement depuis les années soixante-dix marquées par le célèbre modèle de Merton, Black and Scholes [Mer73, BS73a]. Une première modélisation des phénomènes de marché est présentée dès 1900 dans la thèse de Bachelier[Bac95].

Actuellement, les actions, obligations, matières premières, etc... sont utilisées comme sous-jacents pour des milliers de produits dérivés complexes.

Les options vanilles sont, de toute évidence, un des exemples les plus simples de ces produits. Un *call européen* , (*resp. put*) est un contrat qui donne le droit à son détenteur d'acheter (*resp.* de vendre) un nombre d'actions à un prix fixe K à la date future T. L'action se nomme *sous-jacent*, le prix fixe K le *Strike* et la date d'échéance de l'option la *maturité*. Le sous-jacent vaut S_T à l'instant T. L'option sera exercée si $S_T > K$ (*resp.* $K > S_T$), générant un profit $S_T - K$ (*resp.* $K - S_T$). Sinon, l'option ne sera pas exercée, et le profit sera égal à 0.

En supposant un marché liquide et une absence d'opportunité d'arbitrage (*i.e.* il n'est pas possible de faire un bénéfice instantané sans prendre de risque), le prix de l'option au temps T (*le payoff*) est donné par $C_0(S_T) = (S_T - K)^+$ (*resp.* $P_0(S_T) = (K - S_T)^+$). Plus généralement, le prix de l'option à maturité est une fonction de S_T, nommée *fonction payoff*. D'autres fonctions payoff sont étudiées : par exemple, dans le cas d'un call digital, la fonction payoff est donnée par $\mathbb{1}_{(S_T>K)}$. Le prix du sous-jacent à l'instant t ou *prix spot* sera noté S_t.

En considérant l'hypothèse selon laquelle le marché est constitué de deux sous-jacents, le premier noté S_t sur lequel porte l'option et le second, un actif sans risque dont le prix est donné en fonction d'un taux d'intérêt instantané r supposé connu, le prix de l'option à l'instant t s'écrit :

$$P_t = e^{-r(T-t)}\mathbb{E}\left[(S_T - K)^+ \middle| \mathcal{F}_t\right]. \tag{3.1}$$

Ces hypothèses sont vérifiées dans le modèle de Black & Scholes. Selon ce modèle, $S_{t+\delta t}$ évolue en partant de S_t suivant une loi log-normale de *tendance* μ et de variance σ^2 :

$$S_{t+\delta t} = S_t(1 + \mu\delta t) + \sigma S_t N(0, \delta t),$$

où σ est la *volatilité* et $N(0, v)$ une loi normale de moyenne nulle et de variance v. Dans ce cas, $S_{t+\delta t} - S_t$ est indépendant des évènements intervenant avant t. La dynamique

ne dépendant pas de l'incrément en temps δt, il est possible de proposer un processus stochastique continu en temps S_t

$$dS_t = S_t(\mu dt + \sigma dW_t), \tag{3.2}$$

où W_t est un processus Brownien. Comme S_t est un processus de Markov, il existe une fonction à deux variables P, nommée *la fonction prix*, telle que $P_t = P(S_t, t)$ et telle que P soit solution de l'équation aux dérivées partielles (EDP) :

$$\frac{\partial P}{\partial t} + \frac{\sigma^2 S^2}{2}\frac{\partial^2 P}{\partial S^2} + rS\frac{\partial P}{\partial S} - rP = 0, \tag{3.3}$$

pour $t \in [0, T)$ et $S > 0$.

De nombreuses méthodes numériques permettent l'évaluation d'une option européenne : les méthodes de Monte-Carlo, les méthodes d'arbres et les méthodes déterministes basées sur l'EDP. De manière évidente, les méthodes de résolution de l'équation (3.3) sont bien connues. Cependant, des difficultés supplémentaires peuvent être rencontrées :
Tout d'abord, les traders demandent une réponse précise (une erreur de l'ordre de $1e-4$ fois le spot) dans un temps très court (inférieur à la seconde pour une option vanille et jusqu'à l'ordre de la minute pour des produits plus complexes).
Ensuite, la formulation EDP de certains contrats peut s'avérer être complexe :

- les options dites « path-dependent » pour lesquelles des variables de conditionnement sont introduites (par exemple une variable $M_t = \frac{1}{t}\int_0^t S_u du$ dans le cas d'une option asiatique sur la moyenne du sous-jacent) ;
- les options américaines nécessitent la résolution d'inéquations variationnelles ;
- les modèles à volatilité stochastique aboutissent à des EDP en dimension d supérieure à 1 ;
- les options multi sous-jacents nécessite la résolution d'une équation en dimension d où d est le nombre de sous-jacents sur lesquels porte l'option.

Enfin, les modèles dont la dynamique est décrite par des processus de Lévy conduisent à la résolution d'équations intégro-différentielles, [CT03].

Dans ce qui suit, deux types de dynamiques sont abordées : la première dans un modèle à volatilité locale (la volatilité peut dépendre de la valeur du sous-jacent) et la seconde dans un modèle à volatilité stochastique multi-facteurs. Pour ce dernier, une généralisation au cas de processus à saut de type Lévy Ito est proposé. Les contrats étudiés sont des contrats de type européen éventuellement multi sous-jacents.

Les méthodes classiques de résolution d'EDP sont les suivantes :

1. méthodes différences finies, [RM94],
2. méthodes d'éléments finis, [Cia78, Cia91, ZT00],
3. méthodes de volumes finis, [EGH00],
4. méthodes spectrales, [Qua91, BM97],

Notre objectif est de mettre en évidence les cas pour lesquels une méthode des *Sparse Grid* est une « bonne méthode » en terme de précision et de ressources/temps de calcul. Le lien entre les différentes techniques de discrétisation sur une base sparse et la formulation variationnelle du problème aux limites a été établi au chapitre précédent. Cette formulation

fera l'objet d'une étude dans le cadre du modèle à volatilité stochastique présenté au chapitre 5. Une généralisation possible au cas de dynamique à saut de notre modèle à volatilité stochastique sera développée. Cette dynamique suit un processus de Lévy-Ito.

Afin de préciser différentes notions, la suite de ce chapitre donne une description du modèle de Black & Scholes, l'EDP et les formules semi-analytiques qui en découlent.

3.1 L'équation de Black & Scholes

Le modèle de Black & Scholes, plus précisément une extension de ce modèle, suppose l'existence d'un actif sans risque dont le prix au temps t est $S_t^0 = S_0^0 \exp\left(\int_0^t r(s)ds\right)$, où $r(t)$ est le taux d'intérêt. Selon ce modèle, le prix de l'actif risqué satisfait l'équation différentielle stochastique :

$$dS_t = S_t(\mu dt + \sigma_t dW_t), \tag{3.4}$$

où W_t est le mouvement brownien standard sur l'espace de probabilité $(\Omega, \mathcal{A}, \mathbb{P})$. Ici, σ_t est une fonction de t dans le cas d'une *volatilité déterministe* ou de t et S_t dans le cas d'un modèle à *volatilité locale*. En supposant le marché liquide et une absence d'opportunité d'arbitrage, le prix de l'option à l'instant t est donné par

$$P_t = \exp\left(-\int_t^T r(s)ds\right) \mathbb{E}^* \left(P_0(S_T)|F_t\right), \tag{3.5}$$

où l'espérance $\mathbb{E}^*$ est définie sous la *probabilité risque neutre* $\mathbb{P}^*$ (probabilité équivalente à $\mathbb{P}$ et sous laquelle $dS_t = S_t(rdt + \sigma_t dW_t)$, W_t est le mouvement brownien standard sous la probabilité $\mathbb{P}^*$ et F_t est la filtration naturelle de W_t).
A partir de l'Eq.(3.5) et en sachant que S_t est un processus de Markov, le prix de l'option P_t est une fonction de t et de S_t, *i.e.* il existe une fonction de deux variables P, appelée *la fonction prix*, telle que $P_t = P(S_t, t)$.
Si $\sigma_t = \sigma(S_t, t)$, où σ est une fonction suffisamment régulière, la fonction prix P est solution de l'EDP parabolique rétrograde

$$\frac{\partial P}{\partial t} + \frac{\sigma^2(S,t)S^2}{2}\frac{\partial^2 P}{\partial S^2} + r(t)S\frac{\partial P}{\partial S} - r(t)P = 0 \tag{3.6}$$

pour $t \in [0, T)$ et $S > 0$ avec la condition terminale

$$P(S, t = T) = P_0(S) \tag{3.7}$$

pour $S > 0$.

Le problème (3.6), (3.7) est un problème parabolique avec condition finale en temps.

3.1.1 La Formule de Black-Scholes

En notant $P(S,t)$ le prix de l'option de maturité T, de fonction payoff P_0 et en supposant que r et $\sigma > 0$ sont constants, la formule de Black & Scholes est

$$P(S,t) = e^{-r(T-t)}\mathbb{E}^*\left[P_0(Se^{r(T-t)}e^{\sigma(W_T - W_t) - \frac{\sigma^2}{2}(T-t)})\right]. \tag{3.8}$$

En remarquant que, sous $\mathbb{P}^*$, $W_T - W_t$ est une distribution gaussienne centrée de variance $T-t$,

$$P(S,t) = \frac{1}{\sqrt{2\pi}} e^{-r(T-t)} \int_{\mathbb{R}} P_0(Se^{(r-\frac{\sigma^2}{2})(T-t)+\sigma x\sqrt{T-t}}) e^{-\frac{x^2}{2}} dx. \tag{3.9}$$

Dans le cas d'une option vanille, en notant C le prix d'un call et P celui d'un put, une formule explicite peut être déduite de (3.5). Dans le cas du call,

$$\begin{aligned} C(S,t) &= \frac{1}{\sqrt{2\pi}} \int_{-d_2}^{+\infty} \left(Se^{-\frac{\sigma^2}{2}(T-t)+\sigma x\sqrt{T-t}} - Ke^{-r(T-t)} \right) e^{-\frac{x^2}{2}} dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{d_2} \left(Se^{-\frac{\sigma^2}{2}(T-t)-\sigma x\sqrt{T-t}} - Ke^{-r(T-t)} \right) e^{-\frac{x^2}{2}} dx, \end{aligned} \tag{3.10}$$

où

$$d_1 = \frac{\log(\frac{S}{K}) + (r + \frac{\sigma^2}{2})(T-t)}{\sigma\sqrt{T-t}} \quad \text{et} \quad d_2 = d_1 - \sigma\sqrt{T-t}. \tag{3.11}$$

La fonction de répartition de la loi Gaussienne est introduite :

$$\mathcal{N}(d) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{d} e^{-\frac{x^2}{2}} dx. \tag{3.12}$$

La *formule de Black & Scholes* est obtenue à partir de (3.10-3.11).

Proposition 3.1 *Si σ et r sont constants, le prix d'un call s'écrit*

$$C(S,t) = S\mathcal{N}(d_1) - Ke^{-r(T-t)}\mathcal{N}(d_2), \tag{3.13}$$

et le prix d'un put s'écrit

$$P(S,t) = -S\mathcal{N}(-d_1) + Ke^{-r(T-t)}\mathcal{N}(-d_2), \tag{3.14}$$

où d_1 et d_2 sont donnés par (3.11) et $\mathcal{N}$ est donnée par (3.12).

Remarque 3.1 *Si r est une fonction du temps, (3.11) peut être remplacée par*

$$d_1 = \frac{\log(\frac{S}{K}) + \int_t^T r(\tau)d\tau + \frac{\sigma^2}{2}(T-t)}{\sigma\sqrt{T-t}} \quad \text{et} \quad d_2 = d_1 - \sigma\sqrt{T-t}. \tag{3.15}$$

Remarque 3.2 *Si σ est une fonction du temps, (3.11) peut être remplacée par*

$$d_1 = \frac{\log(\frac{S}{K}) + \int_t^T r(\tau)d\tau + \frac{1}{2}\int_t^T \sigma^2(\tau)d\tau}{\sqrt{\int_t^T \sigma^2(\tau)d\tau}} \quad \text{et} \quad d_2 = d_1 - \sqrt{\int_t^T \sigma^2(\tau)d\tau}. \tag{3.16}$$

Les formules du delta $\Delta(t,s) = \partial_s P(t,s)$ et du gamma $\Gamma(t,s) = \partial_{ss} P(t,s)$ sont nécessaires pour la suite :

$$\Delta(t,s) = \sqrt{\frac{1}{2\pi}} \int_{-\infty}^{d_1} e^{-\frac{1}{2}u^2} du, \quad \Gamma(t,s) = \sqrt{\frac{1}{2\pi}} \frac{1}{s\sigma\sqrt{t}} e^{-\frac{1}{2}d_1^2}, \quad s^2\partial_{ss}P(t,s) = \sqrt{\frac{1}{2\pi}} \frac{s}{\sigma\sqrt{t}} e^{-\frac{1}{2}d_1^2}. \tag{3.17}$$

En appliquant le changement de variable $x = \ln s/K$, cette dernière quantité devient :

$$(\partial_{xx}P - \partial_x P) = \sqrt{\frac{1}{2\pi}} \frac{e^x}{\sigma\sqrt{t}} e^{-\frac{1}{2}d_1^2}. \tag{3.18}$$

Chapitre 4

Équations paraboliques en finance

Ce chapitre présente les équations aux dérivées partielles les plus fréquemment rencontrées dans les problèmes d'évaluation de prix d'options.

La première partie de ce chapitre reprend les résultats bien connus sur les liens entre les Équations Différentielles Stochastiques (EDS) et les équations paraboliques.

Les problèmes liés à la dégénérescence du *générateur infinitésimal* du processus de diffusion, c.-à-d. le cas où cet opérateur n'est plus *uniformément elliptique*, sont ensuite abordés. Cette dégénérescence apparaît lorsque le nombre de mouvements browniens, intervenant dans le système d'EDS décrivant la dynamique du processus de diffusion, est strictement inférieur à la dimension de ce processus. Les *solutions de viscosité*, [Bar94], permettent l'étude de ces équations. Cependant, les méthodes de résolution numérique sur des *Sparse Grid* sont liées à la *formulation variationnelle* de ces problèmes. Il est donc nécessaire d'introduire une *formulation faible* dans le cas d'opérateurs dégénérés. Il convient également d'insister sur le choix des conditions aux bords à imposer au problème afin que la solution de celui-ci soit bien définie. Bien que les équations aux dérivées partielles intervenant dans l'évaluation d'options sont le plus souvent posées dans des domaines non bornés, le choix des conditions aux bords pour des domaines tronqués s'avère être important en pratique pour la résolution numérique.

4.1 Processus stochastique et équations aux dérivées partielles

Définition 4.1 (Générateur infinitésimal) *Soit X un processus de Markov, le générateur infinitésimal associé à ce processus de Markov est l'application $\mathcal{L}$, qui est définie pour toute fonction $f : \mathbb{R}^d \to \mathbb{R}^+$ appartenant au domaine de l'opérateur $\mathcal{L}$ par :*

$$(\mathcal{L}f)(x) = \lim_{t \to 0} \frac{\mathbb{E}[f(X_t)] - f(x)}{t}, \qquad \text{où } X_0 = x.$$

L'ensemble des fonctions f telles que la limite existe au point x est noté $D_{\mathcal{L}}(x)$. Le domaine de l'opérateur, noté $D_{\mathcal{L}}$, est l'ensemble des fonctions f pour lesquelles la limite existe pour tout $x \in \mathbb{R}^n$.

Considérons un processus de diffusion $X_s = \left(X_s^1, \ldots, X_s^d\right) \in \mathbb{R}^d$. Le processus $X_s^{t,x}$ est

défini comme la solution de *l'équation différentielle stochastique* (EDS)

$$dX_s^{t,x} = b\left(s, X_s^{t,x}\right) dt + \sigma\left(s, X_s^{t,x}\right) dW_s, \tag{4.1}$$

où $W_t = \left(W_t^1, \ldots, W_t^p\right)^T$ est un processus gaussien adapté à la filtration $\{\mathcal{F}_t, t \geq 0\}$. b est le vecteur de *drift* et σ la matrice de diffusion.

Proposition 4.1 *Soit la matrice* Ξ *telle que* $\Xi_{i,j} = \frac{1}{2}\left[\sigma\ \sigma^\star\right]_{(i,j)}$, *alors* $X_s^{t,x}$ *est un processus de Markov dont le générateur infinitésimal* $\mathcal{L}$ *est donné par*

$$\mathcal{L}_{t,x} = \sum_{i,j=1}^{d} \Xi_{i,j}(t,x) \frac{\partial^2}{\partial x_i \partial x_j} + \sum_{i=1}^{d} b_i(t,x) \frac{\partial}{\partial x_i}. \tag{4.2}$$

Hypothèse 4.1 *Nous supposons que*

1. $\Xi_{i,j}$ *et* b_i *sont bornées sur* $[0,T] \times \mathbb{R}^d$ *et uniformément lipschitziennes par rapport à* x *sur les sous-ensembles compacts de* $[0,T] \times \mathbb{R}^d$.
2. $\Xi_{i,j}$ *hölderienne par rapport à* t, *uniformément pour tout point* (t,x) *appartenant à* $[0,T] \times \mathbb{R}^d$.
3. $\Xi(t,x)$ *est définie positive en tout point de* (t,x) *de* $[0,T] \times \mathbb{R}^d$, *et*

$$\exists\ \alpha > 0, \quad \zeta^T \Xi(t,x) \zeta \geq \alpha \zeta^T \zeta, \quad \forall\ \zeta \in \mathbb{R}^d,\ \forall\ (t,x) \in [0,T] \times \mathbb{R}^d.$$

Proposition 4.2 *Soit* $p\left(t,x,s,\mathcal{A}\right)$, *définie par*

$$p\left(t,x;s,\mathcal{A}\right) = \mathbb{P}\left\{ X_s \in \mathcal{A} |\, X_t^{t,x} = x \right\} = \mathbb{P}\left\{ X_s^{t,x} \in \mathcal{A} \right\}, \tag{4.3}$$

la fonction de transition *du processus de Markov solution de (4.3). Cette fonction de transition définit une densité :*

$$p\left(t,x;s,\mathcal{A}\right) = \int_{\mathcal{A}} p(t,x;s,y) dy \quad \text{où } t¡s, \tag{4.4}$$

pour tout ensemble Borélien $\mathcal{A}$. *De plus,* $p(t,x,s,y)$ *est la solution fondamentale de* $\dfrac{\partial}{\partial t} + \mathcal{L}_{(t,x)}$. *En d'autres termes, pour toute fonction* f *à support compact, la fonction* u *définie par*

$$u\left(t,x\right) = \int p\left(t,x;s,y\right) f(y) dy$$

est solution de

$$\begin{aligned} &\left(\frac{\partial}{\partial t} + \mathcal{L}_{(t,x)}\right) u\left(t,x\right) = 0 \quad x \in \mathbb{R}^d, \quad t < s \leq T \\ &u\left(t,x\right) = f(x). \end{aligned} \tag{4.5}$$

La fonction de densité définie par (4.3) se nomme *fonction de densité de transition* .

Proposition 4.3 ([BL84]) *Si les coefficients du processus de diffusion vérifient l'hypothèse 4.1, alors la fonction de densité de transition est solution de* l'équation parabolique rétrograde

$$\begin{array}{l}\left(\dfrac{\partial}{\partial t}+\displaystyle\sum_{i,j=1}^{n}\Xi_{i,j}(t,x)\dfrac{\partial^2}{\partial x_i\partial x_j}+\sum_{i=1}^{n}b_i(t,x)\dfrac{\partial}{\partial x_i}\right)p\left(t,x;s,y\right)=0\\ p\left(s,x;s,y\right)=\delta\left(x-y\right).\end{array}\tag{4.6}$$

Si $\dfrac{\partial\Xi_{i,j}}{\partial x_i}$, $\dfrac{\partial^2\Xi_{i,j}}{\partial x_i\partial x_j}$ *et* $\dfrac{\partial b_i}{\partial x_i}$ *sont bornées sur* $[0,T]\times\mathbb{R}^d$ *et hölderiennes par rapport à* x *uniformément sur* $[0,T]\times\mathbb{R}^d$, *alors* p *est solution de l'*équation forward

$$\begin{array}{l}-\left(\dfrac{\partial}{\partial t}+\displaystyle\sum_{i,j=1}^{n}\dfrac{\partial^2}{\partial x_i\partial x_j}\left(\Xi_{i,j}(t,x)\ \cdot\right)-\left(\sum_{i=1}^{n}\dfrac{\partial}{\partial x_i}b_i(t,x)\ \cdot\right)\right)p\left(t,x;s,y\right)=0\\ p\left(t,x;t,y\right)=\delta\left(x-y\right).\end{array}\tag{4.7}$$

4.2 Rappels sur les équations aux dérivées partielles

Considérons un ouvert borné Ω de $\mathbb{R}^d$ de frontière lipschitzienne. Nous étudions les propriétés de l'EDP définie, pour toute fonction $f\in L^2\left(\Omega\right)$, par :

$$-a:D^2u+b\cdot\nabla u+cu=\mathcal{L}u=f,\quad x\in\Omega.\tag{4.8}$$

– La notation D^2u représente la matrice Hessienne $d\times d$ de u. Le coefficient (i,j) de cette matrice correspond à la dérivée partielle d'ordre 2 : $\dfrac{\partial^2u}{\partial x_i\partial x_j}$, $i,j=1,\ldots,d$.

– La notation $A:B=\displaystyle\sum_{i,j=1}^{d}A_{i,j}B_{i,j}$, représente le produit scalaire associé à la norme de Frobenius. Ce produit scalaire peut être défini à l'aide de l'opérateur de trace. En effet, $A:B=trace\left(AB^T\right)$.

Nous supposons que

(H1) le champ de matrices $a:\overline{\Omega}\rightarrow\left(\mathbb{R}^{d\times d}\right)$ est tel que, pour tout $x\in\overline{\Omega}$, $a\left(x\right)$ est symétrique semi-définie positive *i.e.*

$$\forall x\in\overline{\Omega},\quad\forall\zeta\in\mathbb{R}^d,\quad\zeta^Ta\left(\boldsymbol{x}\right)\zeta\geq0.\tag{4.9}$$

(H2) les coefficients $a_{i,j}:\overline{\Omega}\rightarrow\mathbb{R}$, $i,j=1,\ldots,d$ appartiennent à $\mathcal{C}^{\mathbf{2}}\left(\bar{\Omega}\right)$. Le champ de vecteurs $b:\overline{\Omega}\rightarrow\mathbb{R}^d$ est tel que chacun des coefficients b_i appartient à $\mathcal{C}^{\mathbf{1}}\left(\bar{\Omega}\right)$, $\forall i=1,..,d$ et la fonction $c:\overline{\Omega}\rightarrow\mathbb{R}$ appartient à $\mathcal{C}^{\mathbf{0}}\left(\bar{\Omega}\right)$.

Proposition 4.4 *L'hypothèse (H1) permet de montrer que* $\zeta^Ta\left(x\right)\zeta=0\Leftrightarrow a\left(x\right)\zeta=0$.

Preuve Soit $\widetilde{a}$ la racine carrée (à valeurs propres positives) de a. La matrice $\widetilde{a}$ est calculée de la manière suivante : a est diagonalisable dans une base orthonormée

$$a\left(x\right)=q(x)^Td(x)q(x),\tag{4.10}$$

où $q(x)$ est une matrice unitaire et $d(x)$ une matrice diagonale à coefficients positifs ou nuls. Définissons $\widetilde{a}$ par

$$\widetilde{a}\,(x) = q(x)^T \sqrt{d(x)} q(x), \tag{4.11}$$

alors $\widetilde{a}$ est symétrique semi-définie positive et $\widetilde{a}\,(x)\,\widetilde{a}\,(x) = a(x)$.

Soit $\zeta \in \mathbb{R}^d$,

$$\zeta^T a\,(x)\,\zeta = 0 \Leftrightarrow \zeta^T \widetilde{a}\,(x)\,\widetilde{a}\,(x)\,\zeta = 0 \Leftrightarrow (\widetilde{a}\,(x)\,\zeta)^T\,(\widetilde{a}\,(x)\,\zeta) = 0 \Leftrightarrow \widetilde{a}\,(x)\,\zeta = 0 \Rightarrow a\,(x)\,\zeta = 0. \tag{4.12}$$

Réciproquement $a\,(x)\,\zeta = 0 \Rightarrow \zeta^T a\,(x)\,\zeta = 0$. ■

4.2.1 Classification des opérateurs différentiels

Introduisons le *polynôme caractéristique* ou *forme caractéristique* $\alpha\,(x) \in \mathcal{P}\left(\mathbb{R}^d, \mathbb{R}\right)$ de degré ≤ 2, défini à x fixé par

$$\alpha\,(x)\,(\zeta) = \zeta^T a\,(x)\,\zeta, \quad \forall \zeta \in \mathbb{R}^d. \tag{4.13}$$

Sous les hypothèses (4.9), l'équation (4.8) est une *équation aux dérivées partielles* dont la *forme caractéristique* est non négative.

Détaillons certaines *formes caractéristiques* non négatives :

1. Lorsque la matrice a est définie positive, l'équation (4.8) est une équation elliptique.
2. Lorsque la matrice a est nulle et le terme de transport ou de convection b ne change pas de signe, (4.8) est une équation hyperbolique du premier ordre.
3. Lorsque la matrice a est de la forme

$$a = \begin{pmatrix} a_0 & 0 \\ 0 & 0 \end{pmatrix},$$

où a_0 est une matrice de taille $(d-1) \times (d-1)$ symétrique définie positive et $b_d > 0$, alors (4.8) est une équation parabolique, dans laquelle la variable x_d joue le rôle du temps. L'Eq.(3.6) est l'exemple le plus connu en finance, l'EDP de Black & Scholes. Dans ce cas, le rôle particulier de la variable « temps » nous conduit à redéfinir le problème (4.8) sous la forme (en changeant les notations $d = d - 1$)

$$\frac{\partial u}{\partial t} + \mathcal{L}u = f \quad \text{ dans } [0,T] \times \Omega, \tag{4.14}$$

où l'opérateur $\mathcal{L}$ de la forme (4.8) est elliptique.

4. Supposons à présent que le bloc $a^m \in \mathbb{R}^{m \times m}$ défini par $a^m_{i,j} = a_{i,j}, \quad \forall 1 \leq i,j \leq m$, est un champ de matrices de $\Omega \to \mathbb{R}^{m \times m}$ symétriques définies positives. Si $m < d$, alors l'équation est *ultra-parabolique*. La solution u du problème $\mathcal{L}u = 0$ n'a, a priori, aucune propriété de régularité.

   Intéressons-nous à une sous-classe de ces opérateurs *ultra-paraboliques* pour laquelle sont établis des résultats sur la régularité de la solution u.

5. **Définition 4.2 (Opérateur *hypoelliptique*)** *Soit $\mathcal{L}$ un opérateur différentiel du second ordre à coefficients réels de classe $\mathcal{C}^\infty$, défini sur $\mathbb{R}^d$. L'opérateur $\mathcal{L}$ est un opérateur hypoelliptique si, pour un ouvert $U \subset \mathbb{R}^d$, $\mathcal{L}u \in \mathcal{C}^\infty(U)$ implique $u \in \mathcal{C}^\infty(U)$.*

Rappelons le théorème établi par Hörmander [Hör67].

Théorème 4.5 *Soit $\mathcal{L} = X_0 + \sum_{j=1}^{m} X_j^2$, où $X_i = \sum_{j=1}^{d} c_{i,j}(x)\,\partial_{x_j}$ $\left(c_{i,j} \in \mathcal{C}^\infty\left(\mathbb{R}^d\right), i = 0, \dots m\right)$ est un champ de vecteurs (opérateur différentiel homogène de premier ordre). $\mathcal{L}$ est hypoelliptique si et seulement si*

$$\forall x \in \mathbb{R}^d, \quad dim\ Lie\ \langle X_0, \dots, X_m \rangle (x) = d. \tag{4.15}$$

La condition (4.15) signifie qu'il existe un entier r tel que les champs de vecteurs X_i et leurs commutateurs $[X_i, X_j]$, $[[X_i, X_j], X_k]$ (jusqu'au rang r) engendrent $\mathbb{R}^d$ en tout point. Le crochet de Lie de deux champs de vecteurs X, Y d'une variété différentielle appliqué à une fonction v indéfiniment dérivable est donné par $[X, Y] = XY(v) - YX(v)$. La condition (4.15) est connue comme hypothèse d'Hörmander forte.

Exemple 4.1 *L'opérateur de la chaleur est un opérateur* hypoelliptique.

Exemple 4.2 ([MDF05]) *Considérons l'exemple particulier d'une équation ultra-parabolique en dimension d'espace d de la forme*

$$\mathcal{L}u = \sum_{i,j=1}^{m} a_{i,j}(t,x)\,\frac{\partial^2 u}{\partial x_i \partial x_j} u + b \cdot \nabla u + c(t,x)\,u = f, \tag{4.16}$$

avec $m < d$ et

$$b \cdot \nabla u = \sum_{i=1}^{m} b_i(t,x)\,\frac{\partial u}{\partial x_i} + \sum_{i,j=1}^{d-1} b_{i,j} x_i \frac{\partial u}{\partial x_j} - \frac{\partial u}{\partial t}. \tag{4.17}$$

Le champ de matrices a est tel que, pour tout point $(t,x) \in [0,T] \times \Omega$, $a(t,x)$ est symétrique définie positive et

$$\exists\ \alpha > 0, \quad \zeta^T a(t,x) \zeta \geq \alpha \zeta^T \zeta, \quad \forall\ \zeta \in \mathbb{R}^m,\ \forall\ (t,x) \in [0,T] \times \mathbb{R}^d.$$

Si la matrice $B = (b_{i,j})_{1 \leq i,j \leq d}$ vérifie les hypothèses suivantes, alors l'opérateur $\mathcal{L}$ de l'équation (4.16) est hypoelliptique.

Nous supposons que B est de la forme

$$\begin{pmatrix} \star & B_1 & 0 & \dots & 0 \\ \star & \star & B_2 & \ddots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \star & \star & \star & \dots & B_m \\ \star & \star & \star & \dots & \star \end{pmatrix}, \tag{4.18}$$

où les blocs B_j correspondent à une matrice de taille $p_{j-1} \times p_j$ de rang p_j. Nous supposons également que le rang des matrices est croissant.

$$p_0 \geq p_1 \geq \cdots \geq p_m \geq 1, \quad et \quad p_0 + p_1 + \ldots p_m = d.$$

Dans le cas particulier d'une matrice a à coefficients constants et $b_i(t,x) = 0$, [LP94] établit que l'opérateur $\mathcal{L}$ vérifie la condition (4.15) avec $X_i = \sum_{j=1}^{m} a_{i,j}\partial_{x_j}$ et $X_0 = \sum_{i,j=1}^{d-1} b_{i,j}x_i\partial_{x_j} - \partial_t$.

Éléments de preuve *Un rapide calcul permet de montrer que, pour tout $i \in \{1,\ldots,d-1\}$, $[X_i, X_0] = \sum_{i,j=1} a_{ii}b_{i,j}\partial_{x_j}$. La condition (4.18) assure l'existence et l'unicité d'une solution au système d'équations :*

$$M\nabla u = [X_i, X_0](u),$$

où $M_{i,j} = a_{i,i}b_{i,j}$. La condition (4.15) est donc vérifiée. ■

Cette dernière classe des équations *ultra-parabolique* et *hypoelliptique* correspond au problème d'évaluation d'options asiatiques. De tels opérateurs interviennent plus généralement dans le cas de modèles de diffusion dégénérée dont un exemple sera traité dans un chapitre ultérieur.

4.2.2 Équation elliptique dégénérée

Cette partie nous permet d'introduire une technique de régularisation du problème initial afin de définir la solution d'une EDP elliptique dégénérée comme la limite d'une suite d'EDP qui sont elles elliptiques. Nous justifions, d'une part, l'espace de Sobolev sur lequel est défini le problème initial et, d'autre part, des conditions aux bords à imposer pour assurer l'existence d'une solution au sens des *distributions*.

4.2.2.1 Passage sous la forme divergence

Il sera nécessaire d'appliquer l'opérateur de divergence à un champ de matrices a. Cette opération notée $\nabla \cdot a$ est définie par

$$\nabla \cdot a : \mathbb{R}^d \to \mathbb{R}^d, \quad (\nabla \cdot a)_j = \sum_{i=1}^{d} \frac{\partial a_{i,j}}{\partial x_i}.$$

Proposition 4.6 *L'opérateur $\mathcal{L}$ de l'équation (4.8) peut s'écrire sous forme de divergence*

$$\mathcal{L}u = -a : D^2u + b \cdot \nabla u + c\ u = -\nabla \cdot (a\nabla u - b\ u) + (\nabla \cdot a)\nabla u - (\nabla \cdot b)\ u + c\ u. \quad (4.19)$$

En supposant vérifiée l'hypothèse (H2) ($a_{i,j} \in \mathcal{C}^2(\overline{\Omega})$, $b_i \in \mathcal{C}^1(\overline{\Omega})$), l'équation (4.19) peut également prendre la forme suivante

$$\mathcal{L}u = -\nabla \cdot [a\nabla u - (b + \nabla \cdot a)\, u] + [c - \nabla \cdot (b + \nabla \cdot a)]\, u = f. \quad (4.20)$$

Preuve Analysons chacun des opérateurs :

$$\begin{array}{rcl} -a : D^2 u & = & -\nabla \cdot (a \nabla u) + (\nabla \cdot a)\, \nabla u \\ & \underbrace{=}_{(H2)} & -\nabla \cdot (a \nabla u - (\nabla \cdot a)\, u) - \nabla \cdot (\nabla \cdot a)\, u, \\ b \cdot \nabla u & \underbrace{=}_{(H2)} & \nabla \cdot (b\ u) - (\nabla \cdot b)\ \ u. \end{array} \tag{4.21}$$

■

Nous aurons par la suite recours au théorème de trace suivant [GR86]

Théorème 4.7 *Considérons l'espace de Sobolev* $H_{div}(\Omega)$ *défini par*

$$H_{div}(\Omega) = \left\{ v \in \left[L^2(\Omega)\right]^d, \nabla \cdot v \in L^2(\Omega) \right\}. \tag{4.22}$$

Si v *appartient à* $H_{div}(\Omega)$, *alors* v *possède une trace normale dans* $H^{-1/2}(\partial\Omega)$, *plus précisément*

$$\forall v \in H_{div}(\Omega), \quad v_{|\partial\Omega} \cdot n \in H^{-1/2}(\partial\Omega), \quad et \quad \left\| v_{|\partial\Omega} \cdot n \right\|_{H^{-1/2}(\partial\Omega)} \leq C \|v\|_{H_{div}(\Omega)}, \tag{4.23}$$

où C *est une constante indépendante de* v.

Nous déduisons de ce théorème le corollaire suivant.

Corollaire 4.8

$$\begin{array}{l} Si \qquad a\nabla u - (b + \nabla \cdot a)\, u \ \in \left[L^2(\Omega)\right]^d \qquad et \quad \nabla \cdot (a\nabla u - (b + \nabla \cdot a)\, u) \ \in L^2(\Omega), \\ \qquad\qquad alors \qquad\qquad (a\nabla u - (b + \nabla \cdot a)\, u) \cdot n \in \ H^{-1/2}(\partial\Omega). \end{array}$$

Multiplions l'équation (4.20) par une fonction test $v : \Omega \to \mathbb{R}$, puis intégrons par parties. Le corollaire précèdent permet de justifier cette étape puisqu'il assure l'existence du terme de bords. La proposition suivante découle de ce calcul :

Proposition 4.9 *Si* u *vérifie l'équation (4.8) où* $f \in L^2(\Omega)$ *et si* $a\nabla u - (b + \nabla \cdot a)\, u \ \in H_{div}(\Omega)$, *alors*

$$\begin{array}{rcl} \forall v \in H^1(\Omega), \quad \displaystyle\int_\Omega \left[a\nabla u - (b + \nabla \cdot a)\, u\right] \cdot \nabla v + \int_\Omega \left[c - \nabla \cdot (b + \nabla \cdot a)\right] u\ v & & \\ - \left\langle (a\nabla u - (b + \nabla \cdot a)\, u) \cdot n, v \right\rangle_{\left(H^{-1/2}(\partial\Omega), H^{1/2}(\partial\Omega)\right)} & = & \displaystyle\int_\Omega f\ v. \end{array} \tag{4.24}$$

Preuve

$$\begin{array}{l} \displaystyle\int_\Omega \left[-\nabla \cdot (a\nabla u - (b + \nabla \cdot a)\, u) + (c - \nabla \cdot (b + \nabla \cdot a))\, u\right] v = \int_\Omega \left[a\nabla u - (b + \nabla \cdot a)\, u\right] \cdot \nabla v \\ \quad + \displaystyle\int_\Omega (c - \nabla \cdot (b + \nabla \cdot a))\, u\ v - \left\langle (a\nabla u - (b + \nabla \cdot a)\, u) \cdot n, v \right\rangle_{\left(H^{-1/2}(\partial\Omega), H^{1/2}(\partial\Omega)\right)}. \end{array}$$

■

4.2.2.2 Problème régularisé

Nous régularisons l'équation (4.8) en ajoutant à l'opérateur $\mathcal{L}$ un opérateur de diffusion $-\varepsilon\Delta$. Nous considérons dans ce paragraphe le problème suivant :

$$\mathcal{L}_\varepsilon u_\varepsilon = -\varepsilon\Delta u_\varepsilon - a : D^2 u_\varepsilon + b \cdot \nabla u_\varepsilon + c u_\varepsilon = f, \quad x \in \Omega. \tag{4.25}$$

La proposition 4.9 s'applique également au problème (4.25), elle implique la proposition suivante.

Proposition 4.10 *Si u_ε vérifie l'équation (4.25) où $f \in L^2(\Omega)$ et si $u_\varepsilon \in H^1(\Omega)$, alors $\forall v \in H^1(\Omega)$*

$$\begin{aligned} \int_\Omega \left[(\varepsilon\, \mathcal{I} + a)\, \nabla u_\varepsilon - (b + \nabla \cdot a)\, u_\varepsilon\right] \cdot \nabla v + \int_\Omega \left[c - \nabla \cdot (b + \nabla \cdot a)\right] u_\varepsilon\ v \qquad (4.26) \\ - \left\langle \left((\varepsilon\mathcal{I} + a)\, \nabla u_\varepsilon - (b + \nabla \cdot a)\, u_\varepsilon\right) \cdot n, v \right\rangle_{\left(H^{-1/2}(\partial\Omega), H^{1/2}(\partial\Omega)\right)} \quad = \quad \int_\Omega f\ v. \end{aligned}$$

4.2.2.3 Conditions aux limites

La *forme caractéristique* permet de définir le sous-ensemble de $\partial\Omega$ noté Σ tel que

$$\Sigma = \left\{ x \in \partial\Omega \,\middle|\, \alpha(x)\,(n(x)) = n(x)^T\, a(x)\, n(x) > 0 \right\}, \tag{4.27}$$

où $n(x)$ est le vecteur unitaire dans la direction normale à $\partial\Omega$, dirigé de l'intérieur vers l'extérieur de Ω. La proposition 4.4 permet de montrer que

$$\Sigma = \{ x \in \partial\Omega \,|\, a(x)\, n(x) \neq 0 \}. \tag{4.28}$$

Notons

$$\Gamma = \partial\Omega \backslash \Sigma = \{ x \in \partial\Omega \,|\, a(x)\, n(x) = 0 \}.$$

Cet ensemble est divisé en deux parties

$$\Gamma = \Gamma_+ \cup \Gamma_- \quad \text{avec} \quad \Gamma_+ \cap \Gamma_- = \emptyset. \tag{4.29}$$

Ces deux parties sont définies à l'aide de la fonction de *Fichera*

$$\beta : x \in \overline{\Omega} \to \mathbb{R}, \quad \beta(x) = (\nabla \cdot a + b) \cdot n(x), \tag{4.30}$$

par

$$\Gamma_+ = \left\{ x \in \Gamma \middle| \beta(x) \geq 0 \right\}, \quad \Gamma_- = \left\{ x \in \Gamma \middle| \beta(x) < 0 \right\}. \tag{4.31}$$

L'ensemble Σ est également divisé en deux sous-ensembles disjoints

$$\Sigma = \Sigma_D \cup \Sigma_N.$$

Nous considérons les conditions aux limites suivantes :
– pour le problème initial (4.8)

$$\begin{cases} u = h & \text{sur } \Sigma_D, \\ \left[a\nabla u - \dfrac{1}{2}(b + \nabla \cdot a)\, u\right] \cdot n = g & \text{sur } \Sigma_N, \\ u = \widetilde{h} & \text{sur } \Gamma_- \ ; \end{cases} \tag{4.32}$$

– pour le problème régularisé (4.25)

$$\left\{\begin{array}{cl} u_\varepsilon = h & \text{sur } \Sigma_D, \\ \left[(\varepsilon\,\mathcal{I}+a)\,\nabla u_\varepsilon - \dfrac{1}{2}\,(b+\nabla\cdot a)\,u_\varepsilon\right]\cdot n = g & \text{sur } \Sigma_N, \\ u_\varepsilon = \widetilde{h} & \text{sur } \Gamma_-, \\ [\varepsilon\,\nabla u_\varepsilon - \gamma\,(b+\nabla\cdot a)\,u_\varepsilon]\cdot n = 0 & \text{sur } \Gamma_+\ , \end{array}\right. \tag{4.33}$$

avec $\gamma \leq \dfrac{1}{2}$.

Hypothèse 4.2 *Nous supposons qu'il existe une fonction de relèvement* $\widetilde{u} \in H^2(\Omega)$ *telle que*

$$\widetilde{u} = h \text{ sur } \Sigma_D, \quad \widetilde{u} = \widetilde{h} \text{ sur } \Gamma_-, \quad \widetilde{u} = 0 \text{ sur } \Gamma_+. \tag{4.34}$$

Cette hypothèse est toujours vérifiée dans le cas d'une condition de Dirichlet homogène *i.e.* $h = 0,\ \widetilde{h} = 0$.

Les conditions vérifiées par :

⋆ $e = u - \widetilde{u}$ sont

$$\left\{\begin{array}{cl} e = 0 & \text{sur } \Sigma_D \cup \Gamma_- \\ \left[a\nabla e - \dfrac{1}{2}\,(b+\nabla\cdot a)\,e\right]\cdot n = \widetilde{g} & \text{sur } \Sigma_N\ , \end{array}\right. \tag{4.35}$$

avec $\widetilde{g} = g - \left[a\nabla\widetilde{u} - \dfrac{1}{2}\,(b+\nabla\cdot a)\,\widetilde{u}\right]\cdot n$ et par

⋆ $e_\varepsilon = u_\varepsilon - \widetilde{u}$ sont

$$\left\{\begin{array}{cl} e_\varepsilon = 0 & \text{sur } \Sigma_D \cup \Gamma_- \\ \left[(\varepsilon\,\mathcal{I}+a)\,\nabla e_\varepsilon - \dfrac{1}{2}\,(b+\nabla\cdot a)\,e_\varepsilon\right]\cdot n = \widetilde{g} - \varepsilon\,\dfrac{\partial\widetilde{u}}{\partial n} & \text{sur } \Sigma_N \\ [\varepsilon\,\nabla e_\varepsilon - \gamma\,(b+\nabla\cdot a)\,e_\varepsilon]\cdot n = -\varepsilon\,\dfrac{\partial\widetilde{u}}{\partial n} & \text{sur } \Gamma_+\ . \end{array}\right. \tag{4.36}$$

4.2.2.4 Formulation variationnelle du problème aux limites régularisé

Soit

$$\mathcal{W}_0 = \left\{w \in H^1(\Omega)\,,\quad w_{\big|_{\Sigma_D\cup\Gamma_-}} = 0\right\}. \tag{4.37}$$

La *formulation faible* du problème régularisé (4.25) avec les conditions aux limites (4.36) consiste à trouver $e_\varepsilon \in \mathcal{W}_0$ tel que $\forall w \in \mathcal{W}_0$

$$\begin{aligned} \int_\Omega [(\varepsilon\,\mathcal{I}+a)\,\nabla e_\varepsilon - (b+\nabla\cdot a)\,e_\varepsilon]\cdot\nabla w + \int_\Omega [c - \nabla\cdot(b+\nabla\cdot a)]\,e_\varepsilon\ w \\ +\frac{1}{2}\int_{\Sigma_N}(b+\nabla\cdot a)\,e_\varepsilon\cdot n\ w + (1-\gamma)\int_{\Gamma_+}(b+\nabla\cdot a)\,e_\varepsilon\cdot n\ w \\ = \int_\Omega \widetilde{f}\ w + \int_{\Sigma_N}\widetilde{g}w - \int_{\Gamma_+\cup\Sigma_N}\varepsilon\frac{\partial\widetilde{u}}{\partial n}w. \end{aligned} \tag{4.38}$$

Nous supposons que les fonctions $\widetilde{g}$ et $\dfrac{\partial \widetilde{u}}{\partial n}$ sont suffisamment régulières pour que les intégrales dans lesquelles ces fonctions interviennent soient définies.

Introduisons la *forme bilinéaire* $B_\varepsilon : \mathcal{W}_0 \times \mathcal{W}_0 \to \mathbb{R}$,

$$\begin{aligned} B_\varepsilon(z,w) &= \int_\Omega \left[(\varepsilon\,\mathcal{I}+a)\,\nabla z - (b+\nabla\cdot a)\,z\right]\cdot\nabla w + \int_\Omega \left[c-\nabla\cdot(b+\nabla\cdot a)\right] z\; w \\ &\quad + \frac{1}{2}\int_{\Sigma_N} (b+\nabla\cdot a)\cdot n\; z\; w + (1-\gamma)\int_{\Gamma_+} (b+\nabla\cdot a)\, z\cdot n\; w, \end{aligned} \tag{4.39}$$

où les intégrations sur Σ_N et Γ_+ sont bien définies.

Lemme 4.11 *La* forme bilinéaire B_ε *est continue sur* $\mathcal{W}_0 \times \mathcal{W}_0$.

Lemme 4.12 *Si il existe une constante* $\underline{c} > 0$ *telle que*

$$\forall x \in \overline{\Omega}, \quad c(x) - \frac{1}{2}\nabla\cdot(b+\nabla\cdot a)\,(x) \geq \underline{c} > 0,$$

alors la forme bilinéaire *est coercive : il existe une constante* $C_\varepsilon > 0$ *telle que*

$$\forall w \in \mathcal{W}_0, \quad B_\varepsilon(w,w) \geq c_\varepsilon \|w\|_{H^1(\Omega)}. \tag{4.40}$$

Preuve

$$\begin{aligned} B_\varepsilon(w,w) &= \int_\Omega (\varepsilon\,\mathcal{I}+a)\,\nabla w\;\nabla w - (b+\nabla\cdot a)\,\frac{\nabla w^2}{2} + \left[c-\nabla\,(b+\nabla\cdot a)\right] w^2 \\ &\quad + \frac{1}{2}\int_{\Sigma_N} (b+\nabla\cdot a)\cdot n w^2 + (1-\gamma)\int_{\Gamma_+} (b+\nabla\cdot a)\cdot n w^2 \\ &= \int_\Omega (\varepsilon\,\mathcal{I}+a)\,\nabla w\cdot\nabla w + \left[c - \frac{1}{2}\nabla\,(b+\nabla\cdot a)\right] w^2 \\ &\quad + \left(\frac{1}{2}-\gamma\right)\int_{\Gamma_+} (b+\nabla\cdot a)\cdot n w^2, \end{aligned} \tag{4.41}$$

et $(b+\nabla\cdot a)\cdot n \geq 0$ sur Γ_+. ■

Remarque 4.1 *Nous disposons également de la minoration suivante, pour tout* $w \in \mathcal{W}_0$,

$$B_\varepsilon(w,w) \geq C\left[\int_\Omega a\nabla w\cdot\nabla w + \underline{c}\int_\Omega w^2 + \left(\frac{1}{2}-\gamma\right)\int_{\Gamma_+} (b+\nabla\cdot a)\cdot n\; w^2\right], \tag{4.42}$$

avec C *indépendante de* ε.

Proposition 4.13 *Le théorème de Lax-Milgram permet de montrer qu'il existe une unique solution* e_ε *du problème aux limites régularisé. De plus,* e_ε *vérifie l'équation*

$$-\left(\varepsilon\;\Delta e_\varepsilon + a : D^2 e_\varepsilon\right) + b\cdot\nabla e_\varepsilon + c\; e_\varepsilon = \widetilde{f} \in L^2(\Omega), \tag{4.43}$$

au sens des distributions, ce qui implique que

$$(\varepsilon\,\mathcal{I}+a)\,\nabla e_\varepsilon \in H_{div}(\Omega). \tag{4.44}$$

Le second membre (4.43) est donné par $\widetilde{f} = f - \mathcal{L}_\varepsilon \widetilde{u}$.

4.2.2.5 Passage à la limite sur le problème régularisé

Introduisons les espaces $\mathcal{V}$ et $\mathcal{V}_0$ définis par

$$\mathcal{V} = \left\{ v \in L^2(\Omega), \quad \int_\Omega a\ \nabla v \cdot \nabla v \leq \infty \right\}, \quad \text{et } \mathcal{V}_0 = \{ v \in \mathcal{V}, \quad v = 0 \ \text{ sur } \Sigma_D \}. \tag{4.45}$$

L'espace $\mathcal{V}$ est muni de la norme $\|\|_{\mathcal{V}}$ définie par

$$\|v\|_{\mathcal{V}}^2 = \|v\|_{L^2(\Omega)}^2 + \int_\Omega a\ \nabla v \cdot \nabla v. \tag{4.46}$$

L'espace $\mathcal{W}_0$ est inclus dans $\mathcal{V}_0$.

Hypothèse 4.3 *Nous supposons que :* $\forall v \in \mathcal{V}_0, v_{|\Sigma_N} \cdot n \in L^2(\Sigma_N)$ *et il existe une constante* $C > 0$ *telle que*

$$\forall v \in \mathcal{V}_0, \quad \|v\|_{L^2(\Sigma_N)} \leq C \|v\|_{\mathcal{V}_0}.$$

Proposition 4.14 *Si les hypothèses 4.2 et 4.3 sont vérifiées, alors il existe une unique fonction* u*, appartenant à* $\mathcal{V}$*, telle que*

$$\begin{cases} -a : D^2 u + b \cdot \nabla u + cu = f, & \left(\mathcal{D}'(\Omega)\right) \\ \left[a \nabla u - \dfrac{1}{2}(b + \nabla \cdot a)\, u \right] \cdot n = g & \left(H^{-1/2}(\Sigma_N)\right) \\ u = h & \left(L^2(\Sigma_D)\right) \\ u = \widetilde{h} & \left(H^{-1/2}(\Gamma_-)\right). \end{cases} \tag{4.47}$$

Cette fonction est obtenue comme la limite de la suite des solutions du problème régularisé (4.25,4.33).

Nous aurons recours au lemme suivant pour la démonstration de cette proposition :

Lemme 4.15 (Estimation a priori) *Si l'hypothèse 4.3 est vérifiée, alors la solution de (4.38) vérifie*

$$\frac{1}{2}\varepsilon \|e_\varepsilon\|_{H^1(\Omega)}^2 + C\|e_\varepsilon\|_{\mathcal{V}}^2 + \left(\frac{1}{2} - \gamma\right) \int_{\Gamma_+} (b + \nabla \cdot a) \cdot n e_\varepsilon^2 \leq \overline{C}. \tag{4.48}$$

où C *et* $\overline{C}$ *sont des constantes indépendantes de* ε*.*

Preuve En choisissant $w = e_\varepsilon$ dans la *formulation faible* (4.38), nous montrons que

$$\begin{aligned} \int_\Omega (\varepsilon\, \mathcal{I} + a) \nabla e_\varepsilon \cdot \nabla e_\varepsilon + \underline{c} \int_\Omega e_\varepsilon^2 + \left(\frac{1}{2} - \gamma\right) \int_{\Gamma_+} (b + \nabla \cdot a) \cdot n e_\varepsilon^2 \\ \leq \left\| \widetilde{f} \right\|_{L^2} \|e_\varepsilon\|_{L^2} + \left| \int_{\Sigma_N} \widetilde{g} e_\varepsilon - \varepsilon \int_{\Gamma_+ \cup \Sigma_N} \frac{\partial \widetilde{u}}{\partial n} e_\varepsilon \right|. \end{aligned} \tag{4.49}$$

Or

$$\left| \int_{\Sigma_N} \left(\widetilde{g} - \varepsilon\, \frac{\partial \widetilde{u}}{\partial n} \right) e_\varepsilon \right| \leq C_1(\widetilde{g}, \widetilde{u}) \|e_\varepsilon\|_{L^2(\Sigma_N)} \underbrace{\leq}_{Hyp:4.3} C \|e_\varepsilon\|_{\mathcal{V}(\Omega)}, \tag{4.50}$$

et

$$\left|\varepsilon \int_{\Gamma_+} \frac{\partial \widetilde{u}}{\partial n} e_\varepsilon\right| \leq C_2\left(\widetilde{u}\right) \varepsilon \|e_\varepsilon\|_{H^1(\Omega)}. \tag{4.51}$$

Nous en déduisons l'existence des constantes positives $C\left(\underline{c}\right)$, C_1 et C_2 telles que

$$\begin{aligned} &\varepsilon\|e_\varepsilon\|^2_{H^1(\Omega)} + C\left(\underline{c}\right)\|e_\varepsilon\|^2_{\mathcal{V}} + \left(\frac{1}{2}-\gamma\right)\int_{\Gamma_+}(b+\nabla\cdot a)\cdot n e_\varepsilon^2 \\ &\qquad \leq \left\|\widetilde{f}\right\|_{L^2(\Omega)}\|e_\varepsilon\|_{L^2(\Omega)} + C_1\|e_\varepsilon\|_{\mathcal{V}} + \varepsilon C_2\|e_\varepsilon\|_{H^1(\Omega)}. \end{aligned} \tag{4.52}$$

En appliquant l'inégalité de Minkowski au membre de droite de (4.52),

$$\left\|\widetilde{f}\right\|_{L^2(\Omega)}\|e_\varepsilon\|_{L^2(\Omega)} + C_1\|e_\varepsilon\|_{\mathcal{V}} \leq \overline{C}\left(\tilde{f}\right)\|e_\varepsilon\|_{\mathcal{V}} \leq \frac{C\left(\underline{c}\right)}{2}\|e_\varepsilon\|^2_{\mathcal{V}} + \overline{C}\left(\tilde{f},\underline{c}\right),$$

où la constante $C\left(\underline{c}\right)$ est celle qui apparaît à gauche de l'inégalité (4.52) et

$$\|C_2\varepsilon\ e_\varepsilon\|_{H^1(\Omega)} \leq \frac{\varepsilon}{2}\|e_\varepsilon\|^2_{H^1(\Omega)} + C_2\varepsilon,$$

nous obtenons

$$\begin{aligned} &\varepsilon\|e_\varepsilon\|^2_{H^1(\Omega)} + C\left(\underline{c}\right)\|e_\varepsilon\|^2_{\mathcal{V}} + \left(\frac{1}{2}-\gamma\right)\int_{\Gamma_+}(b+\nabla\cdot a)\cdot n e_\varepsilon^2 \\ &\qquad \leq \frac{C\left(\underline{c}\right)}{2}\|e_\varepsilon\|^2_{\mathcal{V}} + \overline{C}\left(\tilde{f},\underline{c}\right) + \frac{\varepsilon}{2}\|e_\varepsilon\|^2_{H^1(\Omega)} + C_2\varepsilon. \end{aligned} \tag{4.53}$$

Ce qui permet de conclure. ■

Lemme 4.16 *La suite $\varepsilon\nabla e_\varepsilon$ tend vers 0 dans $L^2(\Omega)$.*

Preuve Ce résultat se déduit du lemme 4.15. En effet, $\varepsilon\|e_\varepsilon\|^2_{H^1(\Omega)} \leq C \Rightarrow \|\varepsilon\nabla e_\varepsilon\|^2_{L^2} \leq C\varepsilon$. ■

Lemme 4.17 *La suite e_ε est une suite de Cauchy dans $\mathcal{V}_0$. Elle admet une limite, notée e appartenant à $\mathcal{V}_0$, lorsque ε tend vers 0.*

Preuve Soient e_ε et $e_{\varepsilon'}$ solutions de la *formulation faible* (4.38), alors

$$\begin{aligned} &\int_\Omega \left[\left(\varepsilon\nabla e_\varepsilon - \varepsilon'\nabla e_{\varepsilon'}\right) + a\nabla\left(e_\varepsilon - e_{\varepsilon'}\right) - (b+\nabla\cdot a)\left(e_\varepsilon - e_{\varepsilon'}\right)\right]\cdot\nabla w \\ &+ \int_\Omega \left[c - \nabla\cdot(b+\nabla\cdot a)\right]\left(e_\varepsilon - e_{\varepsilon'}\right)\ w + \frac{1}{2}\int_{\Sigma_N}(b+\nabla\cdot a)\cdot n\left(e_\varepsilon - e_{\varepsilon'}\right)\ w \\ &+ (1-\gamma)\int_{\Gamma+}(b+\nabla\cdot a)\cdot n\left(e_\varepsilon - e_{\varepsilon'}\right)\ w = -\int_{\Gamma_+\cup\Sigma_N}\left(\varepsilon-\varepsilon'\right)\frac{\partial\widetilde{u}}{\partial n}w. \end{aligned}$$

En reprenant la démonstration de la coercitivité de B_ε, où la fonction test w est choisie telle que $w = e_\varepsilon - e_{\varepsilon'} = \nu_{\varepsilon,\varepsilon'}$ nous obtenons

$$\begin{aligned} &\int_\Omega a\nabla\nu_{\varepsilon,\varepsilon'}\cdot\nabla\nu_{\varepsilon,\varepsilon'} + \left[c - \frac{1}{2}\nabla(b+\nabla\cdot a)\right]\nu^2_{\varepsilon,\varepsilon'} + \left(\frac{1}{2}-\gamma\right)\int_{\Gamma_+}(b+\nabla\cdot a)\cdot n\nu^2_{\varepsilon,\varepsilon'} \\ &= -\left(\int_{\Gamma_+\cup\Sigma_N}\left(\varepsilon-\varepsilon'\right)\frac{\partial\widetilde{u}}{\partial n}\nu_{\varepsilon,\varepsilon'} + \int_\Omega\left(\varepsilon\nabla e_\varepsilon - \varepsilon'\nabla e_{\varepsilon'}\right)\nabla\nu_{\varepsilon,\varepsilon'}\right), \end{aligned} \tag{4.54}$$

avec $(b+\nabla\cdot a)\cdot n>0$ sur Γ_+. Il en découle l'inégalité suivante

$$\left|\int_\Omega a\nabla\nu_{\varepsilon,\varepsilon'}\cdot\nabla\nu_{\varepsilon,\varepsilon'}+\underline{c}\nu^2_{\varepsilon,\varepsilon'}\right| \leq \left(\int_{\Gamma_+\cup\Sigma_N}\left(\varepsilon-\varepsilon'\right)^2\left(\frac{\partial\widetilde{u}}{\partial n}\right)^2\right)^{\frac{1}{2}}\left(\int_{\Gamma_+\cup\Sigma_N}\nu^2_{\varepsilon,\varepsilon'}\right)^{\frac{1}{2}} + \left(\int_\Omega\left(\varepsilon\nabla e_\varepsilon-\varepsilon'\nabla e_{\varepsilon'}\right)^2\right)^{\frac{1}{2}}\left(\int_\Omega\left(\nabla\nu_{\varepsilon,\varepsilon'}\right)^2\right)^{\frac{1}{2}}, \tag{4.55}$$

où la constante $\underline{c}>0$ a été introduite au lemme 4.12. Le terme de droite de cette inégalité tend vers 0 par application du lemme 4.16. La suite e_ε est donc de Cauchy pour la norme $\|\cdot\|_{\mathcal{V}}$. ■

Preuve de la proposition 4.14 Le résultat de la proposition 4.14 se déduit du résultat équivalent sur $e=u-\widetilde{u}$. Ce résultat est établi en passant « à la limite » dans le problème régularisé. Les lemmes précédents nous permettent de montrer les points suivants.

1. La convergence $\varepsilon\nabla e_\varepsilon\to 0$ dans $L^2(\Omega)$ permet de passer à la limite dans l'équation (4.38) en prenant comme fonction test $w\in\mathcal{D}(\Omega)$. Nous obtenons que

$$-a:D^2e+b\cdot\nabla e+c\,e=\widetilde{f}\in L^2(\Omega), \tag{4.56}$$

au sens des distributions, donc $-a:D^2e+b\cdot\nabla e+c\,e\in L^2(\mathbb{R})$ et

$$a\nabla e-(b+\nabla\cdot a)\,e\in H_{\rm div}(\Omega). \tag{4.57}$$

2. La convergence de $e_\varepsilon\to e$ dans $\mathcal{V}_0$ implique que

$$e=0 \text{ sur } \Sigma_D. \tag{4.58}$$

3. La convergence de $e_\varepsilon\to e$ dans $\mathcal{V}_0$ permet de passer à la limite dans la *formulation faible* (4.38). L'hypothèse 4.3 permet de montrer que la convergence dans $\mathcal{V}_0$ implique la convergence $e_{\varepsilon|\Sigma_N}\cdot n\to e_{|\Sigma_N}\cdot n$ dans $L^2(\Sigma_N)$. Pour toute fonction $w\in\mathcal{C}^\infty(\Omega)$ avec $w=0$ sur $\partial\Omega\backslash\Sigma_N$,

$$\int_\Omega\left[a\nabla e-(b+\nabla\cdot a)\,e\right]\cdot\nabla w+\int_\Omega\left[c-\nabla\cdot(b+\nabla\cdot a)\right]e\;w + \frac{1}{2}\int_{\Sigma_N}(b+\nabla\cdot a)\cdot n\;e\;w=\int_\Omega f\;w+\int_{\Sigma_N}\widetilde{g}w. \tag{4.59}$$

Nous déduisons de (4.56) et (4.59) la condition aux limites sur Σ_N :

$$\left[a\nabla e-\frac{1}{2}(b+\nabla a)\,e\right]\cdot n=\widetilde{g},\quad \text{sur } \Sigma_N. \tag{4.60}$$

Cette trace est bien définie d'après (4.57).

4. L'équation (4.57) permet de montrer l'existence de la trace de $a\nabla e-(b+\nabla\cdot a)\,e$ dans $H^{-1/2}(\partial\Omega)$. De plus, la convergence dans $\mathcal{V}_0$ implique

$$\begin{cases} a\nabla e_\varepsilon-(b+\nabla\cdot a)\,e_\varepsilon\to a\nabla e-(b+\nabla\cdot a)\,e & \text{dans } \left(L^2(\Omega)\right)^d \\ \nabla\cdot\left[a\nabla e_\varepsilon-(b+\nabla\cdot a)\,e_\varepsilon\right]\to\nabla\cdot\left[a\nabla e-(b+\nabla\cdot a)\,e\right] & \text{dans } L^2(\Omega). \end{cases} \tag{4.61}$$

Par application du théorème de trace 4.7, nous obtenons

$$[a\nabla e_\varepsilon - (b + \nabla \cdot a)\, e_\varepsilon] \cdot n \to [a\nabla e - (b + \nabla \cdot a)\, e] \cdot n \quad \text{dans } H^{-1/2}(\partial\Omega) \tag{4.62}$$

Sur Γ_-, $a\ n = 0$ et $e_\varepsilon = 0$, ce qui implique $(b + \nabla \cdot a) \cdot n\ e = 0$ sur $H^{-1/2}(\Gamma_-)$ et donc $e = 0$ sur Γ_-.

■

4.2.3 Équation parabolique dégénérée

La technique de régularisation introduite précédemment va nous permettre de définir la solution d'un problème parabolique dégénéré comme la limite d'une suite de problèmes paraboliques. A nouveau, nous justifions, d'une part, l'espace de Sobolev sur lequel est défini le problème initial et, d'autre part, les conditions aux bords vérifiées par le problème « limite ».

Soit u la solution du problème

$$\frac{\partial u}{\partial t} + \mathcal{L}u = f \quad \text{dans }]0,T] \times \Omega, \qquad u(t=0) = u_0 \quad \text{dans } \Omega, \tag{4.63}$$

où l'opérateur $\mathcal{L}$ de la forme (4.14) est elliptique éventuellement dégénéré.

Nous suivons une démarche similaire à celle suivie dans la partie précédente.

Multiplions formellement l'équation (4.63) par une fonction test v régulière et intégrons sur Ω. Après une intégration par partie supposée justifiée,

$$\begin{aligned} \left\langle \frac{\partial u}{\partial t}, v \right\rangle + \int_\Omega [a\nabla u - (b + \nabla \cdot a)\, u] \cdot \nabla v \\ + \int_\Omega [c - \nabla \cdot (b + \nabla \cdot a)]\, u\ v - \int_{\partial\Omega} (a\nabla u - (b + \nabla \cdot a)\, u) \cdot n\ v \ = \ \int_\Omega f\ v. \end{aligned} \tag{4.64}$$

Remarque 4.2 *Nous supposons pour simplifier que les coefficients a et b ne dépendent pas du temps avec $a \in \mathcal{C}^2(\overline{\Omega})$, $b \in \mathcal{C}^1(\overline{\Omega})$ et que $c \in \mathcal{C}^0([0,T] \times \overline{\Omega})$.*

4.2.3.1 Rappels sur les équations paraboliques

Rappelons le théorème fondamental d'existence et d'unicité dans le cas parabolique. Nous aurons, pour cela, besoin des définitions des espaces $L^2((0,T[;X)$ et $\mathcal{C}^k([0,T];X)$.

Définition 4.3 *Soit X un espace de Hilbert défini sur Ω (typiquement $X = L^2(\Omega)$, $H^k(\Omega), \ldots$). Soit un temps final $0 < T \leq \infty$. Soit un entier $k \geq 0$, l'espace de Hölder $\mathcal{C}^k([0,T];X)$ est l'espace des fonctions k fois continûment dérivables de $[0,T]$ dans X. En notant $\|v\|_X$ la norme dans X, alors $\mathcal{C}^k([0,T];X)$ est un espace de Banach pour la norme*

$$\|v\|_{\mathcal{C}^k([0,T];X)} = \sum_{m=0}^{k} \left(\sup_{0 \leq t \leq T} \left\| \frac{\partial^m v}{\partial t^m}(t) \right\|_X \right). \tag{4.65}$$

Soit $L^2(]0,T[;X)$ l'espace des fonctions de $(0,T[$ dans X telles que la fonction $t \to \|v(t)\|_X$ est mesurable et de carré intégrale,

$$\|v\|_{L^2(]0,T[;X)} = \sqrt{\int_0^T \|v(t)\|_X^2 dt} < \infty. \tag{4.66}$$

L'espace $L^2(]0,T[;X)$ *muni de cette norme est un espace de Hilbert pour le produit scalaire*

$$\langle v, w\rangle_{L^2(]0,T[;X)} = \int_0^T \langle v(t), w(t)\rangle_X \, dt. \tag{4.67}$$

Théorème 4.18 (Lions and Magenes [LM68]) *Soient* $\mathcal{V}$ *et* $\mathcal{H}$ *deux espaces de Hilbert tels que* $\mathcal{V} \subset \mathcal{H} \subset \mathcal{V}'$ *avec injections denses. Soit* $\mathcal{B}_t(v,w)$ *une famille de formes bilinéaires continues et coercives dans* $\mathcal{V}$*, uniformément en* t*. Soient un temps final* $T > 0$*, une donnée initiale* $u_0 \in \mathcal{H}$ *et un terme source* $f \in L^2\left(]0,T[;\mathcal{V}'\right)$*. Alors le problème consistant à trouver* $u \in L^2(]0,T[;\mathcal{V}) \cap \mathcal{C}([0,T];\mathcal{H})$ *avec* $\dfrac{\partial u}{\partial t} \in L^2\left(]0,T[;\mathcal{V}'\right)$ *telle que*

$$\begin{gathered} {}_{\mathcal{V}'}\left\langle \frac{\partial u(t)}{\partial t}, v \right\rangle_{\mathcal{V}} + \mathcal{B}_t(u(t),v) = {}_{\mathcal{V}'}(f(t),v)_{\mathcal{V}} \quad \forall\ v\ \in\ \mathcal{V},\ \textit{pour presque tout } t \in [0,T], \\ u(0) = u_0, \end{gathered} \tag{4.68}$$

possède une unique solution. De plus, il existe une constante $C > 0$ *telle que*

$$\|u\|_{L^\infty([0,T],L^2(\Omega))} + \|u\|_{L^2(]0,T[;\mathcal{V})} \le C\left(\|u_0\|_{\mathcal{H}} + \|f\|_{L^2(]0,T[;\mathcal{V}')}\right). \tag{4.69}$$

4.2.3.2 Problème régularisé

Nous supposons que $f \in L^2\left(0,T;L^2(\Omega)\right)$. Nous considérons le problème régularisé associé à l'équation (4.63) en ajoutant à l'opérateur $\mathcal{L}$ un opérateur de diffusion $-\varepsilon\Delta$:

$$\frac{\partial u_\varepsilon}{\partial t} + \mathcal{L}_\varepsilon u_\varepsilon = f, \quad \text{dans }]0,T] \times \Omega, \qquad u_\varepsilon(t=0) = u_0 \text{ dans } \Omega, \tag{4.70}$$

où

$$\mathcal{L}_\varepsilon u_\varepsilon = -\varepsilon\Delta u_\varepsilon - a : D^2 u_\varepsilon + b \cdot \nabla u_\varepsilon + c u_\varepsilon. \tag{4.71}$$

En intégrant contre une fonction test v assez régulière, nous obtenons formellement,

$$\begin{aligned} &\left\langle \frac{\partial u}{\partial t},\ v \right\rangle + \int_\Omega \left[(\varepsilon\,\mathcal{I} + a)\,\nabla u_\varepsilon - (b + \nabla\cdot a)\,u_\varepsilon\right]\cdot\nabla v \qquad (4.72) \\ &+ \int_\Omega \left[c - \nabla\cdot(b + \nabla\cdot a)\right] u_\varepsilon\ v - \int_{\partial\Omega} \left((\varepsilon\mathcal{I} + a)\,\nabla u_\varepsilon - (b + \nabla\cdot a)\,u_\varepsilon\right)\cdot n\ v \ = \ \int_\Omega f\ v. \end{aligned}$$

4.2.3.3 Conditions aux limites

En reprenant les notations du paragraphe 4.2.2.3, nous considérons les conditions aux limites suivantes :

– pour le problème initial (4.63)

$$\forall\ t\ \in (0,T],\ \begin{cases} u = h & \text{sur } \Sigma_D, \\ \left[a\nabla u - \dfrac{1}{2}(b + \nabla\cdot a)\,u\right]\cdot n = g & \text{sur } \Sigma_N, \\ u = \widetilde{h} & \text{sur } \Gamma_-\ ; \end{cases} \tag{4.73}$$

– pour le problème régularisé (4.70)

$$\forall\ t\ \in (0,T],\ \begin{cases} u_\varepsilon = h & \text{sur } \Sigma_D, \\ \left[(\varepsilon\ \mathcal{I} + a)\,\nabla u_\varepsilon - \dfrac{1}{2}\,(b + \nabla\cdot a)\,u_\varepsilon\right]\cdot n = g & \text{sur } \Sigma_N, \\ u_\varepsilon = \widetilde{h} & \text{sur } \Gamma_-, \\ [\varepsilon\ \nabla u_\varepsilon - \gamma\,(b + \nabla\cdot a)\,u_\varepsilon]\cdot n = 0 & \text{sur } \Gamma_+\ , \end{cases} \tag{4.74}$$

avec $\gamma \leq \dfrac{1}{2}$.

Remarque 4.3 *Pour simplifier nous supposons que les données aux bords h, g , $\widetilde{h}$ ne dépendent pas du temps.*

Nous supposons vérifiée l'hypothèse 4.2 d'existence d'une fonction de relèvement. La fonction $\widetilde{u}$ est définie par l'équation (4.34) alors

– $e = u - \widetilde{u}$ vérifie

$$\begin{cases} e = 0 & \text{sur } \Sigma_D \cup \Gamma_- \\ \left[a\nabla e - \dfrac{1}{2}\,(b + \nabla\cdot a)\,e\right]\cdot n = \widetilde{g} & \text{sur } \Sigma_N\ , \end{cases} \tag{4.75}$$

avec $\widetilde{g} = g - \left[a\nabla\widetilde{u} - \dfrac{1}{2}\,(b + \nabla\cdot a)\,\widetilde{u}\right]\cdot n$;

– $e_\varepsilon = u_\varepsilon - \widetilde{u}$ vérifie

$$\begin{cases} e_\varepsilon = 0 & \text{sur } \Sigma_D \cup \Gamma_-, \\ \left[(\varepsilon\ \mathcal{I} + a)\,\nabla e_\varepsilon - \dfrac{1}{2}\,(b + \nabla\cdot a)\,e_\varepsilon\right]\cdot n = \widetilde{g} - \varepsilon\ \dfrac{\partial\widetilde{u}}{\partial n} & \text{sur } \Sigma_N, \\ [\varepsilon\ \nabla e_\varepsilon - \gamma\,(b + \nabla\cdot a)\,e_\varepsilon]\cdot n = -\varepsilon\ \dfrac{\partial\widetilde{u}}{\partial n} & \text{sur } \Gamma_+\ . \end{cases} \tag{4.76}$$

4.2.3.4 Formulation variationnelle du problème aux limites régularisé

Le problème régularisé consiste à trouver la fonction e_ε appartenant à $L^2\,(0,T;\mathcal{W}_0)\cap \mathcal{C}\left([0,T];L^2\,(\Omega)\right)$ avec $\dfrac{\partial e_\varepsilon}{\partial t}$ dans $L^2\left(0,T;\mathcal{W}_0'\right)$ telle que, pour presque tout t dans $]0,T]$, pour tout $w\in\mathcal{W}_0$

$$\left\langle \frac{\partial e_\varepsilon}{\partial t}, w\right\rangle_{\mathcal{W}_0',\mathcal{W}_0} + \mathcal{B}_{\varepsilon,t}\,(e_\varepsilon, w) = \int_\Omega \widetilde{f}\ w + \int_{\Sigma_N}\widetilde{g}w - \int_{\Gamma_+\cup\Sigma_N}\varepsilon\frac{\partial\widetilde{u}}{\partial n}w. \tag{4.77}$$

où $\mathcal{B}_{\varepsilon,t}$ est la forme bilinéaire définie par (4.39) et $e_\varepsilon(t=0) = e_0$.

Hypothèse 4.4 *Il existe une constante $\underline{c} > -\infty$ telle que*

$$\forall\ x\in\overline{\Omega}, \forall\ t\in[0,T]\quad c(t,x) - \frac{1}{2}\nabla\cdot(b + \nabla\cdot a)\,(x) \geq \underline{c} > -\infty.$$

Lemme 4.19 *Sous l'hypothèse précédente, il existe η et C_ε telles que*

$$\forall w\in\mathcal{W}_0,\ \forall\ t\in[0,T]\quad \mathcal{B}_{\varepsilon,t}\,(w,w) + \eta\,(w,w)_{L^2(\Omega)} \geq C_\varepsilon\|w\|^2_{H^1(\Omega)}. \tag{4.78}$$

Remarque 4.4 *La constante η peut être choisie indépendamment de ε.*

Preuve Le résultat se déduit sans difficulté de la démonstration du lemme 4.12. ■

Remarque 4.5 *Sous l'hypothèse précédente, il existe η et C indépendantes de ε telles que*

$$\forall w \in \mathcal{W}_0,\ \forall\ t \in [0,T], \quad \mathcal{B}_{\varepsilon,t}(w,w) + \eta\,(w,w)_{L^2(\Omega)} \geq C\|w\|^2_{\mathcal{V}}, \tag{4.79}$$

où $\|\ \|_{\mathcal{V}}$ est définie dans (4.46).

Rappelons que l'inégalité de Gårding peut remplacer l'hypothèse de coercivité dans le théorème 4.18. En effet, le changement d'inconnue $u(t) = e^{\eta\ t}w(t)$ permet de considérer un problème sur lequel la forme bilinéaire est cette fois coercive. Dans le cas où la forme $\mathcal{B}$ vérifie seulement une inégalité de Gårding, l'*estimation d'énergie* (4.69) devient

$$\|u\|^2_{L^2(0,T;\mathcal{V})\cap\mathcal{C}[0,T];\mathcal{H}} \leq Ce^{2\eta T}\left(\|u_0\|^2_{\mathcal{H}} + \|f\|^2_{L^2(]0,T[;\mathcal{H})}\right). \tag{4.80}$$

Proposition 4.20 *Sous les hypothèses précédentes, le théorème 4.18 permet de montrer qu'il existe une unique fonction e_ε appartenant à $L^2(]0,T[;\mathcal{W}_0)\cap\mathcal{C}[0,T];\mathcal{W}_0$ avec $\dfrac{\partial e_\varepsilon}{\partial t}$ dans $L^2\left(]0,T[;\mathcal{W}_0'\right)$ vérifiant (4.77) et la condition initiale $e_\varepsilon(t=0) = e_0$. La fonction e_ε vérifie*

$$\frac{\partial e_\varepsilon}{\partial t} - \varepsilon\Delta e_\varepsilon - a : D^2 e_\varepsilon + b\cdot\nabla e_\varepsilon + ce_\varepsilon = \widetilde{f}, \tag{4.81}$$

au sens des distributions.

Remarque 4.6 *A l'aide des hypothèses suivantes : $e_0 \in H^1(\Omega)$, $\widetilde{f}$, $\widetilde{g} \in L^2(\Sigma_N)$, $c \in \mathcal{C}^1((0,T)\times\Omega)$ et $\widetilde{u} \in H^2(\Omega)$, nous pouvons montrer que $\dfrac{\partial e_\varepsilon}{\partial t}$ appartient à $L^2((0,T)\times\Omega)$ et que $(\varepsilon\mathcal{I} + a)\nabla e_\varepsilon$ appartient à $L^2(0,T;H_{div}(\Omega))$. De plus, les normes de ces fonctions dans les espaces correspondants sont bornées indépendamment de ε.*

4.2.3.5 Passage à la limite

Nous reprenons les définitions de $\mathcal{V}$ et $\mathcal{V}_0$ données par (4.45) et (4.46).

Théorème 4.21 *Soient un temps final $T > 0$, une donnée initiale $u_0 \in H^1(\Omega)$ et un terme source $f \in L^2((0,T)\times\Omega)$. Si les hypothèses 4.2 et 4.3 sont vérifiées, alors il existe une unique fonction $u \in L^2(0,T;\mathcal{V})\cap\mathcal{C}\left([0,T];L^2(\Omega)\right)$ telle que :*

$$\begin{array}{l}
\dfrac{\partial u}{\partial t} - a : D^2 u + b\cdot\nabla u + cu = f, \quad L^2((0,T)\times\Omega), \\
\left[a\nabla u - \dfrac{1}{2}(b + \nabla\cdot a)\,u\right]\cdot n = g \quad L^2\left(0,T;H^{-1/2}(\Sigma_N)\right), \\
u = h \quad L^2(0,T;\Sigma_D) \quad , \quad u = \widetilde{h}, \quad L^2\left(0,T;H^{-1/2}(\Gamma_-)\right), \\
et\ \ u(0,x) = u_0(x) \quad \left(L^2(\Omega)\right),
\end{array} \tag{4.82}$$

et qui soit obtenue comme la limite de la suite des solutions du problème régularisé (4.72,4.74).

Remarque 4.7 *Sur le bord Γ_+, il n'y a aucune condition à imposer. Sur le bord Γ_-, nous pouvons imposer indifféremment une condition de type Dirichlet ou une condition de Neumann de la forme de celle imposée sur Σ_N.*

Les étapes de la démonstration sont identiques à celle de la proposition 4.14. Nous indiquons les changements les plus significatifs.

Lemme 4.22 (Estimation a priori) *Si l'hypothèse 4.3 est vérifiée, alors il existe deux constantes $C > 0$ et $\overline{C} > 0$ telles que, pour tout t dans $[0,T]$,*

$$e^{-2\eta t}\|e_\varepsilon\|^2_{L^2(\Omega)} + \frac{1}{2}\varepsilon\int_0^t \|\nabla e_\varepsilon(\tau)\|^2_{L^2(\Omega)} e^{-2\eta\tau} C\int_0^t \|e_\varepsilon(\tau)\|^2_{\mathcal{V}} e^{-2\eta\tau} \tag{4.83}$$
$$+\left(\frac{1}{2}-\gamma\right)\int_0^t\int_{\Gamma_+} e^{-2\eta\tau}\,(b+\nabla\cdot a)\cdot n e_\varepsilon(\tau)^2 \le \overline{C}\left(1+\varepsilon^2\right).$$

Preuve Le résultat est obtenu en choisissant $e_\varepsilon\, e^{2\eta\tau}$ comme fonction test puis en intégrant en t et x. ■

Remarque 4.8 *Sous les hypothèses du théorème 4.21, nous pouvons aussi montrer que* $\left\|\dfrac{\partial e_\varepsilon}{\partial t}\right\|_{L^2((0,T)\times\Omega)}$ *est borné indépendamment de ε.*

Lemme 4.23 *La suite $\varepsilon\nabla e_\varepsilon$ tend vers 0 dans $L^2\left((0,T)\times\Omega\right)$.*

Lemme 4.24 *La suite e_ε est une suite de Cauchy dans $L^2(0,T;\mathcal{V}_0)$ et dans $\mathcal{C}\left([0,T];L^2(\Omega)\right)$. Elle admet donc une limite, notée e appartenant à $L^2(0,T;\mathcal{V}_0)\cap\mathcal{C}\left(]0,T];L^2(\Omega)\right)$, lorsque ε tend vers 0.*
Pour tout t dans $]0,T]$, la suite $\frac{\partial e_\varepsilon}{\partial t}$ est une suite bornée dans $L^2\left(]0,T[\times\Omega\right)$. Nous pouvons donc extraire une sous-suite telle que $\frac{\partial e_\varepsilon}{\partial t} \rightharpoonup \frac{\partial e}{\partial t}$ dans $L^2\left(]0,T[\times\Omega\right)$.

Preuve Soient e_ε et $e_{\varepsilon'}$ solutions de la *formulation faible* (4.77), alors en suivant la démonstration du lemme 4.17 avec $\nu_{\varepsilon,\varepsilon'} = e_\varepsilon - e_{\varepsilon'}$ et en choisissant $\nu_{\varepsilon,\varepsilon'}e^{-2\eta\tau}$ comme fonction test, nous obtenons

$$e^{-2\eta t}\left\|\nu_{\varepsilon,\varepsilon'}\right\|^2_{L^2(\Omega)} + \int_0^t e^{-2\eta\tau}\left\|\nu_{\varepsilon,\varepsilon'}\right\|^2_{\mathcal{V}} \tag{4.84}$$
$$\le\left|\int_0^t\int_{\Gamma_+\cup\Sigma_N}\left(\varepsilon-\varepsilon'\right)\frac{\partial\widetilde{u}}{\partial n}e^{-2\eta\tau}\nu_{\varepsilon,\varepsilon'} + \int_0^t\int_\Omega\left(\varepsilon\nabla e_\varepsilon - \varepsilon'\nabla e_{\varepsilon'}\right)e^{-2\eta\tau}\nabla\nu_{\varepsilon,\varepsilon'}\right|,$$

Le terme de droite de cette inégalité tend vers 0 par application du lemme 4.23, ce qui conduit au résultat désiré.

■

Preuve du théorème 4.21 Le résultat de la proposition 4.21 se déduit d'un résultat équivalent sur $e = u - \widetilde{u}$. Ce résultat est établi en passant « à la limite » dans le problème régularisé. Les lemmes précédents nous permettent de montrer les points suivants.

1. La convergence de $e_\varepsilon \to e$ dans $\mathcal{C}\left([0,T], L^2(\Omega)\right)$ implique $e(t=0) = e_0$. La convergence de $e_\varepsilon \to e$ dans $L^2(0,T;\mathcal{V}_0)$ implique $e = 0$ sur Σ_D.

2. La convergence $\varepsilon \nabla e_\varepsilon \to 0$ dans $L^2((0,T)\times\Omega)$ et la convergence faible de $\dfrac{\partial e_\varepsilon}{\partial t} \rightharpoonup \dfrac{\partial e}{\partial t}$ dans $L^2\left(0,T;L^2(\Omega)\right)$ permettent de passer à la limite dans la *formulation faible* (4.77). Pour presque tout t dans $]0,T]$ et pour tout w dans $\mathcal{V}_0$,

$$\int_\Omega \frac{\partial e}{\partial t}\, w + \int_\Omega \left[a\nabla e - (b + \nabla\cdot a)\, e\right]\cdot\nabla w + \int_\Omega \left[c - \nabla\cdot(b+\nabla\cdot a)\right] e\, w \qquad (4.85)$$
$$+\frac{1}{2}\int_{\Sigma_N} (b+\nabla\cdot a)\cdot n\; e\; w = \int_\Omega \widetilde{f}\; w + \int_{\Sigma_N} \widetilde{g} w.$$

   Ceci implique

$$\frac{\partial e}{\partial t} - a : D^2 e + b\cdot\nabla e + c\; e = \widetilde{f}, \qquad (4.86)$$

   au sens des distributions, et

$$\frac{\partial e}{\partial t} - a : D^2 e + b\cdot\nabla e + c\; e \in L^2((0,T)\times\Omega).$$

3. Montrons également que

$$a\nabla e_\varepsilon - (b+\nabla\cdot a)\, e_\varepsilon \rightharpoonup a\nabla e - (b+\nabla\cdot a)\, e \quad \text{dans } L^2(0,T;H_{\mathrm{div}}(\Omega)), \qquad (4.87)$$

   ce qui implique que

$$n\cdot\left[a\nabla e_\varepsilon - (b+\nabla\cdot a)\, e_\varepsilon\right] \rightharpoonup n\cdot\left[a\nabla e - (b+\nabla\cdot a)\, e\right] \quad \text{dans } L^2\left(0,T;H^{-1/2}(\partial\Omega)\right). \qquad (4.88)$$

   Par suite

$$n\cdot\left[a\nabla e - (b+\nabla\cdot a)\, e\right] = \widetilde{g}, \;\; \text{sur } \Sigma_N \quad \text{et } n\cdot\left[b+\nabla\cdot a\right] e = 0, \;\; \text{sur } \Gamma_-, \qquad (4.89)$$

   nous en déduisons Γ_- : $e = 0$.

■

Remarque 4.9 *Les points clés de la démonstration sont présentés dans cette remarque. Si la forme bilinéaire $\mathcal{B}_t$ associée au problème dégénéré vérifie une inégalité de Gårding sur $\mathcal{V}$ avec des constantes indépendantes de ε et si l'opérateur régularisé est continu de $\mathcal{V}_\varepsilon$ dans $\mathcal{V}_\varepsilon'$ où $\mathcal{V}_\varepsilon = \mathcal{V}\cap H^1(\Omega)$, alors le théorème 4.21 s'applique. La solution du problème dégénéré peut être définie par passage à la limite.*

Remarque 4.10 *Le fait que e_ε soit une suite de Cauchy sur $\mathcal{V}$ est une conséquence de l'inégalité de Gårding vérifiée par $\mathcal{B}_t$ sur $\mathcal{V}$. En effet, il suffit de remarquer que*

$$\frac{\partial e_\varepsilon - e_{\varepsilon'}}{\partial t} + \mathcal{L}\left(e_\varepsilon - e_{\varepsilon'}\right) = \frac{1}{2}\left(\varepsilon\nabla e_\varepsilon - \varepsilon'\nabla e_{\varepsilon'}\right). \qquad (4.90)$$

En multipliant par $(e_\varepsilon - e_{\varepsilon'})\, e^{-2\eta\tau}$, puis en intégrant par partie sur $[0,T]\times\Omega$, nous déduisons (4.84) de l'inégalité de Gårding sur $\mathcal{B}_t$.

Chapitre 5

Modèle à volatilité stochastique

Ce chapitre a pour objet l'étude d'un modèle à volatilité stochastique multi-facteurs. L'équation de valorisation, vérifiée par le prix d'une option européenne, est démontrée en exhibant un portefeuille de réplication de l'option. Les résultats d'existence et d'unicité de la solution de cette équation sont également établis à partir de la formulation faible.

5.1 Modèle de diffusion

5.1.1 Description du processus de diffusion

Considérons un sous-jacent dont la dynamique du prix est donnée par l'équation différentielle stochastique

$$dS_t = \mu S_t dt + \sigma_t S_t dW_t, \tag{5.1}$$

où $\mu\ S_t\ dt$ représente le terme de « drift », (W_t) un mouvement brownien et (σ_t) la volatilité. Comme nous l'avons vu précédemment, le modèle le plus simple consiste à considérer une volatilité constante. Toutefois, ce modèle est généralement trop grossier. En effet, les prix de marché observés sur les options vanilles ne peuvent pas être calculés. Ici, (σ_t) est supposée être un processus stochastique à valeurs non négatives satisfaisant une équation différentielle stochastique régie par un second mouvement brownien Z_t, non parfaitement corrélé à W_t. Plus précisément, nous supposons que $\sigma_t = f(Y_t)$, où f est une fonction à valeurs dans $\mathbb{R}^+$ et (Y_t) est le processus « directeur ». Le lecteur trouvera dans [FPS00],[AP05] une description des processus « directeurs » et les fonctions f les plus utilisées. Nous nous intéressons ici au cas d'un processus « directeur » (Y_t) à valeurs dans $\mathbb{R}^n$.

Définition 5.1 (Modèle de diffusion à n facteurs) *Le processus de volatilité (σ_t) est une fonction du processus n-dimensionnel $Y_t = \left(Y_t^1, \ldots, Y_t^n\right)^T$. Plus précisément, (Y_t) est un processus d'Ornstein-Uhlenbeck (OU) de dimension n. Chacun des processus Y_t^i, $1 \leq i \leq n$ vérifie l'équation différentielle stochastique :*

$$dY_t^i = \lambda_i(m_i - Y_t^i)dt + \beta_i dZ_t, \tag{5.2}$$

où λ_i et β_i sont des constantes positives.

Remarque 5.1 *Les n processus d'OU sont parfaitement corrélés entre eux. Les EDS de l'équation (5.2) sont régies par le même mouvement brownien. La matrice de la variance de ce processus est de la forme* $dt\left(\beta\beta^T\right)$, *où* $\beta = (\beta_1, \ldots, \beta_d)^T$. *Cette matrice est de rang* 1.

Une propriété essentielle des processus d'OU repose sur le retour à la moyenne (« mean-reversion ») qui correspond au terme de drift dans l'équation différentielle stochastique (5.2). Pour le processus uni-dimensionnel $\left(Y_t^i\right)$, la loi à l'instant t de Y_t^i sachant Y_0^i est donnée par

$$\mathcal{N}\left(m_i + \left(Y_0^i - m_i\right) e^{-\lambda_i t}, \frac{\beta_i^2}{2\lambda_i}(1 - e^{-2\lambda_i t})\right). \tag{5.3}$$

Remarque 5.2 *Nous supposons le plus souvent que les n processus OU sont de moyenne asymptotique nulle* i.e. $m_i = 0,\ 1 \leq i \leq n$.

Proposition 5.1 *Dans le cas n-dimensionnel, la loi devient*

$$Y_t = \mathcal{N}\left(\begin{pmatrix} m_1 + \left(Y_0^1 - m_1\right) e^{-\lambda_1 t} \\ \vdots \\ m_n + \left(Y_0^n - m_n\right) e^{-\lambda_n t} \end{pmatrix}, \Xi\right), \tag{5.4}$$

avec $\Xi_{i,j} = \dfrac{\beta_i \beta_j}{\lambda_i + \lambda_j}\left(1 - e^{-(\lambda_i + \lambda_j)t}\right)$.

Preuve Considérons le cas de la dimension 1 qui sera ensuite généralisé. La démonstration repose sur les changements de variables suivants : $X_t = Y_t - m$ puis $\widetilde{X}_t = X_t e^{\lambda_i t}$. La dynamique de ce processus est donnée d'après l'Eq(5.2) par

$$d\widetilde{X}_t = \beta_i e^{\lambda_i\ t} dZ_t, \quad \widetilde{X}_t = \mathcal{N}\left(\widetilde{X}_0, \frac{\beta_i^2}{2\lambda_i}\left(e^{2\lambda_i\ t} - 1\right)\right). \tag{5.5}$$

Le retour aux variables initiales donne

$$X_t = \mathcal{N}\left(\left(Y_0 - m_i\right) e^{e^{-\lambda_i\ t}}, \frac{\beta_i^2}{2\lambda_i}\left(1 - e^{-2\lambda_i\ t}\right)\right), \tag{5.6}$$

puis (5.3). Dans le cas multi-dimensionnel, les mêmes changements de variables permettent de montrer que $\left(\widetilde{X}_t\right)$ suit une loi normale de variance $\widetilde{\Xi}$:

$$d\widetilde{X}_t = \begin{pmatrix} \beta_1 e^{\lambda_1\ t} & 0 & \ldots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \beta_{n-1} e^{\lambda_{n-1}\ t} & 0 \\ 0 & \ldots & 0 & \beta_n e^{\lambda_n\ t} \end{pmatrix} dZ_t, \quad \Rightarrow \quad \widetilde{X}_t = \mathcal{N}\left(\widetilde{X}_0, \widetilde{\Xi}\right), \tag{5.7}$$

avec $\widetilde{\Xi}_{i,j} = \dfrac{\beta_i \beta_j}{\lambda_i + \lambda_j}\left(e^{(\lambda_i + \lambda_j)t} - 1\right)$. A nouveau, le retour aux variables initiales donne

$$\widetilde{X}_t = \mathcal{N}\left(X_0, \Xi\right). \tag{5.8}$$

Le passage $\widetilde{\Xi} \to \Xi$ s'effectue à l'aide du calcul de la probabilité de $X_t = x$ sachant $X_0 = x_0$. L'égalité en probabilité, *i.e.* sur les densités gaussiennes, implique le résultat sur les matrices de covariance :

$$(\widetilde{x} - \widetilde{x_0})^T \widetilde{\Xi}^{-1} (\widetilde{x} - \widetilde{x_0}) = (x - x_0)^T \Xi^{-1} (x - x_0) .$$

■

La proposition 5.1 permet de montrer que la loi du processus d'Ornstein-Uhlenbeck n-dimensionnel tend vers une loi limite mesure stationnaire quand $t \to +\infty$:

$$\mathcal{N}\left(\begin{pmatrix} m_1 \\ \vdots \\ m_d \end{pmatrix}, \overline{\Xi}\right), \quad \overline{\Xi}_{i,j} = \frac{\beta_i \beta_j}{\lambda_i + \lambda_j}.$$

Les constantes $\frac{1}{\lambda_i}$ sont les temps caractéristiques de retour à la moyenne, les paramètres λ_i sont nommés *taux de retour à la moyenne*. La matrice $\overline{\Xi}$ est la limite de la variance de Y_t quand $t \to +\infty$. Nous supposerons que les valeurs λ_i sont choisies de telle sorte que cette matrice soit inversible. Nous noterons Π son inverse. Sachant que $\beta_i > 0$ et $\lambda_i > 0$, l'inversibilité de $\overline{\Xi}$ équivaut à $\lambda_1 \neq \lambda_2$, dans le cas $n = 2$.

Remarquons que, si $\lambda_i = \lambda_j$, alors la matrice n'est pas inversible. En effet, $\overline{\Xi} = Diag(\beta)\, \Theta Diag\left(\frac{\beta}{\lambda}\right)$ où $0 < \Theta_{i,j} = \frac{\lambda_i}{\lambda_i + \lambda_j} < 1$. Si $\lambda_i = \lambda_j$, alors Θ a deux lignes égales. Elle n'est donc pas inversible.

Le Brownien Z_t est corrélé à W_t : il peut se décomposer sous la forme d'une combinaison linéaire de W_t et d'un second Brownien $\widetilde{W_t}$ indépendant de W_t :

$$Z_t = \rho W_t + \sqrt{1 - \rho^2}\widetilde{W_t}, \tag{5.9}$$

où le *facteur de corrélation* ρ appartient à $[-1, 1]$.

5.1.2 Équation de valorisation d'une option européenne

Considérons un contrat européen sur le sous-jacent mentionné ci-dessus qui expire à la date T, avec un payoff donné par la fonction $h(S_T)$. Le prix au temps $t < T$ dépend de t, du prix du sous-jacent S_t, et de Y_t. Notons $P(S_t, Y_t, t)$ le prix du contrat et $r(t)$ le taux d'intérêt.

L'option est évaluée en utilisant le principe de non arbitrage et la formule d'Itô multi-dimensionnelle. Le modèle ayant deux facteurs de risque, il n'est pas possible de construire un portefeuille répliquant l'option et contenant une option et une certaine quantité de l'actif. Le marché est dit *incomplet*. Nous construisons un portefeuille de réplication contenant deux options avec des maturités différentes et une certaine quantité du sous-jacent. Le portefeuille est composé d'une quantité a_t du sous-jacent S_t, d'une option de maturité T_1 dont le prix est

$$P_t^{(1)} = P^{(1)}(S_t, Y_t, t),$$

et de b_t options de maturité T_2 ($T_2 > T_1$) dont le prix est

$$P_t^{(2)} = P^{(2)}(S_t, Y_t, t).$$

La valeur du portefeuille est notée c_t. Le principe de non arbitrage implique que, pour tout $t < T_1$,

$$dc_t = a_t dS_t + dP_t^{(1)} + b_t dP_t^{(2)} = r_t c_t dt = r_t(a_t S_t + P_t^{(1)} + b_t P_t^{(2)})dt \tag{5.10}$$

Donnons à présent la dynamique des prix d'options $P_t^{(1)}$ et $P_t^{(2)}$ à l'aide de la formule d'Itô multi-dimensionnelle. D'après cette formule, $dP_t^{(1)}$ et $dP_t^{(2)}$ sont des combinaisons linéaires de dt, dW_t et $d\widetilde{W}_t$:

$$\begin{aligned} dP(S,Y,t) &= \frac{\partial P}{\partial t}dt + \frac{\partial P}{\partial s}dS + \sum_{i=1}^{n}\frac{\partial P}{\partial y_i}dY_i \\ &\quad + \left(\frac{1}{2}\sigma^2 s^2\frac{\partial^2 P}{\partial s^2} + \sum_{i=1}^{n}\sigma S_t\beta_i\rho\frac{\partial^2 P}{\partial s\partial y_i} + \frac{1}{2}\sum_{i,j}\beta_i\beta_j\frac{\partial^2 P}{\partial y_i\partial y_j}\right)dt. \end{aligned} \tag{5.11}$$

En introduisant les dynamiques Eq(5.1) et Eq(5.2), avec $m_i = 0, \quad 1 \le i \le n$,

$$\begin{aligned} dP(S,Y,t) &= \frac{\partial P}{\partial t}dt + \frac{\partial P}{\partial s}\left(\mu S_t dt + S_t\sigma dW_t\right) + \sum_{i=1}^{n}\frac{\partial P}{\partial y_i}\left(-\lambda_i Y_i dt + \beta_i d\widetilde{W}_t\right) \\ &+ \left(\frac{1}{2}\sigma^2 s^2\frac{\partial^2 P}{\partial s^2} + \sum_{i=1}^{n}\sigma S_t\beta_i\rho\frac{\partial^2 P}{\partial s\partial y_i} + \frac{1}{2}\sum_{i,j}\beta_i\beta_j\frac{\partial^2 P}{\partial y_i\partial y_j}\right)dt. \end{aligned} \tag{5.12}$$

Pour construire le portefeuille de réplication c_t neutre au risque, le terme de droite de l'équation (5.10) ne doit pas contenir un terme de risque.

– Le terme de risque $d\widetilde{W}_t$ s'annule si

$$b_t = -\frac{\sum_{i=1}^{n}\beta_i\frac{\partial P^{(2)}}{\partial y_i}}{\sum_{i=1}^{n}\beta_i\frac{\partial P^{(1)}}{\partial y_i}}.$$

– Le terme de risque dW_t s'annule si

$$a_t + \frac{\partial P^{(1)}}{\partial S} + b_t\frac{\partial P^{(2)}}{\partial S} = 0.$$

En remplaçant dans l'équation (5.10) les quantités a_t et b_t par leur valeur calculée ci-dessus, nous obtenons

$$\begin{aligned} &\frac{1}{\sum_{i=1}^{n}\beta_i\frac{\partial P^{(1)}}{\partial y_i}}\left(\begin{array}{ll} & \frac{\partial P^{(1)}}{\partial t} + \frac{1}{2}f(y)^2 s^2\frac{\partial^2 P^{(1)}}{\partial s^2} + \rho\sum_{i=1}^{n}\beta_i s f(y)\frac{\partial^2 P^{(1)}}{\partial s\partial y_i} \\ + & \frac{1}{2}\sum_{i,j=1}^{n}\beta_i\beta_j\frac{\partial^2 P^{(1)}}{\partial y_i y_j} - \sum_{i=1}^{n}\lambda_i y_i\frac{\partial P^{(1)}}{\partial y_i} + r(t)\left(s\frac{\partial P^{(1)}}{\partial s} - P^{(1)}\right) \end{array}\right) = \\ &\frac{1}{\sum_{i=1}^{n}\beta_i\frac{\partial P^{(2)}}{\partial y_i}}\left(\begin{array}{ll} & \frac{\partial P^{(2)}}{\partial t} + \frac{1}{2}f(y)^2 s^2\frac{\partial^2 P^{(2)}}{\partial s^2} + \rho\sum_{i=1}^{n}\beta_i s f(y)\frac{\partial^2 P^{(2)}}{\partial s\partial y_i} \\ + & \frac{1}{2}\sum_{i,j=2}^{n}\beta_i\beta_j\frac{\partial^2 P^{(2)}}{\partial y_i y_j} + \sum_{i=1}^{n}-\lambda_i y_i\frac{\partial P^{(1)}}{\partial y_i} + r(t)\left(s\frac{\partial P^{(2)}}{\partial s} - P^{(2)}\right) \end{array}\right). \end{aligned}$$

En remarquant que le terme de gauche ne dépend pas de T_2 et que le terme de droite ne dépend pas de T_1, nous en déduisons l'existence d'un facteur de prime de risque, représenté par la fonction $g(s,y,t)$, telle que

$$\begin{aligned}\frac{\partial P}{\partial t}+\frac{1}{2}f(y)^2s^2\frac{\partial^2 P}{\partial s^2}+\rho\sum_{i=1}^{n}\beta_i s f(y)\frac{\partial^2 P}{\partial s\partial y_i}+\frac{1}{2}\sum_{i,j=1}^{n}\beta_i\beta_j\frac{\partial^2 P}{\partial y_i y_j}\\ -\sum_{i=1}^{n}\lambda_i y_i\frac{\partial P^{(1)}}{\partial y_i}+r(t)\left(s\frac{\partial P}{\partial s}-P\right) &= g(s,y,t)\sum_{i=1}^{n}\beta_i\frac{\partial P}{\partial y_i}.\end{aligned}$$

Remarque 5.3 *Définir la prime de risque g est équivalent à définir la probabilité* risque neutre *pour $\widetilde{W}_t$. Nous établissons (annexe A) que, sous certaines hypothèses sur la dynamique des variances swap, cette probabilité* risque neutre *pour les produits sur la volatilité implique une prime de risque nulle : $g=0$. La proposition suivante se déduit de ce résultat.*

Proposition 5.2 *En tenant compte de la remarque précédente, le prix d'une option européenne, dont la dynamique du sous-jacent est donnée par l'équation (5.1) et la définition 5.1, vérifie l'équation aux dérivées partielles rétrograde*

$$\begin{aligned}\frac{\partial P}{\partial t}-r(t)P+\frac{1}{2}f(y)^2s^2\frac{\partial^2 P}{\partial s^2}+\rho\sum_{i=1}^{n}\beta_i s f(y)\frac{\partial^2 P}{\partial s\partial y_i}\\ +\frac{1}{2}\sum_{i,j=1}^{n}\beta_i\beta_j\frac{\partial^2 P}{\partial y_i\partial y_j}+r(t)s\frac{\partial P}{\partial s}-\sum_{i=1}^{n}\lambda_i y_i\frac{\partial P}{\partial y_i} &= 0, \quad (t,s,y)\in[0,T)\times\mathbb{R}^+\times\mathbb{R}^n,\\ P(s,y,T) &= h(s), \quad (s,y)\in\mathbb{R}^+\times\mathbb{R}^n,\end{aligned} \tag{5.13}$$

où T est la maturité de l'option et h la fonction payoff.

En notant $\mathcal{L}_t$ le générateur infinitésimal du processus de Markov $(S_t;Y_t)$, l'équation (5.13) devient :

$$\begin{aligned}\frac{\partial P}{\partial t}+\mathcal{L}_t P &= 0 \qquad (t,s,y)\in[0,T[\times\mathbb{R}^+\times\mathbb{R}^n,\\ P(s,y,T) &= h(s), \qquad (s,y)\in\mathbb{R}^+\times\mathbb{R}^n.\end{aligned} \tag{5.14}$$

En suivant ce qui est proposé dans [FPS00], l'opérateur $\mathcal{L}_t$ peut être décomposé en une somme de trois opérateurs différentiels de

$$\begin{aligned}\mathcal{L}_t &= \underbrace{\frac{1}{2}f(y)^2s^2\frac{\partial^2}{\partial s^2}+r(t)\left(\frac{\partial}{\partial s}-P\right)}_{\mathcal{L}^{BS}:\text{ Black\&Scholes }\sigma=\text{f(y)}}+\underbrace{\sum_{i=1}^{n}\rho\beta_i s f(y)\frac{\partial^2}{\partial s\partial y_i}}_{\mathcal{L}^{\rho}:\text{ correlation}}\\ &\quad+\underbrace{\frac{1}{2}\sum_{i,j=1}^{n}\beta_i\beta_j\frac{\partial^2}{\partial y_i\partial y_j}-\sum_{i=1}^{n}\lambda_i y\frac{\partial}{\partial y_i}}_{\mathcal{L}^{OU}:\text{ Ornstein}-\text{Uhlenbeck}}\end{aligned} \tag{5.15}$$

Finalement, si P est solution de l'équation (5.13), alors la formule d'Itô (5.11) permet de montrer que P est une martingale. En effet,

$$dP(S_t,Y_t,t)=(Sf(Y_t)\frac{\partial P}{\partial S}+\sum_{i=1}^{n}\beta_i\rho\frac{\partial P}{\partial y_i})dW_t+\sum_{i=1}^{n}\beta_i\sqrt{1-\rho^2}\frac{\partial P}{\partial y_i}d\widetilde{W}_t$$

Remarque 5.4 *Il est possible d'obtenir (5.13) en utilisant le théorème de Girsanov et la théorie du risque neutre, voir [FPS00] pour une application au cas d'un OU unidimensionnel.*

Quelques pistes pour l'obtention d'une solution semi-analytique de l'équation (5.13) Pascucci & al [MDF05] proposent une solution semi-analytique dans le cas d'un modèle à volatilité dont le processus de volatilité dépend de la moyenne du sous-jacent. Ce problème conduit à la résolution d'une EDP dégénérée de la forme de (5.13). Pascucci & al étudient une classe d'opérateurs hypoelliptiques (4.2.1) pour lesquels ils donnent une solution explicite à l'équation ultra-parabolique associée. Cette solution correspond à (5.13) avec f constante. Une méthode de développement asymptotique est ensuite proposée pour traiter le cas des coefficients variables à partir de la solution obtenue pour f constante.

Nous avons également envisagé l'approche proposée par Fouque & al [FPS00] qui consiste à effectuer un développement asymptotique de la solution par rapport à la variance $\nu^2 = \dfrac{\beta^2}{2\lambda}$ du processus d'Ornstein-Uhlenbeck . Ce développement conduit à la résolution d'EDP en variable d'espace pour chaque terme du développement (parabolique mais également elliptique). Cependant, les n facteurs de notre modèle correspondent à des comportements différents, le premier correspondant à un retour à la moyenne rapide et le dernier à un retour à la moyenne lent. Cette méthode n'a donc pas pu être mise en oeuvre. Notons cependant que les récents résultats proposés dans [SZ07] permettent un développement par rapport à $\frac{1}{\nu}$. En conclusion, il serait intéressant d'étudier une approche utilisant l'un ou l'autre des deux résultats en fonction du facteur considéré.

5.2 Analyse des problèmes de Cauchy

Ce paragraphe a pour objet l'étude de la *formulation variationnelle* de (5.13), afin d'obtenir les résultats d'existence et d'unicité de la solution ainsi qu'une estimation globale d'énergie. Cette estimation est importante pour l'étude d'approximation sur des espaces discrets, typiquement des méthodes de Galerkin. De plus, la *forme bilinéaire* associée à la *formulation variationnelle* nous permet de calculer les coefficients de la matrice de rigidité associée à la méthode de Galerkin.

La *formulation variationnelle* est également importante pour l'étude de problèmes d'inéquations variationnelles pour l'évaluation des options américaines.

Nous présentons, tout d'abord, les difficultés dans le cas du modèle de *Scott*. Jusqu'à présent, la forme de la fonction de volatilité f n'était pas précisée. Nous supposons dans ce qui suit que le processus de volatilité (σ_t) est lié au processus d'Ornstein-Uhlenbeck n-dimensionnel par la fonction

$$f : \mathbb{R}^n \to \mathbb{R}, \quad (x_1, \ldots, x_n)^T \to \exp\left(\frac{1}{2}\sum_{i=1}^{n} x_i\right). \tag{5.16}$$

Le cas $n = 1$ correspond au modèle de Scott.

5.2.1 Modèle de Scott

Dans un premier temps, nous nous limitons au cas $n = 1$. Nous supposons également que $\rho = 0$. Nous suivons le cheminement proposé dans [AT02] et [AFT05].

La variance de la loi limite du processus d'OU unidimensionnel, *i.e.* $\nu^2 = \frac{\beta^2}{2\lambda}$, joue un rôle important dans la suite. Dans le cas du modèle à 1 facteur, le problème (5.13) devient :

$$\begin{aligned} \frac{\partial P}{\partial t} + \frac{1}{2} f(y)^2 s^2 \frac{\partial^2 P}{\partial s^2} + \frac{\beta^2}{2} \frac{\partial^2 P}{\partial y^2} + r(t) \left(s \frac{\partial P}{\partial s} - P \right) - \lambda y \frac{\partial P}{\partial y} &= 0, \quad \text{sur } [0,T[\times \mathbb{R}^+ \times \mathbb{R}, \\ P(s,y,T) &= h(s), \quad \text{sur } \mathbb{R}^+ \times \mathbb{R}. \end{aligned} \tag{5.17}$$

Afin de se ramener au cadre bien connu d'un problème parabolique avec condition initiale, le sens du temps est inversé. En considérant le changement de variable $t \to T - t$, avec la notation $r(t) \to r(T-t)$, (5.17) devient

$$\begin{aligned} \frac{\partial P}{\partial t} - \frac{1}{2} f(y)^2 s^2 \frac{\partial^2 P}{\partial s^2} - \frac{\beta^2}{2} \frac{\partial^2 P}{\partial y^2} - r(t) \left(s \frac{\partial P}{\partial s} - P \right) + \lambda y \frac{\partial P}{\partial y} &= 0, \quad \text{sur } [0,T[\times \mathbb{R}^+ \times \mathbb{R}, \\ P(s,y,0) &= h(s), \quad \text{sur } \mathbb{R}^+ \times \mathbb{R}. \end{aligned} \tag{5.18}$$

En multipliant (5.18) par une fonction test $v \in \mathcal{D}(\Omega)$ (v ne dépend pas de t) et en intégrant par partie, nous obtenons (formellement)

$$\begin{aligned} \frac{\partial}{\partial t} \left(\int_\Omega P v \right) &+ a(P,v) = 0 \\ \text{avec} \quad a(w,v) &= \frac{1}{2} \int_\Omega s^2 f(y)^2 \frac{\partial w}{\partial s} \frac{\partial v}{\partial s} + \int_\Omega f(y)^2 s \frac{\partial w}{\partial s} v - \int_\Omega r s \frac{\partial w}{\partial s} v \\ &\quad + \int_\Omega \frac{\beta^2}{2} \frac{\partial w}{\partial y} \frac{\partial v}{\partial y} + \int_\Omega \lambda y \frac{\partial w}{\partial y} v + \int_\Omega r w \, v. \end{aligned} \tag{5.19}$$

Nous n'avons pas pu démontrer que le problème (5.18) admettait une *formulation faible* sur un espace de Sobolev $\mathcal{V}$. En particulier, il ne nous a pas été possible de trouver un espace $\mathcal{V}$ sur lequel la forme bilinéaire a définie par (5.19) vérifie une inégalité de Gårding. La difficulté est la suivante : suivons la trame permettant habituellement d'aboutir à l'inégalité de Gårding. Posons $w = v$: le terme $\displaystyle\int_\Omega f(y)^2 s \frac{\partial w}{\partial s} v$ devient, après une intégration par partie, $\displaystyle -\int_\Omega f(y)^2 v^2$. L'inégalité de Gårding est vérifiée si ce terme est « compensé » par un autre. Ceci est possible pour certaines fonctions f après un changement d'inconnue (voir [AT02]). Le modèle de Scott ne vérifie pas les hypothèses requises sur f.

Nous souhaitons à présent montrer que le problème associé à la dérivée de P par rapport à s (la fonction delta) est bien posé sur un espace de Sobolev $\mathcal{V}$. D'autre part, $P\big|_{s=0}(y,t) = h(0) e^{-\int_t^T r(\tau)\, d\tau}$.

La fonction δ vérifie

$$\begin{aligned} \frac{\partial \delta}{\partial t} - \frac{1}{2} f(y)^2 \frac{\partial}{\partial s} \left(s^2 \frac{\partial \delta}{\partial s} \right) - \frac{\beta^2}{2} \frac{\partial^2 \delta}{\partial y^2} - r \, s \frac{\partial \delta}{\partial s} + \lambda y \frac{\partial \delta}{\partial y} &= 0, \quad \text{sur } [0,T[\times \mathbb{R}^+ \times \mathbb{R}, \\ \delta(s,y,0) &= h'(s), \quad \text{sur } \mathbb{R}^+ \times \mathbb{R}. \end{aligned} \tag{5.20}$$

5.2.1.1 Le problème de Cauchy (5.20)

Considérons le changement d'inconnue suivant, celui-ci implique que la nouvelle inconnue u tend vers 0 lorsque $|y|$ tend vers ∞,

$$u(s,y,t) = \delta(s,y,t)e^{-(1-\eta)\frac{y^2}{2\nu^2}}, \tag{5.21}$$

où η est un paramètre tel que $0 < \eta < 1$.

Remarquons que la fonction u définie par (5.21) satisfait l'équation parabolique dégénérée

$$\begin{aligned}\frac{\partial u}{\partial t} - \frac{1}{2}f(y)^2\frac{\partial}{\partial s}\left(s^2\frac{\partial u}{\partial s}\right) - \frac{\beta^2}{2}\frac{\partial^2 u}{\partial y^2} - r\ s\frac{\partial u}{\partial s} \\ - (1-2\eta)\,\lambda y\frac{\partial u}{\partial y} - \left((1-\eta)+\eta(\eta-1)\frac{y^2}{\nu^2}\right)\lambda u = \quad 0, \quad \text{sur } (0,T)\times\mathbb{R}^+\times\mathbb{R},\end{aligned} \tag{5.22}$$

avec la condition initiale $u(s,y,0) = h'(s)e^{-(1-\eta)\frac{y^2}{2\nu^2}}, \quad (s,y) \in \mathbb{R}^+\times\mathbb{R}$.

En effet, si $Q = e^{(1-\eta)\frac{y^2}{2\nu^2}}$, alors $\delta = uQ$ et

$$\begin{aligned}\frac{\partial \delta}{\partial y} &= \left(\frac{\partial u}{\partial y} + \frac{(1-\eta)}{\nu^2}yu\right)Q \\ \frac{\partial^2 \delta}{\partial y^2} &= \left(\frac{\partial^2 u}{\partial y^2} + \frac{2(1-\eta)}{\nu^2}y\frac{\partial u}{\partial y} + \left(\frac{(1-\eta)}{\nu^2} + \left(\frac{(1-\eta)}{\nu^2}\right)^2 y^2\right)u\right)Q.\end{aligned} \tag{5.23}$$

De plus,

$$\begin{aligned}&\left(-\lambda y\frac{\partial \delta}{\partial y} + \frac{\beta^2}{2}\frac{\partial^2 \delta}{\partial y^2}\right)Q^{-1} \\ &= \frac{\beta^2}{2}\frac{\partial^2 u}{\partial y^2} + \left(\frac{2(1-\eta)\beta^2}{2\nu^2} - \lambda\right)y\frac{\partial u}{\partial y} + \left(\frac{(1-\eta)\beta^2}{2\nu^2} + \left(\frac{(1-\eta)^2\beta^2}{2\nu^4} - \frac{(1-\eta)}{\nu^2}\lambda\right)y^2\right)u \\ &= \frac{\beta^2}{2}\frac{\partial^2 u}{\partial y^2} + (2(1-\eta)-1)\,\lambda y\frac{\partial u}{\partial y} + \left((1-\eta) + \left((1-\eta)^2 - (1-\eta)\right)\frac{y^2}{\nu^2}\right)\lambda u \\ &= \frac{\beta^2}{2}\frac{\partial^2 u}{\partial y^2} + (1-2\eta)\,\lambda y\frac{\partial u}{\partial y} + \left((1-\eta)+\eta(\eta-1)\frac{y^2}{\nu^2}\right)\lambda u.\end{aligned} \tag{5.24}$$

En introduisant ces relations dans (5.20) puis en divisant par Q, nous obtenons (5.22).

Soient $\Omega = \mathbb{R}^+\times\mathbb{R}$ et $\mathcal{L}$ l'opérateur défini par :

$$\mathcal{L}u = -\frac{1}{2}f(y)^2\frac{\partial}{\partial s}\left(s^2\frac{\partial u}{\partial s}\right) - \frac{\beta^2}{2}\frac{\partial^2 u}{\partial y^2} - rs\frac{\partial u}{\partial s} - (1-2\eta)\,\lambda y\frac{\partial u}{\partial y} - \left(1-\eta\frac{y^2}{\nu^2}\right)(1-\eta)\lambda u. \tag{5.25}$$

En multipliant (5.25) par une fonction test v puis en intégrant par partie, nous définissons une forme bilinéaire qui aura un sens sur un espace de Sobolev à poids $\mathcal{V}$. Cet espace $\mathcal{V}$, défini par

$$\mathcal{V} = \left\{v :\ \left(\sqrt{1+y^2+f(y)^2}v, \frac{\partial v}{\partial y}, sf(y)\frac{\partial v}{\partial s}\right) \in \left(L^2(\Omega)\right)^3\right\}, \tag{5.26}$$

muni de la norme

$$|||v|||_{\mathcal{V}} = \left(\int_\Omega \left(1 + y^2 + f(y)^2\right) v^2 + \left(\frac{\partial v}{\partial y}\right)^2 + s^2 f(y)^2 \left(\frac{\partial v}{\partial s}\right)^2 \right)^{\frac{1}{2}}. \tag{5.27}$$

est un espace de Hilbert. Il vérifie les propriétés suivantes :

1. Notons $\mathcal{D}(\Omega)$ l'espace des fonctions régulières à support compact dans Ω. Alors $\mathcal{D}(\Omega)$ est contenu dans $\mathcal{V}$ et $\mathcal{D}(\Omega)$ est dense dans $\mathcal{V}$.
2. $\mathcal{V}$ est séparable.
3. $\mathcal{V}$ est dense dans $L^2(\Omega)$.

Nous souhaitons appliquer le théorème 4.18 d'existence et d'unicité de la solution au sens faible d'une EDP parabolique. A cette fin, le point 1 est crucial. La méthode de Friedrichs (Théorème 4.2 dans [Fri44]) nous permet de le démontrer :

Montrons que les fonctions de $\mathcal{D}(\Omega)$ sont denses dans $\mathcal{V}$. Ceci nous conduit à introduire une fonction à support compact $\phi_r : \Omega \to \mathbb{R}$ telle que $\phi_r(z) = \phi\left(\frac{z}{r}\right)$ avec

$$\begin{aligned} &\phi_r \in \mathcal{C}^\infty(\Omega), \quad \phi_r(z) = 1 \text{ si } z \in B_r, \quad \phi_r(z) = 0 \text{ si } z \notin B_{2r}, \\ &|\nabla \phi_r(z)| \leq \frac{C}{r} \text{ si } |z| \in [r, 2r]. \end{aligned} \tag{5.28}$$

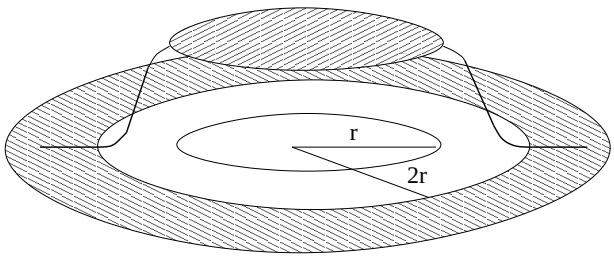

FIG. 5.1 – Fonction ϕ_r

Alors, pour toute fonction v appartenant à $\mathcal{V}$, la fonction $\phi_r\ v$ tend vers v dans $\mathcal{V}$ quand r tend vers ∞. En effet,

$$\begin{aligned} &\int_\Omega \left(sf(y)\frac{\partial}{\partial s}\left[\phi_r\ v\right] - sf(y)\frac{\partial}{\partial s}\left[v\right] \right)^2 \\ &= \int_\Omega \left(s\frac{\partial \phi_r}{\partial s}\ f(y)v + (\phi_r(z) - 1)\ \ f(y)s\frac{\partial v}{\partial s} \right)^2 \\ &\leq 2\int_\Omega \left(s\frac{\partial \phi_r}{\partial s} \right)^2\ f(y)^2 v^2 + 2\int_\Omega (\phi_r(z) - 1)^2\, f(y)^2 s^2 \left(\frac{\partial v}{\partial s}\right)^2 \\ &\leq 2\int_{r\leq|z|\leq 2r} \left(s\frac{\partial \phi_r}{\partial s} \right)^2 f(y)^2 v^2 + 2\int_{|z|>r} f(y)^2 s^2 \left(\frac{\partial v}{\partial s}\right)^2 \\ &\leq 2C\int_{r\leq|z|\leq 2r} f(y)^2 v^2 + 2\int_{|z|>r} f(y)^2 s^2 \left(\frac{\partial v}{\partial s}\right)^2. \end{aligned}$$

Le théorème de Lebesgue permet de montrer que cette dernière quantité tend vers 0 quand r tend vers l'infini. Nous convolons ensuite $\phi_r v$ par une fonction régularisante à support compact pour aboutir au résultat désiré.

Lemme 5.3 *Pour toute fonction v dans $\mathcal{V}$,*

$$\int_{\mathbb{R}} f(y)^2 v^2 \leq 4 \int_{\mathbb{R}} f(y)^2 \left(s \frac{\partial v}{\partial s} \right)^2 .$$

La semi-norme

$$\|v\|_{\mathcal{V}} = \left(\int_{\Omega} \left(1+y^2\right) v^2 + \left(\frac{\partial v}{\partial y} \right)^2 + s^2 f(y)^2 \left(\frac{\partial v}{\partial s} \right)^2 \right)^{\frac{1}{2}}, \tag{5.29}$$

est une norme sur $\mathcal{V}$ équivalente à $|||v|||_{\mathcal{V}}$.

Preuve Voir [AT02]. ■

Soit a la forme bilinéaire de $\mathcal{V} \times \mathcal{V} \to \mathbb{R}$ telle que pour tout $w, v \in \mathcal{V}$,

$$\begin{aligned} a(w,v) &= \frac{1}{2} \int_{\Omega} s f(y) \frac{\partial w}{\partial s} s f(y) \frac{\partial v}{\partial s} - \int_{\Omega} r s \frac{\partial w}{\partial s} v + \int_{\Omega} \frac{\beta^2}{2} \frac{\partial w}{\partial y} \frac{\partial v}{\partial y} \\ &\quad - \int_{\Omega} (1-2\eta) \lambda y \frac{\partial w}{\partial y} v + \int_{\Omega} \left(-\lambda \left((1-\eta) + \eta(\eta-1) \frac{y^2}{\nu^2} \right) \right) w \, v. \end{aligned} \tag{5.30}$$

La théorie de Lions and Magenes [LM68] sur les solutions faibles des problèmes paraboliques nécessite un résultat de continuité de l'opérateur $\mathcal{L}$ de $\mathcal{V}$ dans $\mathcal{V}'$ son dual et une inégalité de Gårding sur la forme bilinéaire a. Nous n'avons pas pu démontrer la continuité du terme $\displaystyle\int_{\Omega} r s \frac{\partial w}{\partial s} v$ dans l'espace $\mathcal{V}$ précédemment défini. Ceci est dû à la dégénérescence de l'opérateur $\mathcal{L}$ mise en évidence dans la remarque suivante :

Remarque 5.5 *En passant en variable logarithmique $x = \log s$,* la partie principale du symbole de Fourier *de $\mathcal{L}^x$ est donnée par :*

$$\mathcal{A}^0_{\mathcal{L}^x}(x, y, \omega_1, \omega_2) = f(y)\omega_1^2 + \beta^2 \omega_2^2. \tag{5.31}$$

En conclusion, $\mathcal{L}^x$ n'est pas uniformément elliptique sur $\mathbb{R}$, puisqu'il n'existe pas de constante $c > 0$ telle que

$$\left| \mathcal{A}^0_{\mathcal{L}}(x, y, \omega_1, \omega_2) \right| \geq c \, |\boldsymbol{\omega}|^2, \qquad \textit{pour } (x,y) \in \mathbb{R}^2, \, \boldsymbol{\omega} \in \mathbb{R}^2.$$

Démontrons que nous nous trouvons dans le cadre d'application du théorème 4.21, en tenant compte de la remarque 4.9. Ce résultat se déduit des deux points suivants :

1. Le premier est énoncé sous la forme d'une proposition :

   Proposition 5.4 *Soit a la forme bilinéaire définie par (5.30). Si $0 < \eta < 1$, alors il existe deux constantes positives C et c (C dépend de η) telles que, pour tout $v \in \mathcal{V}$,*

   $$a(v,v) + c\|v\|^2_{L^2(\Omega)} \geq C \|v\|^2_{\mathcal{V}}. \tag{5.32}$$

2. L'opérateur $\mathcal{L}_\epsilon : \quad v \rightarrow \mathcal{L}v + \epsilon s^2 \dfrac{\partial^2 v}{\partial x^2}$ est continu de $\mathcal{V}_\epsilon$ dans $\mathcal{V}'_\epsilon$ où $\mathcal{V}_\epsilon = \mathcal{V} \cap \left\{ s\dfrac{\partial v}{\partial s} \in L^2(\Omega) \right\}$. De plus, la forme bilinéaire associée à $\mathcal{L}_\epsilon$ définie sur $\mathcal{V}_\epsilon \times \mathcal{V}_\epsilon$ vérifie une inégalité de Gårding.

Preuve de la proposition 5.4 Le résultat est basé sur les intégrations par partie suivantes :

$$-\int_\Omega y\frac{\partial v}{\partial y}v = \frac{1}{2}\int_\Omega v^2, \quad -\int_\Omega rs\frac{\partial v}{\partial s}v = \frac{1}{2}\int_\Omega rv^2$$

Nous en déduisons

$$a(v,v) + \int_\Omega \frac{\lambda + r}{2}v^2 = \int_\Omega \frac{s^2 f(y)^2}{2}\left(\frac{\partial v}{\partial s}\right)^2 + \int_\Omega \frac{\beta^2}{2}\left(\frac{\partial v}{\partial y}\right)^2 + \int_\Omega \lambda\eta(1-\eta)\frac{y^2}{\nu^2}v^2. \quad (5.33)$$

De plus,

$$\begin{aligned} a_\epsilon(v,v) + \int_\Omega \frac{\lambda + r}{2}v^2 = &\int_\Omega \frac{s^2(f(y)^2 + \epsilon)}{2}\left(\frac{\partial v}{\partial s}\right)^2 + \int_\Omega \frac{\beta^2}{2}\left(\frac{\partial v}{\partial y}\right)^2 \\ &+ \int_\Omega \left(\lambda\eta(1-\eta)\frac{y^2}{\nu^2} + \frac{\epsilon}{2}\right)v^2. \end{aligned} \quad (5.34)$$

■

Remarque 5.6 *Si $\eta = 1$, la propriété n'est pas vérifiée, ceci justifie le changement d'inconnue (5.21).*

La continuité de l'opérateur $\mathcal{L}_\epsilon$ de $\mathcal{V}_\epsilon$ dans $\mathcal{V}'_\epsilon$ se déduit de

$$\left|\int_\Omega rs\frac{\partial v}{\partial s}w\right| \leq \left(\int_\Omega \epsilon s^2\left(\frac{\partial v}{\partial s}\right)^2\right)^{\frac{1}{2}}\left(\int_\Omega \epsilon\left(\frac{r}{\epsilon}\right)^2 w^2\right)^{\frac{1}{2}}.$$

En reprenant la démarche générale présentée dans le chapitre 4, le théorème suivant peut être démontré :

Théorème 5.5 *Pour toute fonction $u_0 \in H^1(\Omega)$, il existe une unique fonction u obtenue comme limite des solutions du problème régularisé, dans $L^2(0,T;\mathcal{V}) \cap H^1(]0,T[;L^2(\Omega))$, avec $\frac{\partial u}{\partial t} \in L^2(]0,T[\times\Omega)$ telle que, pour toute fonction régulière $\phi \in \mathcal{D}(0,T)$, et pour toute fonction $v \in \mathcal{V}$,*

$$-\int_0^T \phi'(t)\left(\int_\Omega u(t)v\right)dt + \int_0^T \phi(t)\widetilde{a}(u,v)dt = 0 \quad (5.35)$$

et

$$u(s,y,0) = h'(s)e^{-(1-\eta)\ \frac{y^2}{2\nu^2}}. \quad (5.36)$$

Remarque 5.7 *Dans le cas de l'évaluation d'une option européenne call ou put, la condition initiale $h'(s)$ n'appartient pas à $H^1(\Omega)$. Cependant, la singularité est locale et ne porte que sur une variable. Nous pensons qu'il est possible de généraliser le théorème 5.5 à ce cas.*

Remarque 5.8 *Il est également possible d'étudier la* formulation faible *de l'opérateur obtenu après le changement de variable* $\widetilde{s} = s\, e^{\int_0^t r(v)dv}$. *En notant* $\widetilde{u}\,(t,\widetilde{s},y) = u\,(t,s,y)$ *et* $\widetilde{\mathcal{L}}$ *l'opérateur défini par :*

$$\widetilde{\mathcal{L}}\widetilde{u} = -\frac{1}{2}f(y)^2\frac{\partial}{\partial s}\left(s^2\frac{\partial\widetilde{u}}{\partial s}\right) - \frac{\beta^2}{2}\frac{\partial^2\widetilde{u}}{\partial y^2} - (1-2\eta)\,\lambda y\frac{\partial\widetilde{u}}{\partial y} - \lambda\left((1-\eta)+\eta(\eta-1)\frac{y^2}{\nu^2}\right)\widetilde{u}, \quad (5.37)$$

il est possible d'aboutir à un résultat équivalent au théorème 5.5 pour $T < \infty$.

5.2.1.2 Principe du maximum

Le *principe du maximum faible* est vérifié par la fonction δ. Dans le cas d'un put vanille, le principe du maximum sur la fonction δ et la propriété « $P\,(t, s=0, y) = Ke^{-\int_t^T r(\tau)d\tau}$ ne dépend pas de y » implique l'encadrement sur le prix du put :

$$\left(S - Ke^{-\int_t^T r(\tau)d\tau}\right)^- \leq P(t,S,y) \leq Ke^{-\int_t^T r(\tau)d\tau}, \quad (5.38)$$

ainsi que la *relation de parité Call-Put*

$$C\,(t,S,y) - P\,(t,S,y) = S - Ke^{-\int_t^T r(\tau)d\tau}.$$

5.2.1.3 Le problème de Cauchy pour $P_{|s=0}$ dans le cas où la fonction payoff dépend de y

Considérons à présent le problème au bord $s = 0$ et notons $P_0\,(t,y) = P\,(t,0,y)$. L'équation (5.18) devient

$$\frac{\partial P_0}{\partial t} - \frac{\beta^2}{2}\frac{\partial^2 P_0}{\partial y^2} - \lambda y\frac{\partial P_0}{\partial y} + rP_0 = 0. \quad (5.39)$$

La restriction notée z_0 de la fonction z, définie par $z(s,y,t) = P(s,y,t)e^{-(1-\eta)\frac{y^2}{2\nu^2}}$, à $s = 0$ ($z_0\,(y,t) = z\,(0,y,t)$) vérifie

$$\frac{\partial z_0}{\partial t} - \frac{\beta^2}{2}\frac{\partial^2 z_0}{\partial y^2} - \lambda y\frac{\partial z_0}{\partial y} - (1-2\eta)\,\lambda y\frac{\partial z_0}{\partial y} - \left((1-\eta)+\eta(\eta-1)\frac{y^2}{\nu^2}\right)\lambda z_0 + r\; z_0 = 0, \quad (5.40)$$

sur $(0,T)\times\mathbb{R}$.

Proposition 5.6 *La forme bilinéaire* a_0 *définie sur* $\mathcal{V}_{s_0}$ *par*

$$a_0\,(w,v) = \frac{\beta^2}{2}\int_\Omega \frac{\partial w}{\partial y}\frac{\partial v}{\partial y} - \int_\Omega (1-2\eta)\,\lambda y\frac{\partial w}{\partial y}v + \int_\Omega\left(-(1-\eta)-\eta(\eta-1)\frac{y^2}{\nu^2}+r\right)\lambda w\; v = 0,$$

est continue et vérifie une inégalité de Gårding où

$$\mathcal{V}_{s_0} = \left\{v :\ \left|\sqrt{1+y^2}v, \frac{\partial v}{\partial y} \in \left(L^2\,(\mathbb{R})\right)^2\right\}.$$

Preuve La démonstration est similaire à celle obtenue pour a. Nous obtenons

$$a_0\,(v,v) + \int_\Omega \frac{\lambda}{2}v^2 = \int_\Omega \frac{\beta^2}{2}\left(\frac{\partial v}{\partial y}\right)^2 + \int_\Omega\left(\lambda\eta(1-\eta)\frac{y^2}{\nu^2}+r\right)v^2. \quad (5.41)$$

■

5.2.2 Modèle à n facteurs

Nous détaillons, dans ce paragraphe, le cas multi-facteurs. La difficulté réside ici dans le fait que la matrice de diffusion est dégénérée. L'opérateur n'est donc plus parabolique. Nous verrons que l'espace de Sobolev à poids sur lequel est établie la *formulation variationnelle* est caractérisé par m conditions sur les fonctions qui le composent, avec $m < d+1$.

Nous abordons le même cheminement que précédemment, en traitant l'équation donnant le $\delta = \dfrac{\partial P}{\partial s}$. A nouveau, le prix au bord $s = 0$, vaut $P_{\big|s=0}(t,y) = h(0)e^{-\int_t^T r(\tau)d\tau}$. Nous supposerons $\rho = 0$.

Le changement de variable $t \to T-t$ dans l'équation (5.13) et la dérivation par rapport à s donnent le problème de Cauchy

$$\frac{\partial \delta}{\partial t} - \frac{1}{2}f(y)^2 s^2 \frac{\partial^2 \delta}{\partial s^2} - \frac{1}{2}\sum_{i,j=1}^{n} \beta_i \beta_j \frac{\partial^2 \delta}{\partial y_i \partial y_j} - rs\frac{\partial \delta}{\partial s} + \sum_{i=1}^{n} \lambda_i y_i \frac{\partial \delta}{\partial y_i} = 0, \tag{5.42}$$

avec condition initiale : $\delta(s,y,0) = h'(s)$.

5.2.2.1 Le problème de Cauchy (5.42) :

Soient Π une matrice symétrique définie positive et u la nouvelle inconnue

$$u(s,y,t) = \delta(s,y,t)e^{-(1-\eta)\frac{y^T \Pi y}{2}}, \tag{5.43}$$

où η est un paramètre tel que $0 < \eta < 1$.

Les vecteurs $Y = (y_1, \ldots, y_n)^T$, $\beta = (\beta_1, \ldots, \beta_n)^T, 1 = (1, \ldots, 1)^T$, les matrices diagonales Λ et B telles que $(\Lambda)_{i,i} = \lambda_i$, $(B)_{i,i} = \beta_i$, $1 \le i \le n$ et le produit scalaire dans $\mathbb{R}^n$, noté $x^T y$, nous permettront de caractériser les coefficients de l'EDP. La norme $\|X\|$ associée au produit scalaire et la norme matricielle $\|A\| = \sup_{\|X\|=1} \|AX\|$ seront utilisées pour énoncer certaines hypothèses sur ces coefficients.

Proposition 5.7 *La fonction u définie par (5.43) satisfait l'équation parabolique dégénérée*

$$\begin{aligned} &\frac{\partial u}{\partial t} - \frac{1}{2}f(y)^2 \frac{\partial}{\partial s}\left(s^2 \frac{\partial u}{\partial s}\right) - rs\frac{\partial u}{\partial s} - \frac{1}{2}\nabla_y \cdot \left(\beta\beta^T \nabla_y u\right) + (\Lambda Y)^T (\nabla_y u) \\ &- (1-\eta)\left(\beta^T \Pi Y\right)\beta^T \nabla_y u - \frac{1-\eta}{2}\left[\beta^T \Pi \beta + (1-\eta)\left(\beta^T \Pi Y\right)^2 - 2(\Lambda Y)^T (\Pi Y)\right] u \\ &\qquad = 0, \quad \textit{sur } \mathbb{R}^+ \times \mathbb{R}^n \times (0,T), \\ &\qquad u(s,y,0)e^{(1-\eta)\frac{y^T \Pi y}{2}} = h'(s), \quad \textit{sur } \mathbb{R}^+ \times \mathbb{R}^n. \end{aligned} \tag{5.44}$$

Preuve Notons $Q = e^{(1-\eta)\frac{y^T \Pi y}{2}}$ alors $\delta = uQ$ avec $\nabla_y Q = (1-\eta)\Pi Y\, Q$ et

$$\begin{aligned} \sum_{i=1}^{n} \lambda_i y_i \frac{\partial P}{\partial y_i} &= (\Lambda Y)^T (\nabla_y [u\, Q]) = (\Lambda Y)^T (\nabla_y u\, Q + \nabla_y Q\, u) \\ &= (\Lambda Y)^T (\nabla_y u + (1-\eta)\Pi Y)\, Q, \end{aligned} \tag{5.45}$$

$$\begin{aligned}\frac{1}{2}\nabla_y\cdot\left[\beta\beta^T\nabla_y\left(u\ Q\right)\right] &= \frac{1}{2}\nabla_y\cdot\left[\beta\beta^T\nabla_y u\ Q+\left(1-\eta\right)\beta\beta^T\Pi Y\ u\ Q\right] \\ &= \frac{Q}{2}\left(\nabla_y\cdot\beta\beta^T\nabla_y u+\left(1-\eta\right)\nabla_y\cdot\left[\beta\beta^T\Pi Y\right]u+\left(1-\eta\right)\left(\nabla_y u\right)^T\left(\beta\beta^T\Pi Y\right)\right) \\ &\qquad +\frac{1}{2}\left(\nabla_y Q\right)^T\left(\beta\beta^T\nabla_y u+\left(1-\eta\right)\beta\beta^T\Pi Y\ u\right) \\ &= \frac{Q}{2}\nabla_y\cdot\beta\beta^T\nabla_y u+\left(1-\eta\right)\frac{Q}{2} \\ &\qquad \left\{tr\left(\beta\beta^T\Pi\right)u+2\left(\Pi Y\right)^T\left(\beta\beta^T\nabla_y u\right)+\left(1-\eta\right)\left(\Pi Y\right)^T\left(\beta\beta^T\Pi Y\right)u\right\} \\ &= \frac{Q}{2}\nabla_y\cdot\beta\beta^T\nabla_y u+\frac{Q}{2}\left(1-\eta\right) \\ &\qquad \left\{2\left(\beta^T\Pi Y\right)\beta^T\nabla_y u+\left(\left(1-\eta\right)\left(\beta^T\Pi Y\right)^2+tr\left(\beta\beta^T\Pi\right)\right)u\right\}.\end{aligned} \tag{5.46}$$

En remarquant que $tr\left(\beta\beta^T\Pi\right)=\beta^T\Pi\beta$, l'équation (5.44) est obtenue en introduisant ces relations dans (5.42) puis en divisant par Q. ■

En multipliant (5.44) par une fonction test v et en intégrant par partie, nous obtenons la forme bilinéaire a associée à $\mathcal{L}$ et définie sur l'espace de Sobolev $\mathcal{V}$. Caractérisons cet espace : soient $g:\mathbb{R}^n\to\mathbb{R},\quad y\to\left\|B^{-1}\Lambda Y\right\|$, alors l'espace $\mathcal{V}$

$$\mathcal{V}=\left\{v:\ \left(\sqrt{1+g(y)^2+f(y)^2}v,\left|\beta^T\nabla_y v\right|,sf(y)\frac{\partial v}{\partial s}\right)\in\left(L^2(\Omega)\right)^3\right\}, \tag{5.47}$$

muni de la norme

$$|||v|||_{\mathcal{V}}=\left(\int_\Omega\left(1+g(y)^2+f(y)^2\right)v^2+\left(\beta^T\nabla_y v\right)^2+s^2f(y)^2\left(\frac{\partial v}{\partial s}\right)^2\right)^{\frac{1}{2}}. \tag{5.48}$$

est un espace de Hilbert. Il vérifie les propriétés suivantes :

1. $\mathcal{D}(\Omega)\subset\mathcal{V}$ et $\mathcal{D}(\Omega)$ est dense dans $\mathcal{V}$.
2. $\mathcal{V}$ est séparable.
3. $\mathcal{V}$ est dense dans $L^2\left(\Omega\right)$.

Lemme 5.8 *La semi-norme*

$$\|v\|_{\mathcal{V}}=\left(\int_\Omega\left(1+g(y)^2\right)v^2+\left(\beta^T\nabla_y v\right)^2+s^2f(y)^2\left(\frac{\partial v}{\partial s}\right)^2\right)^{\frac{1}{2}}. \tag{5.49}$$

est équivalente à $|||v|||_{\mathcal{V}}$.

Soit a de $\mathcal{V}\times\mathcal{V}\to\mathbb{R}$ telle que, pour tout w,v appartenant à $\mathcal{V}$, $a\left(w,v\right)=\langle\mathcal{L}w,v\rangle$, alors

$$\begin{aligned}a\left(w,v\right) &= \frac{1}{2}\int_\Omega s^2f(y)^2\frac{\partial w}{\partial s}\frac{\partial v}{\partial s}+\frac{1}{2}\int_\Omega\left(\beta^T\nabla_y w\right)\left(\beta^T\nabla_y v\right) \\ &\quad +\int_\Omega\left(\Lambda Y-\left(1-\eta\right)\left(\beta^T\Pi Y\right)\beta\right)^T\left(\nabla_y w\right)\ v-\int_\Omega rs\frac{\partial w}{\partial s}v \\ &\quad -\frac{1-\eta}{2}\int_\Omega\left(\left(\beta^T\Pi\beta\right)+\left(1-\eta\right)\left(\beta^T\Pi Y\right)^2-2\left(\Lambda Y\right)^T\left(\Pi Y\right)\right)w\ v.\end{aligned} \tag{5.50}$$

Démontrons que nous nous trouvons dans le cadre d'application du théorème 4.21, en tenant compte de la remarque 4.9. Les points suivants vont être démontrés :

1. Nous énonçons le premier sous la forme d'une proposition :

   Proposition 5.9 *Si* Π *est la matrice diagonale définie par* $\Pi_{ii} = \dfrac{\lambda_i}{\beta_i^2}$ *alors, pour tout* $\eta < 1$*, la forme bilinéaire* a*, définie sur* $\mathcal{V} \times \mathcal{V}$ *vérifie une inégalité de Gårding : il existe deux constantes* C *et* c *positives telles que, pour tout* $v \in \mathcal{V}$*,*

   $$a(v,v) + c\|v\|^2_{L^2(\Omega)} \geq C\|v\|^2_{\mathcal{V}}. \tag{5.51}$$

2. L'opérateur $\mathcal{L}_\epsilon : \; v \to \mathcal{L}v + \dfrac{\epsilon}{2}\left(s^2\dfrac{\partial^2 v}{\partial s^2} + \sum_{i=1}^{n}\dfrac{\partial^2 v}{\partial y_i^2}\right)$ est continu de $\mathcal{V}_\epsilon$ dans $\mathcal{V}'_\epsilon$ où $\mathcal{V}_\epsilon = \mathcal{V} \cap \left\{ s\dfrac{\partial v}{\partial s} \in L^2(\Omega),\ \dfrac{\partial v}{\partial y_i} \in L^2(\Omega),\ \forall 1 \leq i \leq n \right\}$. De plus, la forme bilinéaire associée à $\mathcal{L}_\epsilon$ définie sur $\mathcal{V}_\epsilon \times \mathcal{V}_\epsilon$ vérifie une inégalité de Gårding.

Preuve de la proposition 5.9 Le résultat est basé sur les intégrations par partie suivantes :

$$\int_\Omega (\Lambda Y)^T (\nabla_y v)\, v = \int_\Omega \frac{1}{2}(\Lambda Y)^T \left(\nabla_y v^2\right) = -\int_\Omega \nabla_y \cdot [\Lambda Y]\, v^2 = -\frac{1}{2}\int_\Omega tr(\Lambda)\, v^2,$$

$$2\int_\Omega \beta^T \Pi Y\ \beta^T \nabla_y v\ v = \int_\Omega \left(\beta\beta^T \Pi Y\right)^T \left(\nabla_y v^2\right) = -\int_\Omega \nabla_y \cdot \left[\beta\beta^T \Pi Y\right] v^2 = -\int_\Omega tr\left(\beta\beta^T \Pi\right) v^2.$$

Nous en déduisons

$$\begin{aligned} a(v,v) + \int_\Omega (tr(\Lambda) + r)\frac{v^2}{2} &= \frac{1}{2}\int_\Omega s^2 f(y)^2 \left(\frac{\partial v}{\partial s}\right)^2 + \frac{1}{2}\int_\Omega \left(\beta^T \nabla_y v\right)^2 \\ &\quad - \int_\Omega \left(-2(1-\eta)(\Lambda Y)^T (\Pi Y) + (1-\eta)^2 \left(\beta^T \Pi Y\right)^2\right)\frac{v^2}{2}. \end{aligned} \tag{5.52}$$

Nous donnons à présent une condition suffisante sur Π et η pour que la forme bilinéaire a vérifie une inégalité de Gårding.

L'inégalité de Cauchy-Schwarz, appliquée à $\beta^T \Pi Y = (1)^T (B\Pi Y)$ permet d'obtenir une minoration du second terme du membre de droite de (5.52) à l'aide d'une somme de carrés :

$$\begin{aligned} &(1-\eta)^2 \left(\beta^T \Pi Y\right)^2 - 2(1-\eta)(\Lambda Y)^T (\Pi Y) \\ &\quad \underbrace{\leq}_{CS} (1-\eta)^2 \|B\Pi Y\|^2 - 2(1-\eta)\left(B^{-1}\Lambda Y\right)^T (B\Pi Y) \\ &\quad = \left\|(1-\eta)(B\Pi Y) - B^{-1}\Lambda Y\right\|^2 - \left\|B^{-1}\Lambda Y\right\|^2. \end{aligned} \tag{5.53}$$

Si $\Pi = B^{-1}\Lambda B^{-1}$, Π est la matrice diagonale telle que $\Pi_{ii} = \dfrac{\lambda_i}{\beta_i^2}$, alors l'inégalité de Gårding est vérifiée pour tout $\eta < 1$:

$$a(v,v) - \int_\Omega (tr(\Lambda) + r)\frac{v^2}{2} = \frac{1}{2}\int_\Omega s^2 f(y)^2 \left(\frac{\partial v}{\partial s}\right)^2 \left(\beta^T \nabla_y v\right)^2 + (1-\eta)\left\|B^{-1}\Lambda Y\right\|^2 v^2. \tag{5.54}$$

De plus,

$$a_\epsilon(v,v) = a(v,v) + \int_\Omega \frac{\epsilon}{2} s^2 \left(\frac{\partial v}{\partial s}\right)^2 + \frac{\epsilon}{2}\sum_{i=1}^{n}\int_\Omega \frac{\partial v}{\partial x_i}^2 + \int_\Omega \frac{\epsilon}{2} v^2. \tag{5.55}$$

■

Remarque 5.9 *La matrice* Π *choisie correspond à l'inverse de la loi limite dans le cas de processus d'Ornstein-Uhlenbeck non corrélés.*

Montrons à présent que l'opérateur $\mathcal{L}_\epsilon$ est continu de $\mathcal{V}_\epsilon$ dans $\mathcal{V}'_\epsilon$.

Preuve La continuité du terme de convection pose problème si nous considérons le problème initial. Sur le problème régularisé, nous avons

$$\begin{aligned}\left|\int_\Omega (\Lambda Y)^T (\nabla_y w)\, v\right|^2 &\le \int_\Omega \frac{1}{\epsilon^2}\left\|B^{-1}\Lambda Y\right\|^2 v^2 \int_\Omega \|\epsilon B \nabla_y w\|^2 \\ &\le \max_{1\le i\le n}\left(\frac{\beta_i}{\epsilon}\right)^2 \int_\Omega \left\|B^{-1}\Lambda Y\right\|^2 v^2 \int_\Omega \|\epsilon \nabla_y w\|^2.\end{aligned} \tag{5.56}$$

Le résultat sur le terme d'ordre 0 est une conséquence de la majoration suivante

$$\left(\beta^T \Pi Y\right)^2 \le \|B\Pi Y\|^2 \le \left\|B\Pi B\Lambda^{-1}\right\| \left\|B^{-1}\Lambda Y\right\|^2. \tag{5.57}$$

La symétrie de $B\Pi B\Lambda^{-1}$ permet de borner cette norme par la plus grande valeur propre (en valeur absolue).

$$\begin{aligned}\left|\int_\Omega \left((1-\eta)^2\left(\beta^T\Pi Y\right)^2 - 2(1-\eta)(\Lambda Y)^T(\Pi Y)\right) w\, v\right| &\\ \le C\left(\left\|B\Pi B\Lambda^{-1}\right\|\right)\int_\Omega \left\|B^{-1}\Lambda Y\right\|^2 w\, v &\\ \le C\|w\|_{\mathcal{V}}\|v\|_{\mathcal{V}},&\end{aligned} \tag{5.58}$$

■

En reprenant la démarche générale présentée dans le chapitre 4, le théorème suivant peut être démontré :

Théorème 5.10 *Pour toute fonction* $u_0 \in H^1(\Omega)$*, il existe une unique fonction* u *obtenue comme limite des solutions du problème régularisé, dans* $L^2(0,T;\mathcal{V}) \cap H^1(]0,T[;L^2(\Omega))$*, avec* $\frac{\partial u}{\partial t} \in L^2(]0,T[\times\Omega)$ *telle que, pour toute fonction régulière* $\phi \in \mathcal{D}(0,T)$ *et pour toute fonction* $v \in \mathcal{V}$*,*

$$-\int_0^T \phi'(t)\left(\int_\Omega u(t)v\right)dt + \int_0^T \phi(t)a(u,v)dt = 0 \tag{5.59}$$

et

$$u(s,y,0) = h'(s)e^{-(1-\eta)\frac{Y^T\Pi Y}{2}}. \tag{5.60}$$

5.2.2.2 Le problème de Cauchy pour $P_{|s=0}$ dans le cas où la fonction payoff dépend de y

Considérons à présent le problème au bord $s = 0$ et notons $P_0(t,y) = P(t,0,y_1,\ldots,y_n)$. L'équation vérifiée par la fonction prix devient

$$\frac{\partial P_0}{\partial t} - \frac{1}{2}\sum_{i,j=1}^{n}\beta_i\beta_j\frac{\partial^2 P_0}{\partial y_i \partial y_j} - \sum_{i=1}^{n}\lambda_i y_i\frac{\partial P_0}{\partial y_i} + rP_0 = 0. \tag{5.61}$$

La restriction notée z_0 de la fonction z, définie par $z(s,y,t) = P(s,y,t)e^{-(1-\eta)\frac{Y^T\Pi Y}{2}}$, à $s=0$ ($z_0(y,t) = z(0,y,t)$) vérifie

$$\begin{aligned}\frac{\partial u}{\partial t} - \frac{1}{2}\nabla_y\cdot\left(\beta\beta^T\nabla_y z_0\right) + (\Lambda Y)^T(\nabla_y z_0) - (1-\eta)\left(\beta^T\Pi Y\right)\beta^T\nabla_y z_0 \\ -\frac{1-\eta}{2}\left[\beta^T\Pi\beta + (1-\eta)\left(\beta^T\Pi Y\right)^2 - 2(\Lambda Y)^T(\Pi Y)\right]z_0 + rz_0 \\ = 0, \quad \text{sur } \mathbb{R}^+\times\mathbb{R}\times(0,T), \\ z_0(y,0)e^{(1-\eta)\frac{y^T\Pi y}{2}} = h(0), \quad \text{sur } \mathbb{R}^+\times\mathbb{R}.\end{aligned} \tag{5.62}$$

Proposition 5.11 *La forme bilinéaire a_0 définie par*

$$\begin{aligned}a_0(w,v) &= \frac{1}{2}\int_\Omega \beta^T\nabla_y w\beta^T\nabla_y v + \int_\Omega\left(\Lambda Y - (1-\eta)\left(\beta^T\Pi Y\right)\beta\right)^T(\nabla_y w)\ v \\ &-\frac{1-\eta}{2}\int_\Omega\left(\left(\beta^T\Pi\beta\right) + (1-\eta)\left(\beta^T\Pi Y\right)^2 - 2(\Lambda Y)^T(\Pi Y)\right)w\ v + r\ w\ v,\end{aligned}$$

vérifie une inégalité de Gårding sur $\mathcal{V}_{s_0}$ et est continue de $\mathcal{V}_{s_0}$ dans $\mathcal{V}'_{s_0}$, où

$$\mathcal{V}_{s_0} = \left\{v :\ \left|\sqrt{1+g(y)^2}v, \left|\beta^T\nabla_y v\right| \in \left(L^2(\mathbb{R})\right)^2\right\}.$$

Preuve La démonstration est similaire à celle de la proposition 5.9.

$$a_0(v,v) + \frac{1}{2}\int_\Omega tr(\Lambda)v^2 \geq \frac{1}{2}\int_\Omega\left(\beta^T\nabla_y v\right)^2 + \frac{1}{2}c_\eta\int_\Omega\left\|B^{-1}\Lambda Y\right\|^2 v^2 + \int_\Omega rv^2. \tag{5.63}$$

■

Chapitre 6

Approximation numérique de l'équation de valorisation

Deux approches pour la résolution numérique de l'équation de valorisation (5.13) sont présentées dans ce chapitre.

Dans la première partie, la méthode employée tient compte du caractère dégénéré de la diffusion. Un changement de variables permet d'écrire l'équation de valorisation (5.13) sous la forme d'une équation ultra-parabolique et hypoelliptique. La résolution par une méthode de différences finies sparse s'avère être instable : ces instabilités sont liées aux conditions aux bords.

Dans la seconde partie, la méthode numérique est appliquée aux variables de (5.13). Une nouvelle transformation permettra de résoudre un problème de Cauchy avec des conditions aux bords de Dirichlet homogènes. Dans ce cas, les résultats de convergence et les temps de calcul sont conformes à ce qui est attendu pour une méthode de *Sparse Grid*.

6.1 Formulation avec une équation ultra-parabolique

6.1.1 Équation ultra-parabolique

La formulation initiale de l'équation (5.13) ne tient pas compte de la corrélation parfaite entre les $n-$processus d'Ornstein-Uhlenbeck. Celle-ci entraîne une dégénérescence de l'opérateur de diffusion. Deux changements de variables permettant d'aboutir à une équation ultra-parabolique sont présentés. Le premier part du système d'équations stochastiques sur les Y_t^i. Celui-ci donne un sens « physique » à la démarche adoptée. Le second changement de variables intervient, quant à lui, dans l'équation aux dérivées partielles. Le choix des nouvelles variables sera justifié dans la partie suivante.

6.1.1.1 Analyse des processus de diffusion

Supposons que les processus « directeurs » sur la volatilité vérifient le système d'EDS (5.2).

Proposition 6.1 *Soient* $\overline{\beta} = \sum_{i=1}^{n} \beta_i$, *et* X_t^i, $i = 1, \ldots, n$ *les processus définis par*

$$X_t^1 = \frac{\overline{\beta}}{\beta_1} Y_t^1 \quad \text{et} \quad X_t^i = e^{-\lambda_i\, t}\left(X_0^i - X_0^1\right) - (\lambda_i - \lambda_1)\int_0^t e^{\lambda_i (u-t)} X_u^1 du, \tag{6.1}$$

alors

$$Y_t^i = \frac{\beta_i}{\overline{\beta}}\left(X_t^1 + X_t^i\right), \tag{6.2}$$

si cette relation est vérifiée en $t = 0$.

Ce résultat se déduit du lemme suivant

Lemme 6.2 *Si* Y_t^1 *est choisi comme référence, pour tout* $1 \leq i \leq n$,

$$Y_t^i = \frac{\beta_i}{\beta_1}\left[Y_t^1 + e^{-\lambda_i t}\left\{\left(\frac{\beta_1}{\beta_i} Y_0^i - Y_0^1\right) - (\lambda_i - \lambda_1)\int_0^t e^{\lambda_i u} Y_u^1 du\right\}\right]. \tag{6.3}$$

Preuve Soit $\widetilde{Y}_t^i = e^{\lambda_i t} Y_t^i$, alors $d\widetilde{Y}_t^i = \beta_i e^{\lambda_i t} dZ_t = \frac{\beta_i}{\beta_1} e^{(\lambda_i - \lambda_1)t} d\widetilde{Y}_t^1$. Intégrons cette relation :

$$\begin{aligned}
\widetilde{Y}_t^i - \widetilde{Y}_0^i &= \frac{\beta_i}{\beta_1}\int_0^t e^{(\lambda_i-\lambda_1)u} d\widetilde{Y}_u^1 = \left[\frac{\beta_i}{\beta_1} e^{(\lambda_i-\lambda_1)u}\widetilde{Y}_u^1\right]_0^t - \frac{\beta_i}{\beta_1}(\lambda_i - \lambda_1)\int_0^t e^{(\lambda_i-\lambda_1)u}\widetilde{Y}_u^1 du, \\
&= \frac{\beta_i}{\beta_1} e^{(\lambda_i-\lambda_1)t}\widetilde{Y}_t^1 - \frac{\beta_i}{\beta_1}\widetilde{Y}_0^1 - \frac{\beta_i}{\beta_1}(\lambda_i - \lambda_1)\int_0^t e^{(\lambda_i-\lambda_1)u}\widetilde{Y}_u^1 du, \qquad (6.4) \\
& \qquad (6.5)
\end{aligned}$$

Le changement de variables inverse permet de conclure :

$$\beta_1 Y_t^i - \beta_i Y_t^1 = e^{-\lambda_i t}\left(\beta_1 Y_0^i - \beta_i Y_0^1\right) - \beta_i(\lambda_i - \lambda_1)\int_0^t e^{\lambda_i (u-t)} Y_u^1 du. \tag{6.6}$$

■

Introduisons X_t^i donné par (6.1) dans la relation (6.3),

$$Y_i^t = \frac{\beta_i}{\overline{\beta}}\left(X_t^1 + e^{-\lambda_i\, t}\left(\frac{\overline{\beta}}{\beta_i} Y_0^i - X_0^1\right) - (\lambda_i - \lambda_1)\int_0^t e^{\lambda_i (u-t)} X_u^1 du\right) = \frac{\beta_i}{\overline{\beta}}\left(X_t^1 + X_t^i\right), \tag{6.7}$$

où

$$X_t^i = e^{-\lambda_i\, t} X_0^i - (\lambda_i - \lambda_1)\int_0^t e^{\lambda_i u} X_u^1 du \tag{6.8}$$

Les dynamiques des processus X_t^i sont donc :

$$\begin{cases} dX_t^i = -\lambda_i X_t^i dt + \overline{\beta} dW_t & \text{si } i = 1 \\ dX_t^i = \left(-\lambda_i X_t^i - (\lambda_i - \lambda_1) X_t^1\right) dt & \text{sinon .} \end{cases} \tag{6.9}$$

L'équation de valorisation d'une option européenne est obtenue en appliquant le théorème de Feynman-Kac :

$$\begin{aligned}\frac{\partial u}{\partial t} - ru + r\ s\ \frac{\partial u}{\partial s} + \frac{1}{2}f(x)^2 s^2 \frac{\partial^2 u}{\partial s^2} + \rho f(x)\overline{\beta}\frac{\partial^2 u}{\partial s \partial x_1} \\ + \frac{1}{2}\overline{\beta}\frac{\partial^2 u}{\partial x_1^2} - \sum_{i=1}^{n}\left(\left(\lambda_i - \lambda_1\right)x_1 + \lambda_i x_i\right)\frac{\partial u}{\partial x_i} = 0,\end{aligned} \tag{6.10}$$

avec la condition de Cauchy $u\left(T, s, \boldsymbol{x}\right) = h(s)$. La fonction volatilité définie par (5.16) devient d'après (6.7)

$$f\left(y_1, \ldots, y_n\right) = \exp\left(\frac{1}{2}\sum_{i=1}^{n} y_i\right) \rightarrow f\left(x_1, \ldots, x_n\right) = \exp\left(\frac{1}{2}x_1\right)\exp\left(\frac{1}{2\overline{\beta}}\sum_{i=2}^{n}\beta_i x_i\right). \tag{6.11}$$

6.1.1.2 Changement de variables dans l'équation (5.13)

Proposition 6.3 *Considérons le changement de variables suivant*

$$x_1 = \sum_{i=1}^{n} y_i, \quad et \quad x_i = \frac{\beta_i}{\beta_1}y_1 - y_i\ \forall i > 1. \tag{6.12}$$

alors l'équation d'évaluation du prix d'une option européenne (5.13) devient

$$\frac{\partial u}{\partial t} - \frac{1}{2}f(x_1)^2 s^2 \frac{\partial u}{\partial s} - rs\frac{\partial u}{\partial s} + ru - \frac{1}{2}\overline{\beta}^2\frac{\partial u}{\partial x_1} - \overline{\beta}\rho f(x_1)s\frac{\partial^2 u}{\partial s \partial x_1} + \sum_{i,j=1}^{n} L_{i,j}x_j\frac{\partial u}{\partial x_i} = 0, \tag{6.13}$$

avec la condition de Cauchy $u\left(0, s, x\right) = h(s)$. *La matrice* L *est donnée par* $L = Q^T D_\lambda Q^{-T}$, *où*

$$Q = \begin{pmatrix} 1 & \frac{\beta_2}{\beta_1} & \frac{\beta_3}{\beta_1} & \cdots & \frac{\beta_n}{\beta_1} \\ 1 & -1 & 0 & \cdots & 0 \\ 1 & 0 & -1 & \vdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 1 & 0 & \cdots & 0 & -1 \end{pmatrix} \text{ et } D_\lambda = -Diag\left(\lambda_1, \lambda_2, \ldots, \lambda_n\right), \tag{6.14}$$

La démonstration de la proposition 6.3 suit les étapes suivantes : remarquons que

$$y = Q^{-T}x, \quad \nabla_y = Q\nabla_x. \tag{6.15}$$

Lemme 6.4 (Calcul du déterminant)

$$\left|Q^{(d)}\right| = \left(\sum_{i=1}^{n}\beta_i\right)\ \beta_1^{-1} = \frac{\overline{\beta}}{\beta_1} \tag{6.16}$$

Preuve Le développement du déterminant par rapport à la dernière colonne permet d'obtenir la formule de récurrence :

$$\left|Q^{(d)}\right| = \left|Q^{(d-1)}\right| + (-1)^{n-1}\frac{\beta_n}{\beta_1}\begin{vmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 1 & 0 & \cdots & 0 \end{vmatrix} = \left|Q^{(d-1)}\right| + \left((-1)^{n-1}\right)^2\frac{\beta_n}{\beta_1}. \tag{6.17}$$

La somme de 1 à d de ces identités permet de conclure. ■

Lemme 6.5 (Terme d'ordre 2) *Le changement de variables donne*

$$\sum_{i=1}^{d}\beta_i\beta_j\frac{\partial^2 u}{\partial y_i\partial y_j} \to |Q|^{-1}\,\overline{\beta}^2\frac{\partial^2 u}{\partial x_1^2}. \tag{6.18}$$

Preuve Soit $v_1 = (\beta_1, \beta_2, \ldots, \beta_d)^T$, alors :

$$\int_{\mathbb{R}^n}(\nabla_y\varphi)^T v_1\cdot v_1^T\nabla_y\psi = |Q|^{-1}\int_{\mathbb{R}^n}(\nabla_x\varphi)^T Q^T v_1\cdot v_1^T Q\nabla_x\psi. \tag{6.19}$$

Or $\left(Q^T v_1\right)^T = \left(\overline{\beta}, 0\cdots, 0\right)$,

$$\int_{\mathbb{R}^n}(\nabla_y\varphi)^T v_1\cdot v_1^T\nabla_y\psi = |Q|^{-1}\int_{\mathbb{R}^n}\overline{\beta}^2\frac{\partial\varphi}{\partial x_1}\frac{\partial\psi}{\partial x_1}. \tag{6.20}$$

■

Lemme 6.6 (Terme d'ordre 1) *Le changement de variables donne*

$$\sum_{i=1}^{d}\lambda_i y_i\frac{\partial u}{\partial y_i} \to |Q|^{-1}\,(Lx)^T\,\nabla_x u. \tag{6.21}$$

Preuve Il suffit de remarquer que

$$\int_{\mathbb{R}^n}(D_\lambda y)^T\nabla_y\varphi\cdot\psi dy = |Q|^{-1}\int_{\mathbb{R}^n}\left(D_\lambda Q^{-T}x\right)^T Q\nabla_x\varphi\;\psi dx = |Q|^{-1}\int_{\mathbb{R}^n}(Lx)^T\,\nabla_x\varphi\;\psi dx. \tag{6.22}$$

■

Lemme 6.7 (Terme croisé avec x) *Le changement de variables donne*

$$\sum_{i=1}^{d}\rho f(y)s\beta_i\frac{\partial^2 u}{\partial s\partial y_i} \to |Q|^{-1}\,\overline{\beta}\rho f(x)s\frac{\partial^2 u}{\partial s\partial x_1}. \tag{6.23}$$

Preuve

$$\int_{\mathbb{R}^n} \frac{\partial}{\partial s}\left(f(y)v_1^T\nabla_y\varphi\right)\ \psi dy = |Q|^{-1}\int_{\mathbb{R}^n}\frac{\partial}{\partial s}\left(f(x)v_1^TQ\nabla_x\varphi\right)\psi dx$$
$$= |Q|^{-1}\int_{\mathbb{R}^n}\frac{\partial}{\partial s}\left(f(x)\left(\beta^\star,0,\cdots,0\right)\nabla_x\varphi\right)\ \psi dx.$$

■

L'expression de L dans un modèle à 3 facteurs est

$$\begin{aligned} L\ =&Q^TD_\lambda Q^{-T}\\ =&\frac{\beta_1}{\overline{\beta}}\begin{pmatrix}\sum_{i=1}^{3}\lambda_i\beta_i & \beta_1(\lambda_1-\lambda_2)+\beta_3(\lambda_3-\lambda_2) & \beta_1(\lambda_1-\lambda_3)+\beta_2(\lambda_2-\lambda_3)\\ \beta_2(\lambda_1-\lambda_2) & (\lambda_1\beta_2+\lambda_2(\beta_1+\beta_3)) & \beta_2(\lambda_1-\lambda_2)\\ \beta_3(\lambda_1-\lambda_3) & \beta_3(\lambda_1-\lambda_3) & (\lambda_1\beta_3+\lambda_3(\beta_1+\beta_2)).\end{pmatrix}\end{aligned}$$

6.1.2 Résolution par un schéma de différences finies sparse

La méthode de différences finies sur une grille sparse est appliquée à l'équation (6.13). Les remarques suivantes peuvent être faites :

- il est envisageable de raffiner les variables pour lesquelles il y a un terme de diffusion et de discrétiser de manière plus grossière les variables pour lesquelles il n'y a qu'un terme de transport. La solution est a priori régulière pour ces variables.
- le choix de x_1 réduit dans l'algorithme le nombre de calculs nécessaires à la multiplication matrice-vecteur (voir la discussion du paragraphe 9.2.3.2) puisqu'alors $f(x) = f(x_1)$.

Afin de vérifier les hypothèses du théorème 4.21, le domaine $\Omega = \Omega_s \times \Omega_{x_1} \times \Omega_{x_2} \times \cdots \times \Omega_{x_d}$ est construit de manière à ce que les flux soient sortants sur $\partial\Omega_i,\ i > 1$.

6.1.2.1 Résultats

Les résultats sont exprimés par rapport à la *relative Premium*, $RP = \dfrac{Prix}{Spot}100$. Cette quantité est proportionnelle à la volatilité implicite lorsque le Spot est à la monnaie forward $\left(S_0^F = K\exp\left(-\int_0^T r(u)du\right)\right)$. Dans ce cas, le prix d'un call dans le modèle de Black & Scholes vérifie

$$P_{call}^{BS}\left(T,S_0^F\right) \approx \frac{S_0^F\sigma_{impli}}{\sqrt{2\pi T}} \Rightarrow RP_{call} \approx \frac{\sigma_{impli}^{(\%)}}{\sqrt{2\pi T}},$$

où $\sigma_{impli}^{(\%)} = 100 * \sigma_{impli}$ et σ_{impli} est la volatilité implicite Black & Scholes du prix. L'erreur pour une option vanille est exprimée en point de base (bp) de volatilité implicite :

$$\sigma_{impli} = 1e^{-4} = 1\ bp^\sigma, \Rightarrow \sigma_{impli}^{(\%)} = 1\% = 1\ bp^\sigma.$$

La maturité T est de un an. Le taux d'intérêt est constant $r = 0$. Pour les tests, nous nous plaçons à la monnaie, Spot $\approx$ Strike. Le tableau 6.1 donne la *relative premium* d'un

Call dans le modèle SVS 3 facteurs obtenue par une méthode de différences finies sur une Sparse Grid anisotrope.

La prime attendue (obtenue par une méthode de Monte Carlo) est de $13.770(+/-0.01)$. L'erreur, exprimée en bp dans le tableau, est calculée par rapport à ce prix de référence. Les temps de calcul ne sont pas optimaux, cette méthode de simulation ayant été rapidement abandonnée.

Nous nous plaçons dans les conditions équivalentes à la configuration optimale détaillée dans la section suivante.

Prix	Erreur bp	l:= (S, z_1, z_2, z_3)	Nb pts	Time(s)	Time (min)
13.874	-10.4	$(6, 4, 3, 3)$	7540	366.5	$6min07$
13.826	-5.6	$(7, 3, 3, 3)$	11746	629.01	$10min$
13.767	0.3	$(7, 4, 3, 3)$	14738	809.33	$13min29$
13.792	-2.2	$(7, 5, 3, 3)$	17906	983.31	$16min23$
13.790	-2.0	$(7, 4, 4, 4)$	18002	954.512	$15min$
13.804	-3.4	$(7, 5, 5, 5)$	26210	1416.32	$23min$
13.800	-3.0	$(7, 7, 7, 7)$	52994	3348.01	$55min$
13.752	1.8	$(8, 5, 3, 3)$	35058	2781.04	$46min$
13.771	-0.1	$(8, 4, 4, 4)$	35410	1817.69	$30min$
13.764	0.6	$(8, 5, 5, 5)$	50914	3790	$1h03min$
13.753	1.7	$(8, 4, 3, 3)$	29074		$44min$

TAB. 6.1 – Résultat sur l'équation (6.13) avec $\beta = (1.5213, 0.4628, 0.4048)$, $\lambda = (5.322, 2.520, 0.1197)$

6.1.2.2 Commentaires

La convergence de la méthode, lorsque les paramétres de discrétisation varient, n'apparaît pas clairement. Ce phénomène s'explique par l'influence des conditions aux limites sur les bords artificiels. En effet, le choix de λ, par exemple ($\lambda = (30, 1, 0.1)$), induit des flux entrants sur $y_1 = \pm y_1^{max}$ qui polluent la solution à l'intérieur du domaine de calcul.

De plus, les temps de calcul sont importants, car nous ne profitons pas pleinement des *Sparse Grids* à cause des conditions aux bords. Nous remarquons également que le résidu présente une décroissance lente près de la frontière artificielle au cours des itérations du GMRES.

Pour ces raisons, nous avons préféré abandonner cette méthode et proposer une formulation différente avec des conditions aux limites de Dirichlet homogènes sur les frontières artificielles.

6.1.3 Conclusion et perspectives sur la formulation ultra-parabolique

La compétitivité d'une méthode de *Sparse Grid* est fortement liée aux conditions aux bords imposées à l'équation. L'application des méthodes de *Sparse Grid* au problème à convection dominante est une question ouverte. Quelques résultats sur cette question sont proposés dans [SST07]. Nous avons abandonné cette formulation pour une résolution sparse, elle semble néanmoins être intéressante pour d'autres méthodes, par exemple :

Méthode des caractéristiques Nous décrivons dans ce paragraphe une méthode envisagée pour la résolution numérique d'une équation aux dérivées partielles de la forme (6.10) :

Plaçons nous sur la caractéristique $X_t^{t_{n+1},x}$, solution de l'équation différentielle :

$$\begin{aligned} \frac{\partial X^{t_{n+1},x}}{\partial t}(t) &= LX^{t_{n+1},x}(t), \qquad t \in (t_n, t_{n+1}), \\ X^{t_{n+1},x}(t_{n+1}) &= x. \end{aligned} \tag{6.24}$$

alors la fonction v définie par $v(t,s,x) = u\left(t,s,X^{t_{n+1},x}(t)\right)$ vérifie sur l'intervalle (t_n, t_{n+1})

$$\frac{\partial v}{\partial t} = \frac{\partial u}{\partial t} + (Lx)^T \nabla_x u, \tag{6.25}$$

et donc vérifie l'équation

$$\frac{\partial v}{\partial t} + \mathcal{L}v \simeq 0, \tag{6.26}$$

où

$$\mathcal{L}v = -\frac{1}{2}f(x)^2 s^2 \frac{\partial^2 v}{\partial s^2} - r\ s\frac{\partial v}{\partial x} + rv - \frac{1}{2}\overline{\beta}^2 \frac{\partial^2 v}{\partial x_1^2} - \overline{\beta}\rho f(x) s \frac{\partial^2 v}{\partial s \partial x_1}.$$

Discrétisons en temps l'équation (6.26) par un schéma d'Euler implicite :

$$v(t_{n+1},s,x) - v(t_n,s,x) + \Delta_{t_n}\mathcal{L}(v)(t_{n+1},s,x) = 0. \tag{6.27}$$

En notant :

$$v(t_{n+1},s,x) = u^{n+1}(s,x), \quad \text{et} \quad v(t_n,s,x) = u\left(t_n,s,X^{t_{n+1},x}(t_n)\right) = u^n\left(s, X^{t_{n+1},x}(t_n)\right),$$

nous obtenons la semi-discrétisation en temps :

$$u^{n+1}(s,x) - u^n\left(s, X^{t_{n+1},x}\right) + \Delta_{t_n}\mathcal{L}\left(u^{n+1}\right)(s,x) = 0, \quad u^0(s,x) = h(s). \tag{6.28}$$

Remarque 6.1 *Nous nous ramenons entre chaque pas de temps à $N_{x_2} \times \cdots \times N_{x_n}$ équations elliptiques découplées en dimension 2 (N_{x_k} représente le nombre de points de discrétisation sur la variable x_k).*

Remarque 6.2 *Une parallélisation possible de la méthode consiste à découper le domaine en tranches : $\underline{z_k} \leq z_k \leq \overline{z_k}$. La direction k est choisie comme étant celle dans laquelle les variations de la solution sont les plus faibles. De cette manière, le nombre de communications est minimisé.*

Solution semi-analytique du problème Si la fonction de volatilité σ était bornée (inférieurement et supérieurement), l'approche proposée par [MDF05] permettrait d'aboutir à une solution semi-analytique de l'équation (6.10). En effet, l'équation (6.10) appartient à la famille des équations différentielles de Kolmogov définies par :

$$\frac{\partial u}{\partial t} + Lu = \frac{\partial u}{\partial t} + \sum_{i,j=1}^{p_0} a_{i,j}(z)\frac{\partial^2 u}{\partial x_i \partial x_j}u + \sum_{i}^{p_0} a_i(z)\frac{\partial u}{\partial x_i} + \sum_{i=1}^{d} b_{i,j}x_i\frac{\partial u}{\partial x_j} + c(z)\,u = 0, \tag{6.29}$$

où $z = (x,t) \in \mathbb{R}^N \times \mathbb{R}^+$ et $1 \leq p_0 = 2 \leq d = n+1$.

6.2 Résolution de (5.13) par une méthode de différences finies sparse

Nous présentons ici les résultats numériques obtenus par une méthode de différences finies sparse sur l'équation de valorisation (5.13) où la fonction de volatilité f est donnée par (5.16). Avant de présenter les résultats numériques, nous reviendrons sur les problèmes d'approximation et de conditions aux bords. Nous proposons notamment une reformulation de l'équation qui soit cohérente avec la méthode des *Sparse Grid*.

Nous précisons ensuite quelques difficultés intrinsèques à notre équation et quelques paramètres de la méthode de résolution avant d'expliciter les résultats.

6.2.1 Localisation du problème

Le problème initial (5.13) est posé sur un domaine non borné : $\mathbb{R}^+ \times \mathbb{R}^n$. La résolution par une méthode numérique nécessite une formulation approchée du problème (5.13) posé sur un domaine borné : cette étape correspond à la *localisation* du problème. La méthode la plus communément utilisée pour localiser consiste à tronquer le domaine et à résoudre un problème aux bords avec des conditions de type Dirichlet ou Neumann bien choisies. Il n'est pas toujours facile de trouver ces conditions. Dans notre exemple, ceci est particulièrement vrai pour $y_i = y_i^{max}$. Un raffinement de la méthode de localisation permettant de justifier les conditions aux bords est ici présentée.

Pour être cohérent avec l'analyse théorique du problème du chapitre 5, le prix est multiplié par une fonction à décroissance rapide quand $|y_i| \to \infty$. La fonction choisie est la même que celle utilisée en (5.43) pour obtenir la *formulation faible* au théorème 5.10. La fonction u est définie par

$$u_\eta(s, y_1, \ldots, y_n, t) = P(s, y_1, \ldots, y_n, t)\, e^{-(1-\eta)\frac{y^T \Pi y}{2}} = P(s, y_1, \ldots, y_n, t)\ \Gamma(y_1, \ldots, y_n).$$

Ceci nous permet d'imposer des conditions de type Dirichlet homogènes sur les bords $y_i = y_i^{max}$ et $y_i = -y_i^{max}$. Voyons à présent ce qu'il convient de faire pour pouvoir imposer des conditions de Dirichlet homogènes sur le bord $s = s_{max}$.

6.2.1.1 Introduction d'une surprime

A nouveau, nous ne résolvons pas le problème portant sur le prix de l'option, mais l'équation vérifiée par une surprime définie comme la différence entre le prix dans notre modèle et le prix donné par un modèle de Black & Scholes.

Remarque 6.3 *La méthode présentée ci-dessous reste applicable dans le cas de l'équation de valorisation posée sur la variable $x = log(s)$. Dans ce cas, nous pourrions imposer les mêmes conditions sur les deux bords x_{max} et $x_{min} < 0$.*

Proposition 6.8 *La différence de prix $\pi = (P - P_{BS})\,\Gamma(y_1, \ldots, y_n)$ entre :*
- *le prix P d'une option européenne de maturité T (dont la dynamique du sous-jacent est donnée par l'équation (5.1) et la définition 5.1),*
- *le prix d'une option européenne P_{BS}(dont la dynamique du sous-jacent est donnée par l'équation (3.4)),*

de même fonction Payoff, vérifie l'équation aux dérivées partielles :

$$\frac{\partial \pi}{\partial t} - \frac{1}{2} f(y)^2 s^2 \frac{\partial^2 \pi}{\partial s^2} - rs\frac{\partial \pi}{\partial s} - \frac{1}{2}\nabla_y \cdot \left(\beta\beta^T \nabla_y \pi\right) + (\Lambda Y)^T (\nabla_y \pi) - (1-\eta)\left(\beta^T \Pi Y\right)\beta^T \nabla_y \pi$$
$$-\frac{1-\eta}{2}\left[\beta^T \Pi \beta + (1-\eta)\left(\beta^T \Pi Y\right)^2 - 2(\Lambda Y)^T (\Pi Y)\right]\pi = \frac{1}{2}\left(f(y)^2 - 1\right) s^2 \frac{\partial^2 P_{BS}}{\partial s^2}\Gamma, \tag{6.30}$$

sur $\mathbb{R}^+ \times \mathbb{R}^n \times (0,T)$, *avec la condition initiale* $\pi(s,y,0) = 0, \quad$ *sur* $\mathbb{R}^+ \times \mathbb{R}^n$.

Remarque 6.4 *Dans le cas* $\eta = 1$ *(6.30) devient*

$$\frac{\partial u}{\partial t} - \frac{1}{2} f(y)^2 s^2 \frac{\partial^2 u}{\partial s^2} - \rho \sum_{i=1}^{n} \beta_i s f(y) \frac{\partial^2 u}{\partial s \partial y_i} - \frac{1}{2}\sum_{i,j=1}^{n} \beta_i \beta_j \frac{\partial^2 u}{\partial y_i \partial y_j}$$
$$-r(t)\left(s\frac{\partial u}{\partial s} - u\right) - \sum_{i=1}^{n} \lambda_i y_i \frac{\partial u}{\partial y_i} = \frac{1}{2}\left(f(y)^2 - 1\right) s^2 \frac{\partial^2 P_{BS}}{\partial s^2}, \tag{6.31}$$

sur $\mathbb{R}^+ \times \mathbb{R}^n \times (0,T)$, *avec la condition initiale* $\pi(s,y,0) = 0, \quad$ *sur* $\mathbb{R}^+ \times \mathbb{R}^n$.

Remarque 6.5 *Le choix de la fonction de référence n'est pas anodin car le second membre de l'équation (6.31) a une expression analytique donnée par (3.17).*

La restriction de π à $\Omega \times (0,T)$ est approchée par $\widetilde{\pi}$ solution de

$$\frac{\partial \widetilde{\pi}}{\partial t} - \frac{1}{2} f(y)^2 s^2 \frac{\partial^2 \widetilde{\pi}}{\partial s^2} - rs\frac{\partial \widetilde{\pi}}{\partial s} - \frac{1}{2}\nabla_y \cdot \left(\beta\beta^T \nabla_y \widetilde{\pi}\right) + (\Lambda Y)^T (\nabla_y \widetilde{\pi}) - (1-\eta)\left(\beta^T \Pi Y\right)\beta^T \nabla_y \widetilde{\pi}$$
$$-\frac{1-\eta}{2}\left[\beta^T \Pi \beta + (1-\eta)\left(\beta^T \Pi Y\right)^2 - 2(\Lambda Y)^T (\Pi) Y\right]\widetilde{\pi} = \frac{1}{2}\left(f(y)^2 - 1\right) s^2 \frac{\partial^2 P_{BS}}{\partial s^2}\Gamma,$$
$$\widetilde{\pi}(s,y,t) = 0, \quad (s,y) \in \partial\Omega \tag{6.32}$$

avec la condition de Cauchy $\widetilde{\pi}(s,y,T) = 0$ sur Ω.

Les expériences numériques montrent que le choix du paramètre η n'influe pas sur l'erreur de résolution de la méthode numérique dans la zone d'intérêt. En particulier, en choisissant $\eta = 1$ et en imposant des conditions de Dirichlet homogènes sur les bords y, nous obtenons des erreurs comparables. Ce choix, minimisant le nombre d'opérations, a donc été préféré.

6.2.2 Résultats numériques

Les résultats présentés ci-dessous sont obtenus en résolvant numériquement (6.32) par une méthode de différences finies sparse.

6.2.2.1 Paramètres du modèle

La dynamique du processus de volatilité est décrite par trois facteurs. L'équation de valorisation (6.32) est une EDP en dimension 4. Nous nous plaçons à taux d'intérêt nul pour

simplifier l'analyse des résultats et nous considérons des valeurs de paramètres cohérentes par rapport à des données de marché. Les paramètres de modèles sont

$$\rho = -0.5,\ V_0 = 0.2, \quad \lambda = (29.27, 1.46, 0.108)\,, \quad \beta = (1.26, 0.423, 0.421)\,.$$

La fonction payoff est celle d'un call européen, et la maturité est de un an.

L'équation (5.13) n'a pas, à notre connaissance, de solution semi-analytique. Les estimations d'erreurs sont obtenues en comparant l'approximation numérique de (6.32) aux résultats obtenus par une méthode de Monte-Carlo. La discrétisation de la méthode de Monte-Carlo correspond à un schéma d'Euler à 365 pas de temps (1 pas par jour). Les résultats de la méthode de Monte-Carlo sont obtenus pour 500000 tirages. L'estimation de l'erreur (« trust ») est de l'ordre de $8.0e^{-4}$, mais l'erreur de discrétisation en temps doit également être considérée.

La fonction e qui nous permettra de mesurer l'erreur est :

$$e = \left(\sum_{i=-20}^{20} \left(P_{MC} - P_{EDP}\right)^2 (s_i, 0, \ldots, 0) \right)^{\frac{1}{2}}, \quad s_i = K\left(1 + \frac{i}{100}\right). \tag{6.33}$$

L'erreur est donc calculée au voisinage de la monnaie sur les points tels que $y_i = 0,\ 1 \leq i \leq 3$. Le code Monte-Carlo ne permet pas de calculer le prix pour d'autres valeurs de y_i.

6.2.2.2 Paramètres de la méthode numérique

Nous présentons, tout d'abord, les résultats dans une configuration que nous nommerons optimale dans le sens où elle donne le meilleur rapport « précision demandée / temps de calcul ». Nous étudierons ensuite d'autres configurations en changeant les paramètres de l'algorithme. Définissons ces différents paramètres.

- L'introduction de frontières artificielles induit une erreur. Nous considérons l'erreur liée à la localisation suivant les variables y_i. Le choix du domaine dépend de la variance du processus. Cette variance est obtenue au paragraphe 5.1.1. Plus précisément, le domaine est un hypercube caractérisé par les deux sommets $y^{max} = (y_1^{max}, \ldots, y_n^{max})\,,\ y^{min} = -y^{max}$ où

$$y_i^{max} = F_\sigma \left(\frac{\beta_i^2}{2\lambda_i} \left(1 - \exp\left(-2\lambda_i T\right)\right) \right)^{\frac{1}{2}},$$

  avec F_σ est un paramètre de la discrétisation.
- Le schéma de discrétisation en temps introduit une nouvelle source d'erreur. Nous aurons le choix entre les différents θ-schémas, différents pas de temps ou différentes fonctions de distribution de ces pas de temps.
- La tolérance pour la résolution du système linéaire par une méthode itérative est un paramètre responsable d'une troisième source d'erreur. Notons Tol ce paramètre.

Pour être complète, l'analyse doit également considérer l'erreur liée à la méthode de Monte-Carlo. La configuration optimale correspond à $F_\sigma = 6$, un schéma de Crank Nicolson avec 100 pas de temps, une fonction de distribution des pas de temps que l'on précisera par la suite et une tolérance sur le solveur itératif $Tol = 1e^{-4}$. Les tests sont réalisés sur un processeur Intel Core2 6600@2.40GHz muni de 4GB de ram.

6.2.2.3 Résultats sur la configuration optimale

Les résultats et les temps de calcul sont donnés pour des niveaux de discrétisation variant de 5 à 10. Dans le tableau 6.2, figure le nombre de points de chacune des grilles correspondant à ces niveaux de discrétisation en dimension 4.

TAB. 6.2 – Nombre de points d'une grille sparse en dimension 4

Niveau	5	6	7	8	9	10
Nombre de points	770	2562	7938	23298	64538	150018

Le tableau 6.3 contient les résultats obtenus par une résolution de l'équation de valorisation pour différentes valeurs du spot et différents niveaux de précision. A partir du niveau 7, les résultats sont acceptables, en particulier pour des problèmes liés à la calibration. Ce tableau nous permet de calculer l'erreur e estimée à comparer avec l'erreur théorique aux points en $0\left(2^{-2n}n^{d-1}\right)$ vérifiée par la discrétisation sparse. L'erreur théorique et l'erreur de convergence e sont représentées à la figure 6.1 sur une échelle logarithmique. Il apparaît que la vitesse de convergence entre les niveaux 6 et 9 est conforme au comportement théorique. L'erreur pour le niveau 10 est difficile à analyser, l'erreur de discrétisation n'est plus prépondérante. Enfin, le tableau 6.4 donne les temps de calcul en fonction du niveau de discrétisation. Nous en déduisons la figure 6.2 qui représente l'erreur estimée par rapport au temps de calcul exprimée en échelle logarithmique. L'erreur évolue en $\frac{1}{\sqrt{T_c}}$ où T_c est le temps de calcul. La complexité est donc comparable à celle d'une méthode de Monte-Carlo, pour obtenir un seul prix, si l'on considère que le temps de calcul est proportionnel au nombre de tirages.

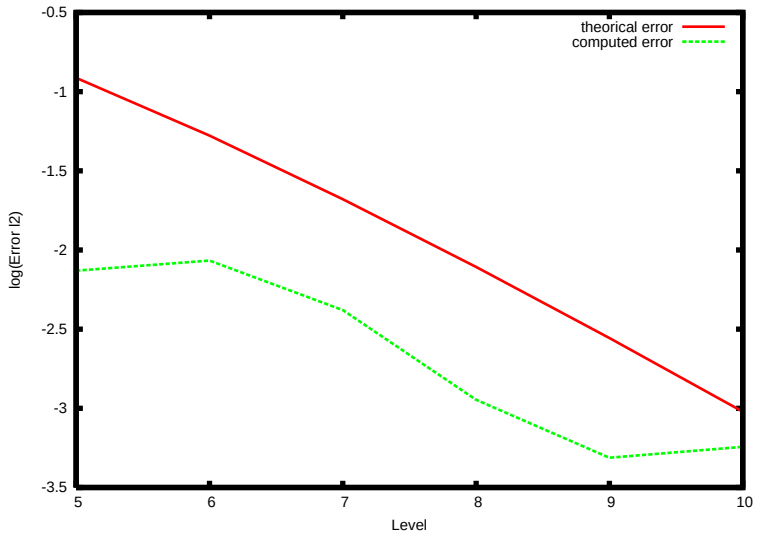

FIG. 6.1 – Erreur e en fonction du raffinement - avec le schéma $CN(100)$ et les paramètres $F_\sigma = 6, Tol = 1e-4$.

TAB. 6.3 – Prix d'un call européen de maturité 1 an (multiplié par 100)- avec le schéma $CN(100)$ et les paramètres $F_\sigma = 6, Tol = 1e-4$.

Spot	MC 500000	6	7	8	9	10
0.8	6.630	6.4989	6.5732	6.5994	6.6248	6.6293
0.81	7.029	6.8732	6.9475	7.0204	7.0225	7.0271
0.82	7.4424	7.2589	7.3819	7.4314	7.433	7.444
0.83	7.8679	7.7593	7.7789	7.8551	7.8697	7.8676
0.84	8.306	8.1677	8.2412	8.2914	8.3061	8.3109
0.85	8.7568	8.5873	8.6608	8.7402	8.755	8.7598
0.86	9.2201	9.142	9.1507	9.2014	9.2163	9.2212
0.87	9.6955	9.5839	9.6564	9.675	9.6899	9.703
0.88	10.1828	10.0369	10.1093	10.1607	10.1757	10.1891
0.89	10.682	10.5008	10.6415	10.6585	10.6735	10.6872
0.9	11.1928	11.1171	11.1163	11.1681	11.2011	11.1972
0.91	11.7151	11.6026	11.6742	11.6895	11.723	11.719
0.92	12.2488	12.0988	12.1703	12.2226	12.2565	12.2523
0.93	12.7934	12.6054	12.753	12.767	12.8014	12.7969
0.94	13.3487	13.2796	13.2701	13.3226	13.3574	13.3528
0.95	13.9144	13.8069	13.8766	13.8893	13.9244	13.9196
0.96	14.4903	14.3445	14.4141	14.4668	14.482	14.4973
0.97	15.0762	14.892	15.0435	15.055	15.0701	15.0855
0.98	15.672	15.617	15.601	15.6536	15.6686	15.6842
0.99	16.2777	16.1842	16.252	16.2624	16.2774	16.2826
1	16.8918	16.7611	16.8288	16.8813	16.8961	16.9013
1.01	17.5172	17.3474	17.4151	17.5099	17.5246	17.5298
1.02	18.1509	17.943	18.096	18.148	18.1627	18.1572
1.03	18.7934	18.7216	18.7008	18.7955	18.7888	18.8045
1.04	19.4446	19.3355	19.4006	19.4521	19.4452	19.4503
1.05	20.1045	19.9582	20.0234	20.1175	20.1105	20.1156
1.06	20.7727	20.5898	20.7407	20.7491	20.7845	20.7896
1.07	21.449	21.4033	21.3808	21.4317	21.4458	21.4614
1.08	22.1334	22.0519	22.1144	22.1227	22.1366	22.1416
1.09	22.8257	22.7088	22.7713	22.8216	22.8353	22.8403
1.1	23.5256	23.3738	23.4363	23.5283	23.5419	23.5365
1.11	24.2332	24.0469	24.1929	24.2426	24.2355	24.2507
1.12	24.948	24.895	24.8738	24.9235	24.9572	24.9621
1.13	25.6699	25.5836	25.6439	25.6528	25.686	25.6809
1.14	26.3989	26.2798	26.340	26.389	26.4021	26.4069
1.15	27.1348	26.9833	27.0436	27.132	27.1448	27.1497
1.16	27.8773	27.6941	27.8332	27.8815	27.875	27.8894
1.17	28.626	28.5683	28.5511	28.5994	28.631	28.6358
1.18	29.381	29.2932	29.3516	29.3622	29.3931	29.3887
1.19	30.142	30.0249	30.0834	30.1309	30.1431	30.1479
1.2	30.909	30.7633	30.8218	30.9053	30.9173	30.9132

TAB. 6.4 – Erreur attendue & erreur estimée - avec le schéma $CN(100)$ et les paramètres $F_\sigma = 6, Tol = 1e-4$.

Niveau	5	6	7	8	9	10
théorique	$1,22 \cdot 10^{-1}$	$5,27 \cdot 10^{-2}$	$2,09 \cdot 10^{-2}$	$7,81 \cdot 10^{-3}$	$2,78 \cdot 10^{-3}$	$9,53 \cdot 10^{-4}$
estimée	$7.40 \cdot 10^{-3}$	$8.57 \cdot 10^{-3}$	$4.17 \cdot 10^{-3}$	$1.13 \cdot 10^{-3}$	$4.87 \cdot 10^{-4}$	$5.70 \cdot 10^{-4}$
Tps (s)	1.74	7.71	37.8	228	1163	3056

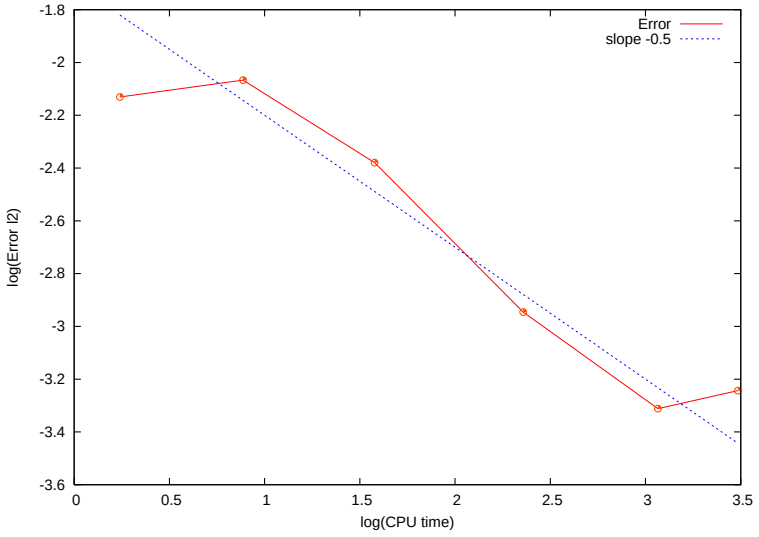

FIG. 6.2 – Erreur e en fonction du temps de calcul - avec le schéma $CN(100)$ et les paramètres $F_\sigma = 6, Tol = 1e-4$.

6.2.2.4 Résultats sur différentes configurations

Semi-discrétisation en temps Il est difficile d'observer l'erreur liée à la discrétisation en temps, voir la figure 6.3. Les autres erreurs dominent l'erreur liée à la discrétisation en temps. Différentes explications peuvent être avancées. La plus pertinente est sans doute de remarquer que nous comparons deux méthodes de résolution numériques avec une discrétisation en temps. En particulier, le pas de temps de la méthode de Monte-Carlo est de un jour : il y a 360 pas de temps dans l'exemple considéré.

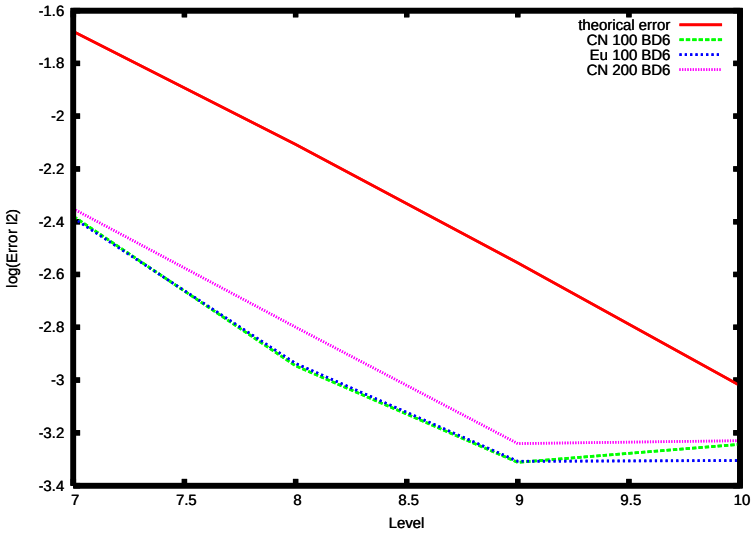

FIG. 6.3 – Erreur en fonction de différentes discrétisations en temps avec les paramètres $F_\sigma = 6$ et une précision de $Tol = 1e^{-4}$.

Le tableau 6.5 reprend les valeurs de la figure 6.3. Il indique également le comportement du solveur itératif en fonction du nombre de pas de temps. Il faut bien noter qu'avec ce schéma de discrétisation, le temps de calcul n'est pas proportionnel au nombre de pas de temps. En effet, si des pas de temps sont plus petits, la résolution du système linéaire nécessite moins d'itérations. A titre indicatif, le passage de 100 à 200 pas de temps pour un schéma de Crank-Nicolson entraîne, au niveau 10, une augmentation du temps de calcul de l'ordre de 20%, car celui-ci passe de 51 minutes à 61 minutes.

TAB. 6.5 – Erreur estimée & nombre d'itérations moyen pour le solveur GMRES - avec les schémas CN(100) et CN(200) et les paramètres $F_\sigma = 6$, $Tol = 1e^{-4}$. Le nombre d'itérations donne une bonne indication de l'évolution du temps de calcul en fonction du nombre de pas de temps

Niveau	7	8	9	10
Erreur estimée	$4.174 \cdot 10^{-3}$	$1.133 \cdot 10^{-3}$	$4.878 \cdot 10^{-4}$	$5.708 \cdot 10^{-4}$
Itérations	9	13	25	27
Erreur estimée	$4,4533 \cdot 10^{-3}$	$1,586 \cdot 10^{-3}$	$5,745 \cdot 10^{-4}$	$5,8895 \cdot 10^{-4}$
Itérations	5	8	13	16

La distribution des pas de temps suit une loi de puissance, le $i^{\text{ème}}$ pas de temps est

donné par :

$$(\Delta t)_i = \frac{1}{T}\left(\left(\frac{i}{N_T}\right)^{\alpha} - \left(\frac{i-1}{N_T}\right)^{\alpha}\right), \tag{6.34}$$

où N_T est le nombre de pas de temps, T le temps final et $i \in \{1, \dots N_T\}$. Nous choisissons $\alpha = 1.5$. La figure 6.4 donne les pas de temps pour deux valeurs de $N_T = \{100, 200\}$ et $\alpha = 1.5$. Ce choix a permis d'équilibrer le nombre d'itérations de la méthode GMRES entre les pas de temps. Dans le cas $N_T = 100$, les pas de temps sont dans l'intervalle $[0.00102, 0.0151131890912513]$.

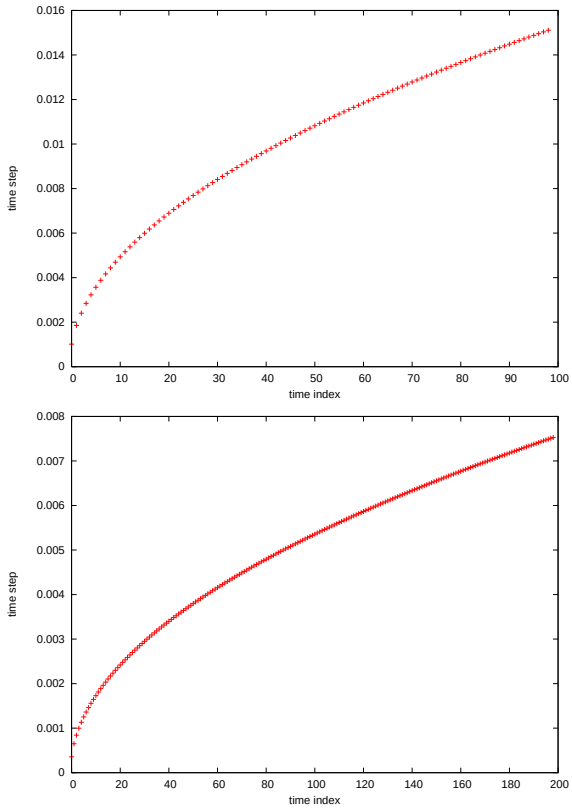

FIG. 6.4 – Discrétisation en temps pour 100 et 200 pas de temps.

Remarque 6.6 *Nous avons étudié (voir § 2.6) les propriétés du schéma d'Euler implicite et, plus généralement, les schémas dissipatifs. Dans le cas présent, le payoff étant une fonction à une seule variable, la régularisation du schéma en temps n'est donc pas un facteur prépondérant, contrairement au cas d'options multi sous-jacents. En effet, si les coefficients de l'EDP sont des fonctions à variables séparées, la discrétisation sur une grille sparse n'ajoute pas d'erreur par rapport à une méthode classique.*

Localisation Nous avons fait varier F_σ dans l'ensemble $\{4, 6, 10\}$ afin de mesurer l'impact des conditions aux bords sur l'erreur. Notons que ce choix influe également sur le nombre d'itérations du solveur GMRES. La figure 6.5 indique les erreurs d'approximation pour les différents facteurs F_σ. Le tableau 6.6 indique l'influence de cette localisation sur le nombre d'itérations de la méthode GMRES.

TAB. 6.6 – Erreur estimée et temps de calcul - pour le schéma $CN(100)$ et le paramètre $Tol = 1e^{-4}$

Niveau	5	6	7	8	9	10
$F_\sigma = 4$	$1,57 \cdot 10^{-2}$	$8,19 \cdot 10^{-3}$	$4,22 \cdot 10^{-3}$	$2,22 \cdot 10^{-3}$	$1,41 \cdot 10^{-3}$	$1,10 \cdot 10^{-3}$
Tps (s)	1.85	8.41	48.86	292	1502	3782
$F_\sigma = 6$	$7.40 \cdot 10^{-3}$	$8.57 \cdot 10^{-3}$	$4.17 \cdot 10^{-3}$	$1.13 \cdot 10^{-3}$	$4.87 \cdot 10^{-4}$	$5.70 \cdot 10^{-4}$
Tps (s)	1.74	7.71	37.8	228	1163	3056

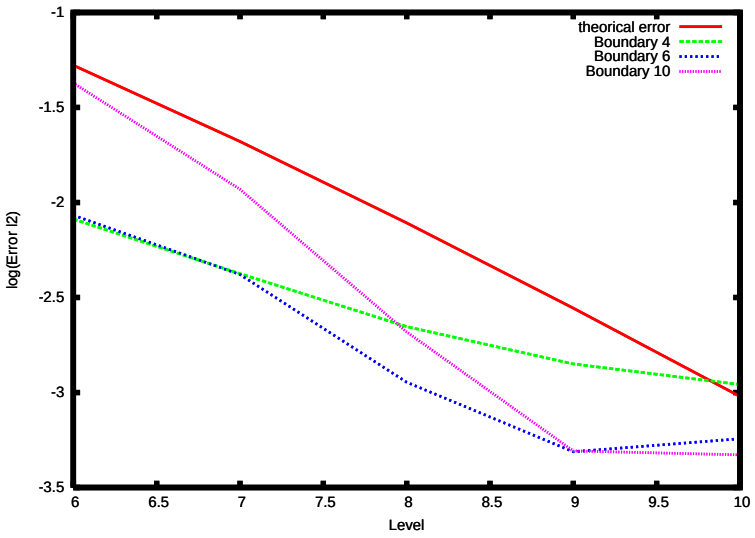

FIG. 6.5 – Erreur en fonction de F_σ- $CN(100)$, $Tol = 1e^{-4}$.

Couches limites La figure 6.6 représente une coupe de la solution à l'instant final au point $y_i = 0$, $i = 2, 3$ ou $i = 1, 2$ ou $i = 1, 3$.

Remarquons qu'il s'agit d'un phénomène de couches limites sur chacune de ces coupes. La condition aux bords de type Dirichlet homogènes contraint la solution à avoir de forts gradients au voisinage du bord artificiel. Une nouvelle fois, il est clair que le choix des conditions de Dirichlet homogènes n'est pas le plus adapté au problème. Il est toutefois fortement souhaitable pour appliquer la méthode de résolution numérique. Nous avons observé graphiquement que la taille de ces couches limites dépend du niveau de discrétisation et que la solution n'est que faiblement polluée dans la région d'intérêt.

La figure 6.6 pourrait nous inciter à utiliser un niveau de raffinement plus faible pour la variable y_1. Cependant, il faut savoir que $\dfrac{\partial u}{\partial y_1}$ n'est pas petit au voisinage de la maturité (t petit). Les trois variables y_i jouent un rôle similaire à différentes échelles de temps : la surface donnant la dépendance de la solution par rapport à (s, y_1) près de la maturité (t petit) ressemble à la coupe de (s, y_3) plus loin de la maturité.

Solveur itératif Ces résultats peuvent être complétés par l'étude des paramètres du solveur itératif : d'une part, la précision de l'erreur Tol et, d'autre part, le paramètre de redémarrage de la méthode GMRES. Celui-ci influe sur le nombre d'itérations et, par conséquent, sur le temps de calcul. Dans le tableau suivant figure le nombre moyen d'itérations observé pour différentes configurations du solveur itératif. Ces deux tableaux justifient notre choix de configuration dites optimale.

TAB. 6.7 – Erreur estimée - en fonction de Tol - avec un schéma $CN(100)$ et les paramètres $F_\sigma = 6, N = 7$.

GMRES	$1e-3$	$1e-4$	$1e-5$	$1e-6$
Estimée	$2,7986 \cdot 10^{-3}$	$4,1744 \cdot 10^{-3}$	$3,8821 \cdot 10^{-3}$	$4.0278 \cdot 10^{-3}$
Itérations	4	8	12	18

TAB. 6.8 – Itérations du GMRES pour les différents paramètres : $F_\sigma, k,\ M$- avec $Tol = 1e^k$ et un schéma $CN(100)$. M est le paramètre de restart du GMRES.

Niveau	$(6,4,10)$	$(6,6,10)$	$(4,3,10)$	$(4,6,10)$	$(4,6,5)$	$(4,6,20)$
6	4	8	6	10	13	9
7	8	18	11	17	21	16
8	13	25	17	27	34	29
9	25	43	30	45	55	46
10	27	45	33	47	59	48

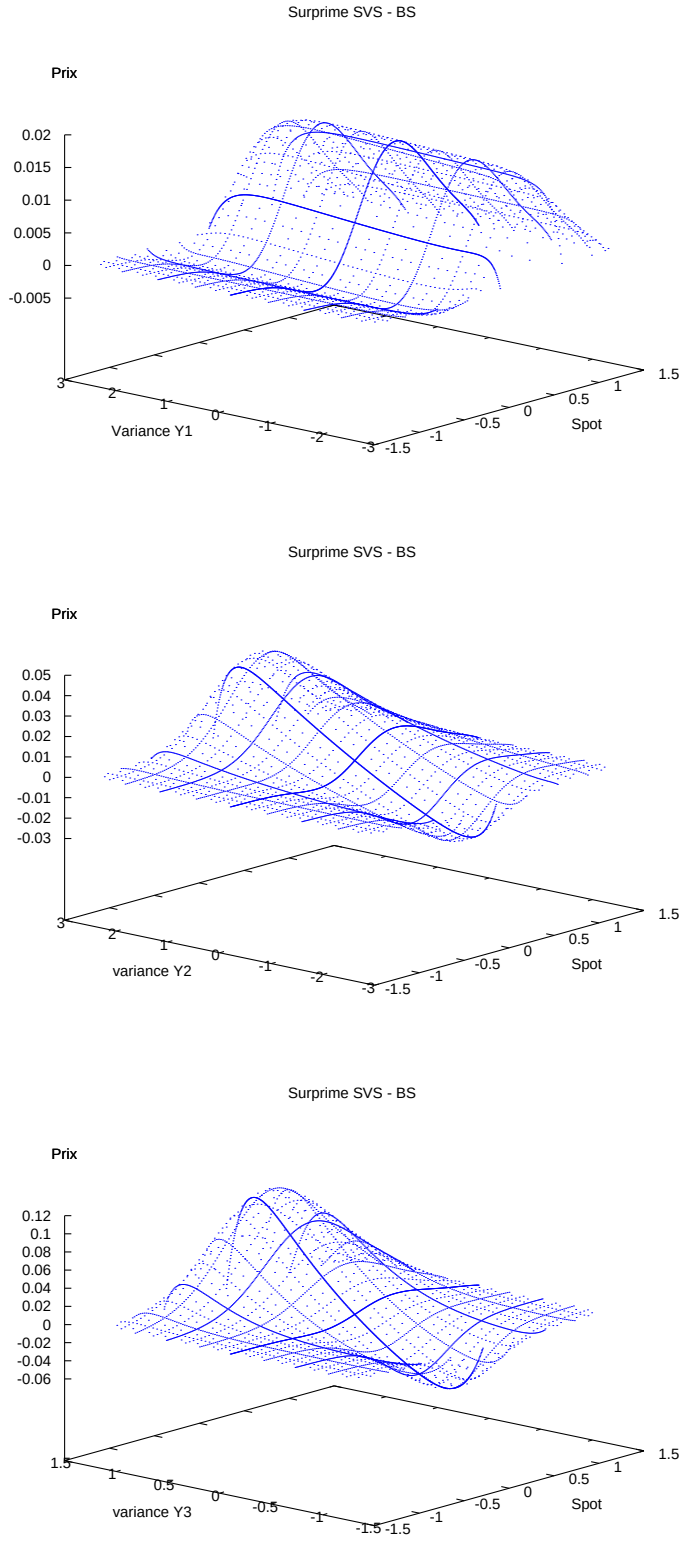

FIG. 6.6 – Représentation de la surprime en coupe par rapport au spot et chacune des variables y_i dans l'ordre décroissant des vitesses de retour à la moyenne. Des couches limites sont observées au voisinage des frontières artificielles.

Chapitre 7

Modèle à volatilité stochastique avec saut

Dans ce chapitre, le modèle à volatilité stochastique, présenté au chapitre 5, est étendu au cas de processus à saut.

Différents modèles à volatilité stochastique avec saut sont décrits dans la première partie. La notion de diffusion de Lévy est définie dans la deuxième partie. Cette notion décrit la dynamique de notre modèle de diffusion. L'équation de valorisation d'une option européenne est établie dans la troisième partie. Les résultats numériques obtenus par une méthode de différence finie sur une grille sparse font l'objet de la quatrième partie.

7.1 Modèle à volatilité stochastique & Processus à saut

7.1.1 Processus de Lévy

7.1.1.1 Définition

Définition 7.1 (Processus de Lévy, extrait de [CT03]) *Soit* $(\Omega, \mathcal{F}, \{F_t\}_{t\geq 0}, P)$ *un espace de probabilité muni de la filtration* $\{F_t\}_{t\geq 0}$. *Le processus* $\{X_t\}_{t\geq 0} \subset \mathbb{R}^d$ *adapté à la filtration* $\mathcal{F}_t$ *tel que* $X_0 = 0$ *est un* processus de Lévy *s'il vérifie les propriétés suivantes :*

- *les incréments sont indépendants : pour toute suite croissante de temps :* $t_0 < t_1 < \cdots < t_n$, *les variables aléatoires* $X_{t_0}, X_{t_1} - X_{t_0}, \ldots, X_{t_n} - X_{t_{n-1}}$ *sont indépendantes ;*
- *stationnarité des incréments : la loi de* $X_{t+h} - X_t$ *ne dépend pas de* t *;*
- *le processus est continu en probabilité :* $\forall \epsilon > 0, \lim_{h\to 0} \mathbb{P}\{|X_{t+h} - X_t| \geq \epsilon\} = 0.$

Remarque 7.1 *Soit* $\{X_t\}$ *un processus de Lévy, alors il existe une unique version de* X_t *qui soit* cad-lag *(continu à droite et admettant une limite à gauche) et qui soit un processus de Lévy (voir [Pro04a], [Sat99])*

Nous supposerons que les processus de Lévy considérés sont cad-lag.

Définition 7.2 (Mesure aléatoire de Poisson) *Le* saut *de* X_t *à* $t \geq 0$ *est défini par*

$$\Delta X_t = X_t - X_{t^-} \ . \tag{7.1}$$

Soit $\boldsymbol{B}_0$ *une famille d'ensembles boréliens* $U \subset \mathbb{R}^d$ *telle que l'adhérence* $\bar{U}$ *ne contienne pas* 0. *Pour* $U \in \boldsymbol{B}_0$, *nous pouvons définir la mesure*

$$N(t,U) = N(t,U,\omega) = \sum_{s:0<s\leq t} \mathcal{X}_U(\Delta X_s)\,, \tag{7.2}$$

où $\mathcal{X}_U$ *est la fonction caractéristique de* U. *En d'autres termes,* $N(t,U)$ *est le nombre de sauts de taille* $\Delta X_s \in U$ *qui se réalisent avant le temps ou au temps* t. *Nous nommons* $N(t,U)$ *la* mesure aléatoire de Poisson *de* X_t.

Définition 7.3 *Soit* $(X_t)_{t\geq 0}$ *un processus défini sur* $\mathbb{R}^d$. *La mesure* ν *de* $\mathbb{R}^d \to \mathbb{R}$ *définie par :*

$$\nu(A) = \mathbb{E}\left[\#\left\{t \in [0,1] \,\middle|\, \Delta X_t \neq 0,\ \Delta X_t \in A\right\}\right], \quad \forall A \in \boldsymbol{B}\left(\mathbb{R}^d\right), \tag{7.3}$$

est la mesure de Lévy de X *:* $\nu(A)$ *est l'espérance du nombre de sauts de taille appartenant à* A, *par unité de temps.*

7.1.1.2 Propriétés

Nous rappelons la propriété d'un processus de Lévy qui permet d'obtenir la formulation déterministe sous forme d'équation intégro-différentielle des problèmes d'évaluation d'options. Cette propriété est une conséquence de la définition d'un processus de Lévy qui est un processus markovien. Ceci nous permet de définir une fonction caractéristique. Dans le cas d'un processus de Lévy, cette fonction est définie explicitement. Nous étudierons, par la suite, le lien entre la fonction caractéristique et le générateur infinitésimal (voir la définition 4.1).

Définition 7.4 *Soit* $(X_t)_{t\geq 0}$ *un processus de Markov,* la fonction caractéristique Φ_t *est définie par*

$$\Phi_t(\omega) = \Phi_{X_t}(\omega) = \mathbb{E}\left[\exp\left(\imath\,\langle \omega, X_t\rangle\right)\right].$$

Proposition 7.1 *Si* $(X_t)_{t\geq 0}$ *est un processus de Lévy sur* $\mathbb{R}^d$, *alors il existe une fonction continue, nommée* exposant caractéristique, $\psi : \mathbb{R}^d \to \mathbb{R}$ *telle que*

$$\Phi_t(\omega) = \exp\left(t\psi(\omega)\right),\ \omega \in \mathbb{R}^d. \tag{7.4}$$

Le reste de ce paragraphe permet de décrire cet exposant caractéristique.

Théorème 7.2 (Décomposition de Lévy) *Soit* $(X_t)_{t\geq 0}$ *un processus de Lévy sur* $\mathbb{R}^d$. *Alors* X_t *admet la décomposition*

$$X_t = \alpha t + \Sigma W_t + \int_{z\in B_R} z\widetilde{N}(t,dz) + \int_{z\notin B_R} zN(t,dz)\,, \tag{7.5}$$

où W_t *est un Brownien et* $N(t,dz)$ *une mesure aléatoire de Poissson indépendante de* W_t *et* $\alpha \in \mathbb{R}^d$, $\Sigma \in \mathbb{R}^{d\times m}$, $B_R = \left\{z \in \mathbb{R}^d \middle|\, |z_i| \leq R \quad \forall 1 \leq i \leq d\right\}$. *Nous avons noté*

$$\widetilde{N}(dt,dz) = N(dt,dz) - \nu(dz)dt \tag{7.6}$$

la mesure aléatoire de Poisson compensée *de* $(X_t)_{t\geq 0}$. *Pour tout ensemble* $A \in \boldsymbol{B}_0$ *le processus*

$$M_t = \widetilde{N}(t, A) \quad \textit{est une} \text{ martingale.} \tag{7.7}$$

Si $\alpha = 0$ *et* $R = \infty$, *alors* X_t *est une* martingale de Lévy.

Preuve Voir, par exemple, [Kal02]. ■

Proposition 7.3 *Si* $\mathbb{E}\left[X_t^2\right] < \infty$ *pour tout* $t \geq 0$, *nous pouvons choisir* $R = \infty$ *et écrire*

$$X_t = \alpha t + \sigma W_t + \int_{\mathbb{R}^d} z\widetilde{N}(t, dz),$$

sinon nous posons $R = 1$. *Notons, pour simplifier,*

$$\widetilde{N}(ds, dz) = \begin{cases} N(ds, dz) - \nu(dz)ds & si \;\; z \in B_R, \\ N(ds, dz) & si \;\; z \notin B_R. \end{cases} \tag{7.8}$$

Remarque 7.2 *[Pro04a]. Un processus de Lévy est un processus de Markov fort.*

Théorème 7.4 (Représentation de Lévy Khintchine) *Soit* $(X_t)_{t\geq 0}$ *un processus de Lévy sur* $\mathbb{R}^d$ *muni du triplet caractéristique* (Ξ, ν, α), *où* $\Xi = \Sigma\Sigma^T$. *Alors*

$$\psi(\omega) = \imath \langle \alpha, \omega \rangle - \frac{1}{2}\omega\Xi\omega + \int_{\mathbb{R}^d} \left(\exp\left(\imath \langle \omega, z \rangle - 1 - \imath \langle \omega, z \rangle 1_{|z|<R}\right)\nu(dz)\right), \quad \forall \omega \in \mathbb{R}^d. \tag{7.9}$$

Preuve La démonstration se déduit de la décomposition de Lévy (7.5) ,voir [CT03]. ■

7.1.2 Processus de Lévy d'activité finie

Un processus de Lévy à saut pur, *i.e.* $\sigma = 0$, est à activité finie si

$$\int_{\mathbb{R}^d} \nu(dz) = \lambda < \infty. \tag{7.10}$$

Concrètement, un processus de Lévy d'activité finie réalise un nombre fini de sauts dans tout intervalle de temps fini. L'exemple le plus classique de processus d'activité finie est le *processus de Poisson composé* introduit par Merton [Mer76]. Pour ce processus, l'intégrale (7.10) représente l'intensité du processus de Poisson. Conditionnellement au fait qu'un saut se produise, le modèle de Merton suppose que l'amplitude du saut suit une loi normale de moyenne μ et de variance η^2. La mesure de Lévy du processus de Merton est donnée par

$$\nu(dz) = \frac{\lambda}{\sqrt{2\pi\eta^2}} \exp\left(-\frac{(z-\mu)^2}{2\eta^2}\right) dz. \tag{7.11}$$

Bien entendu, il est possible de choisir n'importe quelle distribution $F(z)$ pour la taille des sauts. Dans le cadre de processus de Poisson, la mesure de Lévy est donnée par

$$\nu(dz) = \lambda dF(z). \tag{7.12}$$

En reprenant la représentation de Lévy Khintchine du théorème 7.4 et sans tenir compte de la correction de drift,

$$\psi(\omega) = \int_{\mathbb{R}^d} (\exp(\imath \langle \omega, x \rangle) - 1) \lambda dF(x) = \lambda (\Phi_J(\omega) - 1), \tag{7.13}$$

où Φ_J est la fonction caractéristique de la distribution de saut

$$\Phi_J(\omega) = \int_{\mathbb{R}^d} \exp(\imath \langle \omega, x \rangle) dF(x).$$

7.1.3 Extension au modèle à volatilité stochastique

Dans cette section, nous présentons quelques extensions des processus de Lévy au cas de modèles à volatilité stochastique. Bates [Bat00] propose une extension du modèle de Heston [Hes93]. Dans ce cas le processus de saut est indépendant du processus de volatilité. Pour obtenir des estimateurs robustes des paramètres historiques des sauts, il peut être judicieux de remettre en question cette hypothèse d'indépendance entre le processus de saut et le processus de volatilité. Nous considérons deux approches. La première consiste à faire porter la volatilité stochastique sur l'intensité de saut. Grossièrement, lorsque le niveau de volatilité est élevé, l'événement « le sous-jacent saute » a une probabilité plus importante de se réaliser. Nous verrons deux exemples de cette approche et certaines de ses limites. La seconde approche consiste à faire porter la volatilité stochastique sur la taille du saut. Lorsque les niveaux de volatilité sont élevés, les amplitudes de saut du sous-jacent sont plus grandes. Nous retiendrons cette dernière approche.

7.1.3.1 Extension du modèle de Bates

Le premier exemple d'un modèle où l'intensité du processus à saut dépend de la volatilité est donné par le modèle présenté par Pan [Pan02]. Ce modèle correspond à une extension du modèle de Bates. Dans ce modèle à volatilité stochastique, l'intensité du processus de Poisson est proportionnelle au niveau d'un processus de Cox, Ingersoll et Ross (CIR) qui est aussi la volatilité dans la dynamique du prix du sous-jacent.

$$\begin{aligned} dS_t &= (r - \lambda\pi V_t) S_t dt + \sqrt{V_t} S_t dW_t^1 + J S_t dN_t \\ dV_t &= (\theta - \kappa V_t) dt + \beta \sqrt{V_t} dW_t^2 \\ \left\langle dW_t^1, dW_t^2 \right\rangle &= \rho dt, \end{aligned} \tag{7.14}$$

où $1+J$ est une distribution lognormale de moyenne μ et de variance η^2, N_t est un processus de Poisson d'intensité λV_t et la correction de drift π est donnée par

$$\pi = \exp\left(\mu - \frac{\sigma^2}{2}\right) - 1.$$

7.1.3.2 Modèle BNS

Un autre modèle ayant une structure semblable à celle d'Heston est celui proposé par Barndorff-Nielsen et Shephard [BNS01, BNS02]. Dans celui-ci la variance $V_t = \sigma_t^2$ est un processus d'Orstein-Uhlembeck positif dirigé par un processus de Lévy.

$$\begin{aligned} S_t &= \exp(X_t), \\ dX_t &= \left(\mu + \beta\sigma_t^2\right) dt + \sigma_t dW_t + \rho dZ_t, \qquad (7.15) \\ d\sigma_t^2 &= -\lambda dt + dZ_t. \qquad (7.16) \end{aligned}$$

En pratique, $\rho < 0$, $\lambda > 0$, $(W_t)_{t\geq 0}$ est un mouvement brownien standard et $(Z_t)_{t\geq 0}$ est un processus de Lévy sans drift et il est indépendant de W. Le terme $\beta\sigma^2$ correspond à une prime de risque sur la volatilité. Le terme ρdZ_t permet de tenir compte d'un effet de levier (leverage). Cet effet, observé sur le marché action, est constaté en observant la corrélation entre les accroissements journaliers des sous-jacents et ceux de la volatilité. Lorsque le sous-jacent baisse, la volatilité augmente et réciproquement.

Dans le modèle BNS, le prix d'une option européenne peut être obtenue par une méthode de transformée de Fourier. Ce prix est fonction de la transformée de Laplace du processus de Lévy Z_t. Cependant, ce modèle ne permet pas de retrouver les « smiles » observés sur le marché des options vanilles, voir par exemple [CT03] page 490.

7.1.3.3 Modèle avec changement de temps

L'approche proposée par Pane peut être généralisée à des processus de Lévy par le concept de changement de temps [CGMY03]. Exposons rapidement cette idée qui consiste à introduire un temps stochastique défini par

$$Y_t = \int_0^t \sigma(u)du,$$

où σ, la volatilité, est un processus stochastique. Il est possible de construire à partir d'un processus de Lévy $(X_t)_{t\geq 0}$ un processus avec changement de temps Z_t défini par

$$Z_t = X_{Y_t}.$$

Le prix d'un sous-jacent est alors modélisé par

$$S_t = \exp(Z_t) = \exp(X_{Y_t}), \qquad (7.17)$$

où $(X_t)_{t\geq 0}$ est un processus de Lévy.

On a encore un effet levier ; quand la volatilité est élevée, le temps « passe plus vite » et donc la probabilité d'un événement « le spot saute » augmente.

En pratique ces modèles sont souvent présentés dans le cas où X_t est indépendant de Y_t. Ceci permet d'appliquer la méthode de calcul des prix proposée dans [CGMY03] et basée sur des techniques de FFT introduites dans [CM98a]. Dans le cas d'une dépendance entre les deux processus $(X_t)_{t\geq 0}$ et $(Y_t)_{t\geq 0}$ une solution pour le problème déterministe est proposée dans [CW02].

7.1.3.4 Modèle à volatilité stochastique avec saut

Dans le modèle présenté ici, nous considérons une autre approche. Les observations ont montré que, suivant la définition donnée ci-dessous, la fréquence des sauts ne dépend pas du niveau de volatilité du marché.

Le modèle suivant généralise les modèles de Lévy exponentiels au cas des processus à volatilité stochastique. Nous allons supposer que le prix du sous-jacent et la volatilité sont solutions de l'équation différentielle stochastique suivante : (nous notons $X_t = \left(X_t^1, Y_t^1, \ldots, Y_t^n\right)^T$, et $X_t^1 = \log S_t$)

$$dX_t = \alpha\left(t, X_t\right) dt + \Sigma(t, X_t) dW_t + \int_{\mathbb{R}^d} \gamma(t, X_{t^-}, z) \widetilde{N}(dt, dz) \tag{7.18}$$

avec

- Σ est la matrice de volatilité de taille $d \times m$ correspondant au modèle à volatilité stochastique du chapitre 5. En reprenant les notations de ce chapitre :

$$\Sigma d\left\langle W_t, W_t\right\rangle \Sigma^T = \Xi dt = \begin{pmatrix} \sigma^2 & \rho\sigma\beta_1 & \cdots & \rho\sigma\beta_n \\ \rho\sigma\beta_1 & \beta_1^2 & \cdots & \beta_1\beta_n \\ \vdots & \cdot & \ddots & \vdots \\ \rho\sigma\beta_n & \beta_1\beta_n & \cdots & \beta_n^2 \end{pmatrix} dt, \text{ avec } \sigma = f\left(Y_t^1, \ldots, Y_t^n\right).$$

- γ est l'amplitude de saut. Nous supposerons que γ est une fonction linéaire de z dont le coefficient est donné par la volatilité :

$$\gamma_i\left(t, X_{t^-}, z\right) = \Sigma_i\left(t, X_{t^-}\right) z_i, \quad 1 \leq i \leq d. \tag{7.19}$$

  Dans nos applications, nous ne considérons que des sauts sur le sous-jacent. Dans ce cas, $\gamma_i = 0$, pour $i > 1$.
- α correspond au terme de drift et dépend du choix de la probabilité sur laquelle le prix est calculé.

Nous verrons dans la deuxième partie de ce chapitre que cette dynamique est celle d'une diffusion de Lévy. Avant cela, nous donnons une écriture plus habituelle à ce modèle à volatilité stochastique en explicitant le processus de saut.

Définition 7.5 *Considérons une extension du modèle présenté au § 5.1.*

$$S_t = S_0 \exp\left(\int_0^t r(u)du + X_t^1\right), \tag{7.20}$$

$$dX_t^1 = \sigma_t dW_t + \widetilde{\eta}_t dN_t, \tag{7.21}$$

$$\sigma_t = f\left(Y_t^1, \ldots, Y_t^n\right) = \sigma_0(t) \exp\left(1/2 \sum_{i=1}^n Y_t^i\right), \tag{7.22}$$

$$\widetilde{\eta}_t = \gamma_1\left(X_t, \eta\right) = c\ \sigma_t\ \eta, \tag{7.23}$$

$$dY_t^i = -\lambda_i Y_t^i dt + \beta_i dZ_t, \tag{7.24}$$

avec $d\left\langle W_t, Z_t\right\rangle = \rho dt$, *et* η *une variable gaussienne de moyenne* μ *et de variance* ν *et* $c = \left(\frac{1}{260}\right)^{\frac{1}{2}}$ *un facteur d'échelle.* n *est le nombre de facteur et la fonction prix est une fonction de* $[0, T] \times \Omega, \quad \Omega \subset \mathbb{R}^d$ *avec* $d = n + 1$.

7.2 Diffusion de Lévy

Nous allons considérer des processus de diffusion définis comme solutions d'équations différentielles stochastiques de la forme de (7.18).

Ce paragraphe reprend les principaux résultats énoncés dans [ØS07]. Une généralisation de la représentation de Lévy Khintchine au cas des *diffusions de Lévy* est ensuite ajoutée. Elle permet de proposer, dans la partie suivante, une formulation sous forme d'équation intégro-différentielle du problème d'évaluation d'options.

7.2.1 Définition & propriétés

Définition 7.6 (Diffusion de Lévy) *Considérons l'équation différentielle stochastique de Lévy définie sur* $\mathbb{R}^d$ *:*

$$dX_s^{t,x} = \alpha\left(t, X_s^{t,x}\right) ds + \Sigma\left(t, X_s^{t,x}\right) dW_s + \int_{\mathbb{R}^d} \gamma(t, X_{s^-}^{t,x}, z)\widetilde{N}(ds, dz) \tag{7.25}$$

avec la condition de Cauchy $X_t^{t,x} = x \in \mathbb{R}^d$ *et les fonctions* $\alpha : [0,T] \times \mathbb{R}^d \to \mathbb{R}^d$, $\Sigma : [0,T] \times \mathbb{R}^d \to \mathbb{R}^{d\times m}$ *et* $\gamma : [0,T] \times \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^{d\times \ell}$ l'amplitude de saut. *Ici* W_t *est un Brownien de dimension* m *et* $\widetilde{N}(dt, dz)$ *un processus de sauts quadratique.*

Si la solution $\left(X_s^{t,x}\right)_{s\geq t}$ *de (7.25) existe, alors* $\left(X_s^{t,x}\right)_{s\geq t}$ *est une diffusion de Lévy.*

Nous notons $\Sigma d\left\langle W_t, W_t\right\rangle \Sigma^T = \Xi dt$.

Remarque 7.3 *Nous ne détaillerons pas la construction du processus à saut* $\widetilde{N}(dt, dz)$*. Notons simplement que pour tenir compte de l'effet leverage, il peut être nécessaire de construire ce processus à l'aide de copules de Lévy. Le lecteur trouvera dans [CT03] une présentation de ces copules et dans [FS06] la construction des mesures de Lévy associées.*

Théorème 7.5 (Existence et unicité de la solution de l'EDS de Lévy (7.25)) *Notons*

$$\nu_k(x) = \int_{\mathbb{R}^{d-1}} \nu\left(dz_1, \ldots, dz_{k-1}, x, dz_{k+1}, \ldots, dz_d\right). \tag{7.26}$$

En supposant que les coefficients α, Σ, γ *satisfont les conditions suivantes*

– *Croissance linéaire : il existe une constante* $C < \infty$ *telle que*

$$\|\Sigma(t,x)\|^2 + |\alpha(t,x)|^2 + \int_{\mathbb{R}} \sum_{k=1}^{\ell} |\gamma_k(t,x,z)|^2 \nu_k(dz_k) \leq C(1+|x|^2), \quad \forall x \in \mathbb{R}^d;$$

– *Régularité Lipschitz : il existe une constante* $\widetilde{C} < \infty$ *telle que*

$$\begin{aligned} &\|\Sigma(t,x) - \Sigma(t,y)\|^2 + |\alpha(t,x) - \alpha(t,y)|^2 \\ &+ \sum_{k=1}^{\ell} \int_{\mathbb{R}} |\gamma_k(t,x,z_k) - \gamma_k(t,y,z_k)|^2 \nu_k(dz_k) \leq \widetilde{C}|x-y|^2, \quad \forall x, y \in \mathbb{R}^d. \end{aligned}$$

Alors il existe un unique processus $(X_t)_{t\geq 0}$ *cad-lag et adapté à la filtration* $\mathcal{F}_t$ *tel que*

$$\mathbb{E}\left[|X(t)|^2\right] < \infty \qquad \forall t \geq 0.$$

Par la suite nous notons $X_t^{0,x} = X_t$.

7.2.2 Formule d'Itô

Nous donnons à présent deux versions de la formule d'Itô : tout d'abord sur $\mathbb{R}$ puis sa généralisation sur $\mathbb{R}^d$. Le lecteur trouvera les démonstrations de ces résultats dans [ØS07]. Notons que les formules de Dynkin [ØS07] pour l'évaluation d'options américaines dans ces modèles sont un corollaire de ce théorème.

Théorème 7.6 (Formule d'Itô en dimension 1) *Supposons que $X_t \in \mathbb{R}$ est une* diffusion de Lévy *solution de (7.25) avec $d = 1$.*

Soient $u \in C^2(\mathbb{R}^2)$ et le processus $(Y_t)_{t\geq 0}$ défini par $Y_t = u(t, X_t)$. Alors Y_t est également une diffusion de Lévy *et*

$$\begin{aligned} dY_t = &\left\{ \frac{\partial u}{\partial t}(t, X_t) + \frac{\partial u}{\partial x}(t, X_t)\alpha(t, X_t) + \frac{1}{2}\Xi(t, X_t)\frac{\partial^2 u}{\partial x^2}(t, X_t) \right\} dt \\ &+ \int_{|z|<R} \left\{ u(t, X_{t^-} + \gamma(t, X_{t^-}, z)) - u(t, X_{t^-}) - \frac{\partial u}{\partial x}(t, X_{t^-})\gamma(t, X_{t^-}), z) \right\} \nu(dz)dt \\ &+ \frac{\partial u}{\partial x}(t, X_t)\Sigma(t, X_t)dW_t + \int_{\mathbb{R}} \{u(t, X_{t^-} + \gamma(t, X_{t^-}, z)) - u(t, X_{t^-})\} \widetilde{N}(dt, dz). \quad (7.27) \end{aligned}$$

Dans le cas multi-dimensionel, la formule devient :

Théorème 7.7 (Formule d'Itô multi-dimensionnelle) *Soit $X_t \in \mathbb{R}^d$ une* diffusion de Lévy *solution de (7.25).*

Soient $u \in C^{1,2}([0, T]) \times \mathbb{R}^d; \mathbb{R})$ et $Y_t = u(t, X_t)$. Alors

$$\begin{aligned} dY_t = &\left\{ \frac{\partial u}{\partial t}(t, X_t) + \sum_{i=1}^{n} \frac{\partial u}{\partial x_i}(t, X_t)\alpha_i(t, X_t) + \frac{1}{2}\sum_{i,j=1}^{n} \Xi_{i,j}(t, X_t)\frac{\partial^2 u}{\partial x_i \partial x_j}(t, X_t) \right\} dt \\ &+ \sum_{k=1}^{l} \int_{|z|<R} u(t, X_{t^-} + \gamma_k(t, X_{t^-}, z_k)) - u(t, X_{t^-}) - \frac{\partial u}{\partial x}(t, X_{t^-})\gamma_k(t, X_{t^-}, z_k)\nu(dz)dt \\ &+ \sum_{i=1}^{n} \frac{\partial u}{\partial x_i}(t, X_t)\sigma_i(t, X_t)dW_t + \sum_{k=1}^{l} \int_{\mathbb{R}} u\left(t, X_{t^-} + \gamma_k(t, X_{t^-}), z_k - u\left(t, X_{t^-}\right)\right) \widetilde{N}(dt, dz). \end{aligned} \quad (7.28)$$

7.3 Équation de valorisation

Considérons à nouveau le modèle donné à la définition 7.5 avec $X_t = \left(X_t^1, Y_t^1, \ldots, Y_t^n\right)$. Nous ne sommes pas dans le cadre d'application du théorème 7.5. En particulier il est connu [Jou04] que dans le cas du modèle de Scott ($n = 1$), nous ne disposons pas du résultat suivant :

$$\mathbb{E}\left[|X(t)|^2\right] < \infty \qquad \forall t \geq 0. \quad (7.29)$$

Afin de calculer le prix d'une option européenne de maturité T et de payoff $h\left(S_T\right)$, nous allons supposer que dans (7.20),

$$\sigma_t = f\left(Y_t^1, \ldots, Y_t^n\right) = \max\left(\sigma_0(t)\exp\left(1/2\sum_{i=1}^{n} Y_t^i\right), \kappa\right).$$

La volatilité et l'amplitude de saut sont donc bornées et nous pouvons appliquer le théorème 7.5.

Nous supposons que le mesure de Lévy vérifie :

$$\int_{\mathbb{R}} e^{2\kappa z}\nu(dz) \leq \infty. \tag{7.30}$$

Il existe une probabilité $\mathbb{Q}$, équivalente à la probabilité historique, sous laquelle les prix actualisés de tous les produits financiers sont des $\mathbb{Q}$-martingales si :

$$\mathbb{E}\left[\exp\left(\int_0^T \sigma_t^2 dt\right)\right] < \infty, \tag{7.31}$$

$$\mathbb{E}\left[\exp\left(\int_0^T \int_{\mathbb{R}} \gamma_1(X_t, z)^2 \ \nu(dz)dt\right)\right] < \infty, \tag{7.32}$$

avec $\gamma_1(X_t, z) = \sigma_t z$. (7.31) et (7.32) sont une conséquence de $\sigma_t \leq \kappa$ et de (7.30). Sous la probabilité $\mathbb{Q}$, le sous-jacent actualisé $\exp\left(-\int_0^t r(u)du\right) S_t = \exp\left(X_t^1\right) = \widehat{S}_t$ est une martingale.

L'existence de la probabilité $\mathbb{Q}$ est obtenue en suivant ce qui est proposé dans [LL97] page 139.

Nous supposons également que la fonction payoff est Lipschitzienne :

$$|h(x) - h(y)| \leq c\,|x - y|\,, \tag{7.33}$$

avec $c > 0$. Cette condition est vérifiée dans le cas d'un call ou d'un put avec $c = 1$.

Proposition 7.8 *Le prix d'une option européenne est définie comme l'espérance conditionnelle du payoff $h(S_T)$ actualisé sous la probabilité risque neutre $\mathbb{Q}$.*

$$P_t = \mathbb{E}\left[\exp\left(-\int_t^T r(u)du\right) h(S_T)\middle|\, \mathcal{F}_t\right].$$

La propriété de Markov, implique

$$P\left(t, S, Y^1, \ldots, Y^n\right) = \mathbb{E}\left[\exp\left(-\int_t^T r(u)du\right) h(S_T)\middle|\, S_t = S, Y_t^1 = Y_1, \ldots Y_t^n = Y_n\right].$$

Les dynamiques des processus S_t et Y_t^i sont données par la définition 7.5. En tenant compte de la remarque 5.3, $P\left(t, S, Y^1, \ldots, Y^n\right)$ vérifie l'équation intégrodifférentielle rétrograde

$$\begin{array}{rcll} \dfrac{\partial P}{\partial t} + \mathcal{L}_D P + \mathcal{L}_J P & = & 0, & (t, s, y) \in [0, T) \times \mathbb{R}^+ \times \mathbb{R}^n, \\ P(s, y, T) & = & h(s), & (s, y) \in \mathbb{R}^+ \times \mathbb{R}^n, \end{array} \tag{7.34}$$

où T est la maturité de l'option et h la fonction payoff et

$$\mathcal{L}_D P = \frac{1}{2} f(y)^2 s^2 \frac{\partial^2 P}{\partial s^2} + \rho \sum_{i=1}^{n} \beta_i s f(y) \frac{\partial^2 P}{\partial s \partial y_i} + \frac{1}{2} \sum_{i,j=1}^{n} \beta_i \beta_j \frac{\partial^2 P}{\partial y_i \partial y_j} + r(t) \left(s \frac{\partial P}{\partial s} - P \right) - \sum_{i=1}^{n} \lambda_i y_i \frac{\partial P}{\partial y_i},$$

$$\mathcal{L}_J P = \int_{\mathbb{R}} \left(P \left(s e^{\gamma_1(s,y,z)} \right) - P(s) - \left(e^{\gamma_1(s,y,z)} - 1 \right) \frac{\partial P}{\partial s} \right) \nu(dz).$$

Éléments de preuve Nous nous plaçons ici dans le cas d'un processus d'activité finie, ce qui nous permet d'imposer $R = \infty$.

Cette hypothèse implique (voir le lemme B.4) que

$$\alpha_1(x) = -\frac{f(y)^2}{2} - \int_{\mathbb{R}} \left(e^{\gamma_1(x,z)} - 1 - \gamma_1(x,z) \right) \nu(dz). \tag{7.35}$$

Sous la probabilité $\mathbb{Q}$, le générateur infinitésimal $\mathcal{L}_x$ devient

$$\begin{aligned} \mathcal{L}_x u = {} & \frac{f(y)^2}{2} \left(\frac{\partial^2 u}{\partial x_1^2} - \frac{\partial u}{\partial x_1} \right) \\ & + \int_{\mathbb{R}} \left[u(x_1 + \gamma_1(x,z), y_1, \ldots, y_n) - u(x) - \left(e^{\gamma_1(x,z)} - 1 \right) \frac{\partial u}{\partial x_1} \right] \nu(dz) \\ & + \rho \sum_{i=1}^{n} \beta_i f(y) \frac{\partial^2 u}{\partial x_1 \partial y_i} + \frac{1}{2} \sum_{i,j=1}^{n} \beta_i \beta_j \frac{\partial^2 u}{\partial y_i \partial y_j} - \sum_{i=1}^{n} \lambda_i y_i \frac{\partial u}{\partial y_i}. \end{aligned} \tag{7.36}$$

Nous en déduisons le générateur de $\left(S_t, Y_t^1, \ldots, Y_t^n\right) = (S_t, Y_t)$

$$\begin{aligned} \mathcal{L}_s u = {} & \frac{f(y)^2}{2} s^2 \frac{\partial^2 u_1}{\partial s^2} + \int_{\mathbb{R}} \left[u \left(s e^{\gamma_1(s,y,z)}, y \right) - u(s,y) - \left(e^{\gamma_1(s,y,z)} - 1 \right) s \frac{\partial u}{\partial s} \right] \nu(dz) \\ & + r s \frac{\partial u}{\partial s} + \rho \sum_{i=1}^{n} s \beta_i f(y) \frac{\partial^2 u}{\partial s \partial y_i} + \frac{1}{2} \sum_{i,j=1}^{n} \beta_i \beta_j \frac{\partial^2 u}{\partial y_i \partial y_j} - \sum_{i=1}^{n} \lambda_i y_i \frac{\partial u}{\partial y_i}. \end{aligned} \tag{7.37}$$

Le prix actualisé

$$\widehat{P}\left(t, S_t, Y_t^1, \ldots, Y_t^n\right) = \exp\left(\int_t^T r(u) du \right) P\left(t, S_t, Y_t^1, \ldots, Y_t^n\right),$$

est une martingale, sous la probabilité $\mathbb{Q}$. La décomposition suivante est obtenue en appliquant la formule d'Itô du théorème 7.7 à $\widehat{P}$,

$$d\widehat{P} = a(t) dt + dM_t^C + dM_t^J, \tag{7.38}$$

avec

$$a(t) = \frac{\partial \widehat{P}}{\partial t} + \mathcal{L}_s \widehat{P}, \tag{7.39}$$

$$dM_t^C = \sum_{i=1}^{d} \frac{\partial \widehat{P}}{\partial x_i}(t, X_{t^-}^i) \Sigma(t, X_t) dW_t, \tag{7.40}$$

$$dM_t^J = \int_{\mathbb{R}} \left\{ \widehat{P}\left(t, S_{t^-} e^{\gamma_1(X_{t^-}, z)} \right) - \widehat{P}(t, S_{t^-}, Y_t) \right\} \widetilde{N}(dt, dz). \tag{7.41}$$

Nous allons montrer que sous les hypothèses de la proposition 7.8, M_t^C et M_t^J sont des martingales de carré intégrables. Ceci implique que le processus $a(t) = 0$ $\mathbb{Q}$-presque sûrement, pour tout $t \in [0, T)$ et donc que P vérifie l'équation (7.34).

La démonstration reprend ce qui est proposé dans [Vol05].

Lemme 7.9 *Le prix actualisé $\widehat{P}$ est une fonction Lipschitzienne de S.*

Ce lemme permet de montrer que :

$$\begin{aligned}\mathbb{E}\left[\left\langle M_t^J, M_t^J\right\rangle\right] =& \mathbb{E}\left[\int_0^T \int_{\mathbb{R}} \left(\widehat{P}\left(t, S_{t^-} e^{\gamma_1(Y_t,z)}\right) - \widehat{P}\left(t, S_{t^-}\right)\right)^2 \nu(dz)dt\right] \\ &\lesssim \mathbb{E}\left[\int_0^T \int_{\mathbb{R}} S_{t^-}^2 \left(e^{\gamma_1(Y_t,z)} - 1\right)^2 \nu(dz)dt\right] < \infty.\end{aligned}$$

Ce résultat est une conséquence de (7.30) et implique que M_t^J est une martingale de carré intégrable.

Le lemme 7.9 implique également :

$$\mathbb{E}\left[\int_0^T S_{t^-}^2 \sigma_t^2 \left|\frac{\partial \widehat{P}}{\partial s}\right|^2 dt\right] \leq C^2\kappa^2 \mathbb{E}\left[\int_0^T S_{t^-}^2 dt\right] < \infty. \tag{7.42}$$

Lemme 7.10 *Le prix actualisé $\widehat{P}$ est de régularité Lipschitz par rapport à y_k avec une constante qui dépend linéairement de S :*

$$\left|\widehat{P}\left(t, S, y_1^1, \ldots, y_{k-1}^1, y_k^1, y_{k+1}^1, \ldots, y_d^1\right) - \widehat{P}\left(t, S, y_1^1, \ldots, , y_{k-1}^1, y_k^2, , y_{k+1}^1, \ldots, y_n^1\right)\right| \leq cS\left|y_k^1 - y_k^2\right|.$$

Ce lemme permet de montrer que

$$\mathbb{E}\left[\int_0^T \left|\frac{\partial \widehat{P}}{\partial y_i}\right|^2 dt\right] \leq c^2 \mathbb{E}\left[\int_0^T S_{t^-}^2 dt\right] < \infty. \tag{7.43}$$

Le résultat précédent et (7.42) permettent de conclure que M_t^C est une martingale de carré intégrable, en effet :

$$\mathbb{E}\left[\left\langle M_t^C, M_t^C\right\rangle\right] = \mathbb{E}\left[\int_0^T \left(\sum_{i=1}^d \frac{\partial \widehat{P}}{\partial x_i}(t, X_{t^-}^i)\Sigma(t, X_t)\right)^2 dt\right] < \infty. \tag{7.44}$$

■

preuve du lemme 7.9

$$\begin{aligned}\left|\widehat{P}\left(t, S_1, Y_1, \ldots, Y_d\right) - \widehat{P}\left(t, S_2, Y_1, \ldots, Y_d\right)\right| =& \left|\mathbb{E}\left[h\left(S_1 \exp\left(\int_t^T r(u)du + X_{T-t}^1\right)\right)\right]\right. \\ &\left. -\mathbb{E}\left[h\left(S_2 \exp\left(\int_t^T r(u)du + X_{T-t}^1\right)\right)\right]\right| \\ \leq& c\left|S_1 - S_2\right| \mathbb{E}\left[\exp\left(X_{T-t}^1\right)\right] \\ \leq& C\left|S_1 - S_2\right|. \end{aligned} \tag{7.45}$$

■

preuve du lemme 7.10 Nous allons donner les éléments de démonstration de ce lemme dans le cas sans saut, nous admettrons la généralisation de ce résultat. Le lecteur intéressé trouvera dans [Pro04b] les éléments permettant d'obtenir la démonstration dans le cas d'un processus à saut. Afin de simplifier l'exposé on se ramène au cas d'un seul facteur. Nous notons $Z_T^{t,s,y} = \left(S_T^{t,s,y}, Y_T^{t,s,y}\right)$ le flot du processus.

$$\begin{aligned}\left|\widehat{P}\left(t,s,y^1\right) - \widehat{P}\left(t,s,y^2\right)\right| =& \left|\mathbb{E}\left[h\left(S_T^{t,s,y_1}\right)\right] - \mathbb{E}\left[h\left(S_T^{t,s,y_2}\right)\right]\right| \\ \leq& c\mathbb{E}\left[\left|S_T^{t,s,y_1} - S_T^{t,s,y_2}\right|\right] \\ \leq& c\left(\mathbb{E}\left[\left(S_T^{t,s,y_1} - S_T^{t,s,y_2}\right)^2\right]\right)^{1/2}\end{aligned}$$

$$\begin{aligned}\mathbb{E}\left[\left(S_T^{t,s,y_1} - S_T^{t,s,y_2}\right)^2\right] \leq& \mathbb{E}\left[\int_t^T \left(S_u^{t,s,y_1} f\left(Y_u^{t,s,y_1}\right) - S_u^{t,s,y_2} f\left(Y_u^{t,s,y_2}\right)\right)^2 du\right] \\ \leq& 2\mathbb{E}\left[\int_t^T \left(S_u^{t,s,y_1} - S_u^{t,s,y_2}\right)^2 f\left(Y_u^{t,s,y_1}\right)^2 du\right] \\ &+ 2\mathbb{E}\left[\int_t^T {S_u^{t,s,y_2}}^2 \left(f\left(Y_u^{t,s,y_1}\right) - f\left(Y_u^{t,s,y_2}\right)\right)^2 du\right] \\ \leq& 2\kappa^2\mathbb{E}\left[\int_t^T \left(S_u^{t,s,y_1} - S_u^{t,s,y_1}\right)^2\right] \\ &+ 2\mathbb{E}\left[\int_t^T {S_u^{t,s,y_2}}^2 \left(f\left(Y_u^{t,s,y_1}\right) - f\left(Y_u^{t,s,y_2}\right)\right)^2 du\right] \qquad (7.46)\end{aligned}$$

La régularité Lipschitz locale de f implique

$$\begin{aligned}\mathbb{E}\left[\left(S_T^{t,s,y_1} - S_T^{t,s,y_2}\right)^2\right] \leq& 2\kappa^2\mathbb{E}\left[\int_t^T \left(S_u^{t,s,y_1} - S_u^{t,s,y_2}\right)^2\right] \\ &+ 2c^2\mathbb{E}\left[\int_t^T {S_u^{t,s,y_2}}^2 \left|Y_u^{t,s,y_1} - Y_u^{t,s,y_2}\right|^2 du\right] \qquad (7.47)\end{aligned}$$

Y_t est un processus d'Ornstein-Uhlenbeck ce qui implique $\mathbb{E}\left[\left(Y_u^{t,s,y_1} - Y_u^{t,s,y_2}\right)^2\right] \leq c_1\left(y_1 - y_2\right)^2$ et

$$\mathbb{E}\left[\left(S_T^{t,s,y_1} - S_T^{t,s,y_2}\right)^2\right] \lesssim \kappa^2\mathbb{E}\left[\int_t^T \left(S_u^{t,s,y_1} - S_u^{t,s,y_2}\right)^2 du\right] + s^2\left|y_1 - y_2\right|^2 \qquad (7.48)$$

Le Lemme de Gronwall permet de conclure que

$$\mathbb{E}\left[\left(S_T^{t,s,y_1} - S_T^{t,s,y_2}\right)^2\right] \lesssim s^2\left|y_1 - y_2\right|^2 \exp\left(\kappa^2(T-t)\right). \qquad (7.49)$$

On en déduit

$$\left|\widehat{P}\left(t,s,y_1\right) - \widehat{P}\left(t,s,y_2\right)\right| \lesssim s\left|y_1 - y_2\right|. \qquad (7.50)$$

■

7.4 Approximation numérique par différences finies sparse

Nous abordons dans ce paragraphe la résolution numérique du problème (7.34) issu du modèle donné par la définition 7.5.

7.4.1 Localisation du problème

A nouveau, nous ne résolvons pas l'équation (7.34) portant sur le prix de l'option, mais l'équation vérifiée par une surprime définie comme la différence entre le prix dans notre modèle et le prix donné par un modèle de Black & Scholes.

Proposition 7.11 *La différence de prix $\pi = (P - P_{BS})$ entre :*
- *le prix P d'une option européenne de maturité T, et dont la fonction payoff est donnée par : $h : \mathbb{R} \to \mathbb{R},\ s \to h(s)$, qui vérifie l'équation (7.34),*
- *le prix d'une option européenne P_{BS}(dont la dynamique du sous-jacent est donnée par l'équation (3.4)), de même fonction Payoff h ;*

vérifie l'équation aux dérivées partielles :

$$\frac{\partial \pi}{\partial t} - \mathcal{L}_D \pi - \mathcal{L}_J \pi = \frac{1}{2}\left(f(y)^2 - 1\right) s^2 \frac{\partial^2 P_{BS}}{\partial s^2} + \mathcal{L}_J P_{BS}, \tag{7.51}$$

sur $\mathbb{R}^+ \times \mathbb{R}^n \times (0,T)$, *avec la condition de Cauchy* $\pi(s,y,0) = 0, \quad$ *sur* $\mathbb{R}^+ \times \mathbb{R}^n$.

Soit $\Omega =]0, s_{max}[\times \prod_{i=1}^{d}]-y_i^{max}, y_i^{max}[$. Nous approchons $\pi_{|\Omega\times(0,T)}$ par $\widetilde{\pi}$ solution de

$$\begin{aligned} \frac{\partial \widetilde{\pi}}{\partial t} - \mathcal{L}_D \widetilde{\pi} - \mathcal{L}_J \widetilde{\pi} &= \frac{1}{2}\left(f(y)^2 - 1\right) s^2 \frac{\partial^2 P_{BS}}{\partial s^2} + \mathcal{L}_J P_{BS}, \\ \widetilde{\pi}(s,y,t) &= 0, \quad s > s_{max}, \end{aligned} \tag{7.52}$$

avec la condition de Cauchy $\widetilde{\pi}(s,y,0) = 0$ sur Ω.

7.4.2 Discrétisation en temps

Afin de résoudre numériquement le problème (7.52), une discrétisation en temps semi-implicite est appliquée. Cette approche consiste à utiliser le θ-schéma suivant

$$\frac{u^{n+1} - u^n}{\Delta t} = \mathcal{L}\left(\theta u^{n+1} + (1-\theta)u^n\right) + \mathcal{L}_J u^n.$$

Dans le cas de l'équation (7.52), le schéma discrétisé en temps est donné par

$$\widetilde{\pi}^{n+1} - \Delta t \theta \mathcal{L}_D \widetilde{\pi}^{n+1} = \widetilde{\pi}^n + \Delta t\,(1-\theta)\,\mathcal{L}_D \widetilde{\pi}^n + \mathcal{L}_J \widetilde{\pi}^n + \frac{1}{2}\left(f(y)^2 - 1\right) s^2 \frac{\partial^2 P_{BS}^{n+\frac{1}{2}}}{\partial s^2} + \mathcal{L}_J P_{BS}^{n+\frac{1}{2}} \tag{7.53}$$

7.4.3 Discrétisation de l'opérateur $\mathcal{L}_J$

La principale difficulté liée à la discrétisation du terme de saut $\mathcal{L}_J$ provient du fait que l'intensité des sauts dépend de $y_1, \ldots, y_n$. La difficulté est accentuée par le fait que l'opérateur n'est pas tensoriel. Les deux méthodes de discrétisation pour un opérateur intégral sont rappelées ci-dessous.

7.4.3.1 Panorama des méthodes de discrétisation des opérateurs intégraux

1. La discrétisation proposée dans le cadre d'une méthode de différence finie pour les processus de Lévy par Cont [CV05] consiste à approcher l'opérateur intégral par une formule des trapèzes avec le même pas Δx_1 que le niveau de discrétisation sur la première variable. Les auteurs proposent également une méthode de calcul de $\mathcal{L}_J u$ basée sur des techniques de FFT.

   Dans la méthode des Sparse Grid, ce pas varie en fonction du point de grille. Une méthode de quadrature-interpolation est ici préférée.

   Algorithme 7.1 (Méthode de quadrature-interpolation) *Supposons que, pour f suffisamment régulière,*

   $$\int_{\mathbb{R}} f(z)\nu\,(dz) \approx \sum_{k=1}^{N_k} \omega_k f\,(\zeta_k)\,.$$

   L'opérateur

   $$\mathcal{L}_{J_0}(P)(s) = \int_{\mathbb{R}} \left(P\left(se^z\right) - P(s)\right)\nu(dz),$$

   est approché par

   $$\mathcal{L}_{J_0}(P)(s_j) \approx \sum_{k=1}^{N_k} \omega_k \left(P\left(s_j e^{\zeta_k}\right) - P(s_j)\right) \tag{7.54}$$

   $$\approx \sum_{k=1}^{N_k} \omega_k \left(P_{i_k} \frac{s_{i_k+1} - s_j e^{\zeta_k}}{s_{i_k+1} - s_{i_k}} + P_{i_k+1} \frac{s_j e^{\zeta_k} - s_{i_k}}{s_{i_k+1} - s_{i_k}} - P_j\right),$$

   où i_k vérifie $s_{i_k} \leq s_j e^{\zeta_k} \leq s_{i_k+1}$. N_k est le nombre de points de quadrature, qui sera par la suite un paramètre de notre méthode de résolution numérique.

2. La discrétisation proposée dans le cadre d'une méthode de Galerkin pour les processus de Lévy par Schwab [MvPS04], consiste à calculer les coefficients de la matrice de rigidité associée à l'opérateur $\mathcal{L}_J$. Ce calcul peut être abordé de front, ou par Fourier dans le cas où la transformée de Fourier des fonctions de base est connue. Pour appliquer cette méthode basée sur l'égalité de Parseval, le symbole de Fourier q ne doit pas dépendre de x. L'approximation du coefficient de la matrice de rigidité par une méthode de quadrature peut s'avérer être inefficace ou très difficile à mettre en oeuvre. Dans le cas de base d'ondelettes la taille du support des fonctions de bases n'est pas constante.

Nous utilisons la méthode de collocation présentée au § 2.5.2 et l'algorithme 7.1 pour l'approximation de l'opérateur intégral.

7.4.3.2 Choix des méthodes numériques en fonction des opérateurs intégraux

Prenons quelques exemples issus de modèles réels et précisons le type de d'opérateurs discrets qu'il est possible d'utiliser.

1. Si la taille des sauts γ ne dépend pas de s et y, les contributions des sauts correspondent à un processus de Lévy. Le symbole de Fourier ne dépend pas de x. Il est alors possible d'appliquer toutes les méthodes précédemment citées.

2. Si γ ne dépend que de s, la méthode de quadrature 7.1 peut être utilisée. L'opérateur de saut porte sur une seule variable.
3. Dans le cas du modèle de Bâtes (7.14), l'opérateur intégral est de la forme :

$$\mathcal{L}_J(P) = v \int_{\mathbb{R}} \left[P\left(se^z, v\right) - P(s,v) - se^z \frac{\partial P}{\partial s}(s,v) \right] \nu(dz).$$

   $\mathcal{L}_J(P)$ est donc tensoriel. Dans ce cas :
   – pour la méthode de différence finie, les opérateurs sont composés ;
   – pour la méthode de Galerkin, nous effectuons le produit tensoriel des matrices associées aux opérateurs portant sur s et v, comme dans le cas d'opérateurs différentiels
4. Dans le cas de la proposition 7.11, nous devons discrétiser l'opérateur de saut par une méthode de collocation.

7.4.4 Résultats numériques

7.4.4.1 Paramètres du modèle

La dynamique du processus de volatilité est décrite par 3 facteurs. L'équation de valorisation (6.32) est une équation intégro-différentielle en dimension 4. Nous nous plaçons à taux d'intérêt nul pour simplifier l'analyse des résultats et nous considérons des valeurs de paramètres estimés à partir du marché. Les paramètres de modèles sont

$$\rho = -0.5,\ V_0 = 0.1936, \quad \lambda = (29.27, 2.46, 0.108)\,, \quad \beta = (1.26, 0.423, 0.421)\,.$$

Le processus de saut est un processus de Poisson d'intensité $\lambda_J = 1$ et dont l'amplitude de saut est donnée par une loi normale de moyenne $\mu = -5$ et de variance $\nu = 0.5$. La variance du processus X_t (logarithme du processus S_t actualisé) sachant $Y_t = y$ est donnée par

$$var(X_t) = f(y)^2 \left(1 + \lambda_J \frac{1}{260} \left(\mu^2 + \nu^2 \right) \right). \tag{7.55}$$

Avec les paramètres ci-dessus, la partie « saut » du processus représente 10% de la variance totale de X_t. Cette remarque justifie de manière succincte le schéma semi-implicite. Les résultats se détériorent nettement si cette proportion augmente.

Le fonction payoff est celle d'un call européen, et la maturité est de 1 an.

7.4.4.2 Résultat dans la configuration optimale

Nous reprenons la configuration de § 6.2.2.2. Le nombre de points de discrétisation de l'opérateur intégral est fixé égal à 5. La méthode de collocation donne de bons résultats de convergence sur le jeu de paramètres testés. Le temps de calcul augmente significativement par rapport au cas sans saut. Ceci est lié au coût de la méthode de collocation.

Nous indiquons dans le tableau 7.1 les prix calculés par la méthode de différence finie et de collocation sparse en fonction du niveau de raffinement. Le tableau 7.4.4.2 donne l'erreur e, définie en (6.33), estimée en fonction du niveau de raffinement. Il indique également le temps de calcul pour ces différents niveaux. Ce tableau est à comparer avec le tableau 6.6 dans le cas d'une équation différentielle.

La vitesse de convergence est sensiblement conforme au résultat théorique d'après la figure 7.1 qui représente sur une échelle logarithmique l'erreur de convergence en fonction du niveau de discrétisation.

Dans le tableau 7.3 figurent les erreurs aux points calculés par une résolution de l'équation de valorisation. Ce tableau montre qu'à partir du niveau 7 l'erreur se situe dans l'intervalle de confiance de la méthode de Monte-Carlo. A titre indicatif, le temps de calcul de la méthode de Monte Carlo sur la même machine est, pour 500000 tirages, de l'ordre de $5min$.

L'interprétation des résultats est difficile car nous ne disposons pas de la solution exacte. A titre d'illustration, nous voyons sur le tableau 7.4 que pour une discrétisation sparse de niveau 8 fixée, l'erreur calculée par rapport à la simulation Monte Carlo a tendance à augmenter quand le pas de temps diminue. Cependant, les erreurs sont contenues dans l'intervalle de confiance de la méthode de Monte Carlo. Il est donc difficile de voir l'ordre de la discrétisation en temps.

Pour conclure sur les exemples numériques, nous donnons les résultats dans le cas d'un processus à saut dont la variance de l'amplitude de saut est plus grande $\nu = 2$. Les résultats donnés par les tableaux 7.5 sont plus difficiles à analyser. Les erreurs restent du même ordre de grandeur, mais le tableau 7.6 ne montre pas clairement la convergence par rapport au niveau de discrétisation.

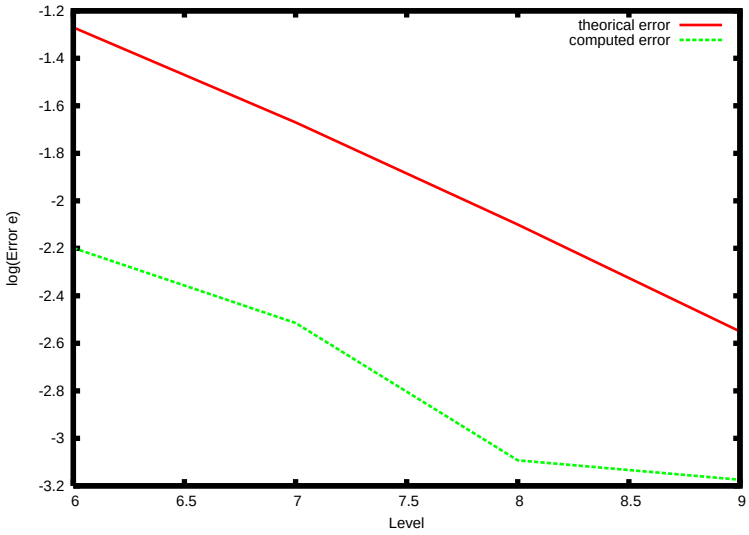

FIG. 7.1 – Erreur e en fonction du raffinement - avec un schéma implicite-explicite $\theta = 0.5$ et 100 pas de temps - pour les paramètres $\mathcal{F}_\sigma = 6, Tol = 1e-4, N_k = 5$.

TAB. 7.1 – Prix d'un call européen de maturité 1 an (multiplié par 100)- avec un schéma implicite-explicite $\theta = 0.5$ et 100 pas de temps et les paramètres $F_\sigma = 6, Tol = 1e-4, N_k = 5$.

	MC	$L6$	$L7$	$L8$	$L9$
0.8	7.082	6.994	7.051	7.058	7.088
0.81	7.492	7.366	7.423	7.496	7.493
0.82	7.913	7.750	7.875	7.915	7.911
0.83	8.347	8.286	8.270	8.345	8.359
0.84	8.792	8.693	8.749	8.788	8.803
0.85	9.250	9.110	9.167	9.244	9.258
0.86	9.719	9.695	9.671	9.711	9.725
0.87	10.200	10.136	10.190	10.190	10.205
0.88	10.692	10.587	10.642	10.682	10.696
0.89	11.196	11.050	11.186	11.185	11.199
0.9	11.712	11.690	11.659	11.699	11.734
0.91	12.238	12.174	12.227	12.225	12.260
0.92	12.775	12.669	12.722	12.762	12.797
0.93	13.324	13.175	13.313	13.310	13.345
0.94	13.882	13.865	13.829	13.869	13.904
0.95	14.451	14.392	14.442	14.438	14.473
0.96	15.029	14.928	14.979	15.018	15.031
0.97	15.618	15.475	15.613	15.608	15.621
0.98	16.216	16.209	16.170	16.208	16.221
0.99	16.823	16.776	16.824	16.818	16.831
1	17.440	17.352	17.400	17.438	17.450
1.01	18.066	17.938	17.986	18.067	18.079
1.02	18.701	18.533	18.668	18.705	18.717
1.03	19.344	19.313	19.273	19.352	19.342
1.04	19.995	19.926	19.972	20.008	19.998
1.05	20.655	20.549	20.594	20.672	20.662
1.06	21.323	21.180	21.309	21.303	21.334
1.07	21.999	21.987	21.949	21.984	21.995
1.08	22.683	22.636	22.679	22.673	22.683
1.09	23.374	23.293	23.336	23.370	23.379
1.1	24.097	23.958	24.001	24.073	24.083
1.11	24.778	24.631	24.752	24.785	24.775
1.12	25.490	25.466	25.433	25.466	25.494
1.13	26.210	26.154	26.196	26.192	26.219
1.14	26.936	26.851	26.892	26.924	26.933
1.15	27.668	27.555	27.596	27.663	27.671
1.16	28.407	28.266	28.377	28.408	28.399
1.17	29.152	29.120	29.095	29.126	29.151
1.18	29.903	29.846	29.886	29.884	29.908
1.19	30.660	30.578	30.618	30.648	30.656
1.2	31.422	31.317	31.357	31.417	31.424

TAB. 7.2 – Erreur estimée avec un schéma implicite-explicite $\theta = 0.5$ et 100 pas de temps, et - les paramètres $F_\sigma = 6, Tol = 1e-4, N_k = 5$.

	$L6$	$L7$	$L8$	$L9$
e	$6.33 \cdot 10^{-3}$	$3.06 \cdot 10^{-3}$	$8.08 \cdot 10^{-4}$	$6.69 \cdot 10^{-4}$
Tps(s)	25.96	150.6	878	5029
Tps(m)	25.96	$2min30s$	$14min39s$	$83min43s$

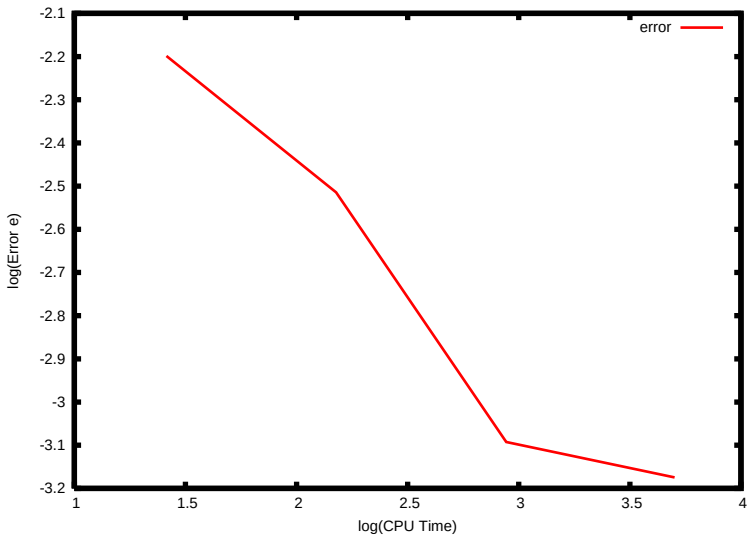

FIG. 7.2 – Erreur e en fonction du temps de calcul - avec un schéma implicite-explicite $\theta = 0.5$ et 100 pas de temps - pour le paramètre $\mathcal{F}_\sigma = 6, Tol = 1e-4, N_k = 5$.

TAB. 7.3 – Erreur aux points multipliée par $1e4$ pour un call européen de maturité 1 an - avec un schéma implicite-explicite $\theta = 0.5$ et les paramètres $F_\sigma = 6, Tol = 1e-4, N_k = 5$. L'erreur est donnée en fonction du niveau de raffinement L et du pas de temps.

L	6	7	8	8	8	9	9	Trust
N_T	100	100	100	200	300	100	200	MC
0.8	8.87	3.13	2.48	2.70	2.80	0.53	0.36	5.04
0.81	12.59	6.85	0.43	0.21	0.11	0.13	0.04	5.21
0.82	16.38	3.84	0.11	0.12	0.21	0.24	0.41	5.37
0.83	6.10	7.68	0.17	0.40	0.49	1.24	1.06	5.54
0.84	9.99	4.37	0.40	0.63	0.73	1.01	0.83	5.70
0.85	13.94	8.33	0.62	0.85	0.94	0.79	0.61	5.87
0.86	2.36	4.78	0.80	1.03	1.12	0.61	0.43	6.04
0.87	6.42	0.96	0.95	1.18	1.27	0.45	0.27	6.21
0.88	10.52	5.07	1.08	1.31	1.40	0.32	0.14	6.38
0.89	14.66	1.06	1.18	1.41	1.49	0.22	0.04	6.55
0.9	2.16	5.25	1.26	1.49	1.58	2.20	2.02	6.72
0.91	6.39	1.09	1.32	1.56	1.64	2.15	1.97	6.89
0.92	10.64	5.34	1.37	1.60	1.68	2.11	1.93	7.07
0.93	14.91	1.05	1.39	1.62	1.70	2.10	1.91	7.24
0.94	1.66	5.29	1.34	1.57	1.65	2.14	1.96	7.41
0.95	5.87	0.82	1.24	1.47	1.55	2.24	2.05	7.58
0.96	10.07	5.01	1.11	1.34	1.41	0.20	0.02	7.75
0.97	14.25	0.44	0.95	1.17	1.24	0.34	0.16	7.93
0.98	0.65	4.59	0.74	0.97	1.04	0.53	0.34	8.10
0.99	4.76	0.06	0.51	0.73	0.80	0.74	0.56	8.27
1	8.82	4.00	0.23	0.46	0.52	0.99	0.80	8.44
1.01	12.83	8.01	0.08	0.15	0.21	1.27	1.09	8.62
1.02	16.77	3.25	0.43	0.21	0.15	1.60	1.41	8.79
1.03	3.11	7.13	0.81	0.58	0.53	0.16	0.35	8.96
1.04	6.91	2.37	1.21	0.99	0.94	0.24	0.06	9.13
1.05	10.64	6.09	1.65	1.43	1.38	0.67	0.49	9.30
1.06	14.29	1.38	2.00	2.22	2.27	1.12	0.94	9.47
1.07	1.15	4.96	1.48	1.70	1.75	0.43	0.61	9.64
1.08	4.66	0.36	0.95	1.17	1.21	0.07	0.11	9.82
1.09	8.07	3.77	0.40	0.61	0.65	0.60	0.41	9.99
1.1	13.90	9.60	2.33	2.55	2.59	1.37	1.55	10.16
1.11	14.66	2.57	0.71	0.49	0.46	0.24	0.42	10.33
1.12	2.45	5.72	2.45	2.66	2.69	0.34	0.16	10.49
1.13	5.52	1.38	1.81	2.02	2.05	0.90	0.72	10.66
1.14	8.48	4.34	1.17	1.37	1.40	0.29	0.47	10.83
1.15	11.34	7.20	0.53	0.73	0.76	0.31	0.14	11.00
1.16	14.10	2.98	0.10	0.10	0.13	0.75	0.93	11.17
1.17	3.17	5.65	2.56	2.77	2.79	0.12	0.29	11.34
1.18	5.73	1.70	1.86	2.06	2.08	0.50	0.33	11.50
1.19	8.19	4.15	1.17	1.37	1.39	0.42	0.59	11.67
1.2	10.56	6.52	0.52	0.72	0.73	0.21	0.03	11.84

TAB. 7.4 – Erreur estimée avec un schéma semi-implicite avec $\theta = 0.5$ en fonction du raffinement en espace et du nombre de pas de temps - pour les paramètres $\mathcal{F}_\sigma = 6, Tol = 1e-4, N_k = 5$.

	$L8-100$	$L8-200$	$L8-300$	$L9-100$	$L9-200$
	$8.08 \cdot 10^{-4}$	$9.040 \cdot 10^{-4}$	$9.319 \cdot 10^{-4}$	$6.69 \cdot 10^{-4}$	$6.081 \cdot 10^{-4}$

TAB. 7.5 – Prix d'un call européen de maturité 1 an (multiplié par 100)- avec un schéma semi-implicite avec $\theta = 0.5$, 100 et 300 pas de temps et les paramètres $F_\sigma = 6, Tol = 1e-4, N_k = 5$.

L	MC	6	7	8	9	8
N_T	MC	100	100	100	100	300
0.80	7.145	7.061	7.117	7.123	7.153	7.118
0.81	7.556	7.433	7.489	7.565	7.561	7.560
0.82	7.980	7.817	7.944	7.986	7.980	7.980
0.83	8.416	8.360	8.340	8.418	8.431	8.413
0.84	8.864	8.767	8.821	8.863	8.876	8.857
0.85	9.323	9.185	9.239	9.320	9.333	9.314
0.86	9.794	9.776	9.747	9.789	9.802	9.783
0.87	10.277	10.217	10.269	10.270	10.283	10.264
0.88	10.771	10.668	10.721	10.762	10.775	10.757
0.89	11.277	11.130	11.268	11.267	11.280	11.261
0.90	11.794	11.777	11.741	11.783	11.817	11.777
0.91	12.321	12.261	12.312	12.310	12.344	12.304
0.92	12.860	12.756	12.807	12.848	12.883	12.843
0.93	13.409	13.261	13.401	13.398	13.432	13.392
0.94	13.969	13.957	13.917	13.958	13.992	13.952
0.95	14.539	14.484	14.533	14.529	14.563	14.523
0.96	15.119	15.020	15.069	15.110	15.122	15.104
0.97	15.710	15.567	15.706	15.701	15.713	15.696
0.98	16.308	16.306	16.262	16.303	16.314	16.297
0.99	16.917	16.872	16.919	16.914	16.925	16.908
1.00	17.535	17.448	17.495	17.534	17.545	17.529
1.01	18.161	18.034	18.080	18.164	18.175	18.159
1.02	18.797	18.629	18.765	18.803	18.813	18.798
1.03	19.441	19.413	19.369	19.451	19.439	19.445
1.04	20.093	20.026	20.070	20.108	20.096	20.102
1.05	20.753	20.649	20.692	20.773	20.761	20.767
1.06	21.421	21.280	21.409	21.404	21.434	21.398
1.07	22.097	22.090	22.049	22.086	22.094	22.080
1.08	22.780	22.738	22.780	22.775	22.784	22.770
1.09	23.471	23.395	23.436	23.472	23.480	23.467
1.10	24.170	24.060	24.101	24.177	24.185	24.171
1.11	24.875	24.733	24.853	24.888	24.877	24.883
1.12	25.587	25.570	25.534	25.569	25.596	25.564
1.13	26.307	26.258	26.298	26.296	26.321	26.290
1.14	27.033	26.955	26.995	27.029	27.035	27.023
1.15	27.765	27.659	27.698	27.768	27.774	27.762
1.16	28.504	28.370	28.480	28.513	28.502	28.507
1.17	29.249	29.225	29.198	29.231	29.254	29.226
1.18	30.000	29.950	29.989	29.990	30.011	29.984
1.19	30.757	30.682	30.721	30.753	30.759	30.748
1.20	31.519	31.421	31.460	31.522	31.528	31.517

TAB. 7.6 – Erreur estimée avec un schéma semi-implicite avec $\theta = 0.5$ en fonction du raffinement en espace et du nombre de pas de temps - pour le paramètre $\mathcal{F}_\sigma = 6, Tol = 1e-4, N_k = 5$.

$L6-100$	$L7-100$	$L8-100$	$L9-100$	$L8-300$	$L9-300$
$6.09 \cdot 10^{-3}$	$2.88 \cdot 10^{-3}$	$6.48 \cdot 10^{-4}$	$7.56 \cdot 10^{-4}$	$8.37 \cdot 10^{-4}$	$8.37 \cdot 10^{-4}$

Chapitre 8

Options multi sous-jacents

Les méthodes développées dans ce chapitre peuvent s'appliquer à toute fonction payoff bornée et appartenant à $H^\epsilon\left(\mathbb{R}_+^d\right)$, $\epsilon > 0$. Elles peuvent donc s'étendre à de nombreux modèles conduisant à la résolution d'équations aux dérivées partielles en dimension supérieure à 3. La méthode et les résultats numériques sont présentés dans le cas d'un modèle multi sous-jacents.

8.1 Modèle multi sous-jacents

Considérons d actifs risqués dont les prix au temps t sont notés $S_{i,t}$, $i = 1, \ldots, d$. Nous supposons que, pour tout i, $1 \leq i \leq d$, $S_{i,t}$ satisfait l'équation différentielle stochastique

$$dS_{i,t} = S_{i,t}\left(\mu_i dt + \sigma_i dW_{i,t}\right). \tag{8.1}$$

Ici

- $(W_{i,t})$, $1 \leq i \leq d$, sont des mouvements browniens éventuellement corrélés définis sur l'espace de probabilité $(\Omega, \mathcal{A}, \mathbb{P})$. Nous notons $\rho_{i,j}$ le facteur de corrélation entre $(W_{i,t})$ et $(W_{j,t})$. Nous avons $-1 \leq \rho_{i,j} \leq 1$ et $\rho_{i,i} = 1$.
- les volatilités σ_i, $1 \leq i \leq d$, sont positives et supposées constantes.

Le modèle de Black-Scholes suppose l'existence d'un actif non-risqué évoluant au taux sans risque r supposé constant par souci de simplification. Il peut toutefois dépendre du temps.

Nous considérons une option européenne sur un panier contenant ces d actifs risqués, de maturité T et dont la fonction payoff est notée $h\left(S_1, \ldots, S_d\right)$. Il est possible de trouver une probabilité risque neutre $\mathbb{P}^*$ sous laquelle le prix de l'option au temps t est donné par

$$P_t = e^{-r(T-t)}\mathbb{E}^*(h(S_{1,T}, \ldots, S_{d,T})|F_t). \tag{8.2}$$

8.1.1 Problèmes aux limites

En suivant ce qui est proposé dans § 4.1, il est possible de montrer que le prix de l'option est solution d'une équation aux dérivées partielles parabolique avec $1+d$ variables. L'opérateur différentiel suivant apparaît naturellement à partir de la définition 4.1.

Proposition 8.1 *Soit $\mathcal{L}$ l'opérateur différentiel :*

$$\mathcal{L} : f \to \frac{1}{2}\sum_{i=1}^{d}\sum_{j=1}^{d}\Xi_{i,j}S_iS_j\frac{\partial^2 f}{\partial S_i\partial S_j} + r\sum_{i=1}^{d}S_i\frac{\partial f}{\partial S_i}, \tag{8.3}$$

avec

$$\Xi_{i,j} = \sigma_i\sigma_j\rho_{i,j}. \tag{8.4}$$

Alors, pour toute fonction $u : (S_1 \ldots, S_d, t) \mapsto u(S_1, \ldots, S_d, t)$, $u \in \mathcal{C}^{2,1}(\mathbb{R}^d_+ \times [0,T))$ vérifiant $|S_i\frac{\partial u}{\partial S_i}| \leq C(1+|S_i|)$, $i = 1, \ldots, d$, avec C indépendant de t, le processus

$$M_t = e^{-rt}u(S_{1,t} \ldots, S_{d,t}, t) - \int_0^t e^{-r\tau}\left(\frac{\partial u}{\partial t} + \mathcal{L}u - ru\right)(S_{1,\tau}, \ldots, S_{d,\tau}, \tau)d\tau$$

est une martingale sous la filtration F_t.

Nous rapellons les résultats suivants, voir [AP05] :

Théorème 8.2 *Considérons une fonction continue $P \in \mathcal{C}^{2,1}(\mathbb{R}^d_+ \times [0,T))$, telle que $|S_i\frac{\partial P}{\partial S_i}| \leq C(1+S_i)$ avec C indépendant de t. Supposons que P satisfait*

$$\left(\frac{\partial P}{\partial t} + \mathcal{L}P - rP\right)(S_1, \ldots, S_d, t) = 0, \qquad t < T,\ (S_1, \ldots, S_d) \in \mathbb{R}^d_+ \tag{8.5}$$

et vérifie la condition de Cauchy

$$P(S_1, \ldots, S_d, T) = h(S_1, \ldots, S_d), \qquad (S_1, \ldots, S_d) \in \mathbb{R}^d_+, \tag{8.6}$$

alors le prix de l'option européenne donné par (8.2) vérifie

$$P_t = P(S_{1,t} \ldots, S_{d,t}, t). \tag{8.7}$$

8.1.2 Formulation variationnelle du problème de Cauchy (8.5)

Avant de préciser les changements à effectuer pour résoudre le problème par une méthode de Galerkin sur une base d'ondelettes sparse, nous rappelons le cadre standard d'application de la méthode de Galerkin pour le problème (8.5).

Nous supposons que la fonction payoff est de carré intégrable. Le changement de variable $T - t \to t$ permet de se ramener au problème parabolique avec condition initiale suivant :

$$\begin{array}{rl} \left(\frac{\partial P}{\partial t} - \mathcal{L}P + rP\right)(S_1, \ldots, S_d, t) = 0, & 0 < t \leq T,\ (S_1, \ldots, S_d) \in \mathbb{R}^d_+, \\ P(S_1, \ldots, S_d, 0) = h(S_1, \ldots, S_d), & (S_1, \ldots, S_d) \in \mathbb{R}^d_+, \end{array} \tag{8.8}$$

où $\mathcal{L}$ est donnée par (8.3), avec Ξ et r ne dépendent pas de t. S'ils dépendent de t, nous obtenons un problème analogue.

Supposons que les coefficients sont constants, l'opérateur $\mathcal{L}$ s'écrit sous forme de divergence :

$$\mathcal{L}u = \frac{1}{2}\sum_{i=1}^{d}\frac{\partial}{\partial S_i}\left(\sum_{j=1}^{d}\Xi_{i,j}S_iS_j\frac{\partial u}{\partial S_j}\right) + \sum_{j=1}^{d}\left(rS_j - \frac{1}{2}\sum_{i=1}^{d}\frac{\partial}{\partial S_i}\left(\Xi_{i,j}S_iS_j\right)\right)\frac{\partial u}{\partial S_j}. \tag{8.9}$$

En multipliant $-\mathcal{L}u + ru$ par une fonction test v, puis en intégrant sur $\mathcal{Q} = \mathbb{R}_+^d$ et en intégrant par partie, nous obtenons la forme bilinéaire

$$\begin{aligned} a_t(u,v) &= \frac{1}{2}\sum_{i=1}^{d}\sum_{j=1}^{d}\int_Q \Xi_{i,j}S_iS_j\frac{\partial u}{\partial S_j}\frac{\partial v}{\partial S_i} \\ &\quad -\sum_{j=1}^{d}\int_Q\left(rS_j - \frac{1}{2}\sum_{i=1}^{d}\frac{\partial}{\partial S_i}\left(\Xi_{i,j}S_iS_j\right)\right)\frac{\partial u}{\partial S_j}v + r\int_Q uv. \end{aligned} \tag{8.10}$$

Nous introduisons l'espace de Hilbert

$$\mathcal{V} = \left\{v :\ v \in L^2(Q), S_i\frac{\partial v}{\partial S_i} \in L^2(Q), i = 1,\dots,d\right\}, \tag{8.11}$$

muni de la norme

$$\|v\|_{\mathcal{V}} = \left(\|v\|^2_{L^2(Q)} + \sum_{i=1}^{d}\|S_i\frac{\partial v}{\partial S_i}\|^2_{L^2(Q)}\right)^{\frac{1}{2}}. \tag{8.12}$$

Théorème 8.3 *Si Ξ est définie positive, pour tout $h \in L^2(Q)$, il existe un unique P dans $L^2(0,T;\mathcal{V}) \cap \mathcal{C}^0([0,T];L^2(Q))$, avec $\dfrac{\partial P}{\partial t} \in L^2\left(0,T;\mathcal{V}'\right)$ tel que, pour toute fonction régulière $\phi \in \mathcal{D}(0,T)$, pour tout $v \in \mathcal{V}$,*

$$-\int_0^T \phi'(t)\left(\int_Q P(t)v\right)dt + \int_0^T \phi(t)a_t(P,v)dt = 0 \tag{8.13}$$

et

$$P(t=0) = h. \tag{8.14}$$

Cette formulation ne nous paraît pas être la plus adaptée dans le cas de problèmes en dimension supérieure à 3 pour la raison suivante : si les techniques « classiques » de localisation sont étendues au cas d'une dimension grande, le volume « intéressant » (*i.e.* le volume au voisinage de la singularité de la condition initiale) diminue exponentiellement avec la dimension.

La figure 8.1 illustre cette remarque dans le cas d'une option panier en dimension trois dont le payoff est $\left(K - \sum S_i/d\right)^+$. Sur le bord $S_2 = S_3 = 0$, la solution est obtenue en résolvant une équation en dimension 1. La localisation est choisie de sorte que $S_1^{max} = C\ K \approx C\ d\ S_1^0$. La zone « intéressante » sur la solution se situe entre les deux plans hachurés. Ce volume diminue avec la dimension.

La formulation suivante tient compte de cette remarque. Considérons le changement de variable

$$s \to x = \frac{s}{s^\star + s}, \tag{8.15}$$

où $s^\star$ est une constante positive à déterminer. Ce changement de variable ramène le domaine $\mathbb{R}_+$ sur le domaine borné $[0,1]$, et le volume « d'intérêt » de la figure 8.1 reste au centre du domaine, figure 8.2.

Ce changement de variable permet de ne pas imposer de condition sur le bord du domaine. L'équation aux dérivées partielles y dégénère et les conditions sont imposées « naturellement » par cette équation.

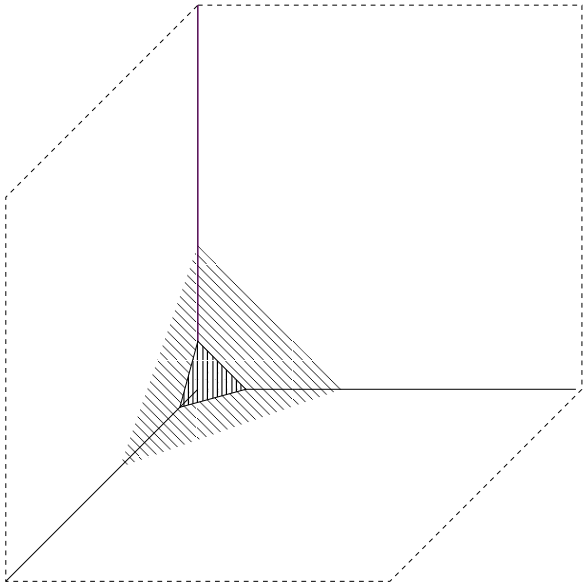

FIG. 8.1 – Volume d'intérêt pour une option basket en dimension 3 - le volume se situe entre les deux plans hachurés.

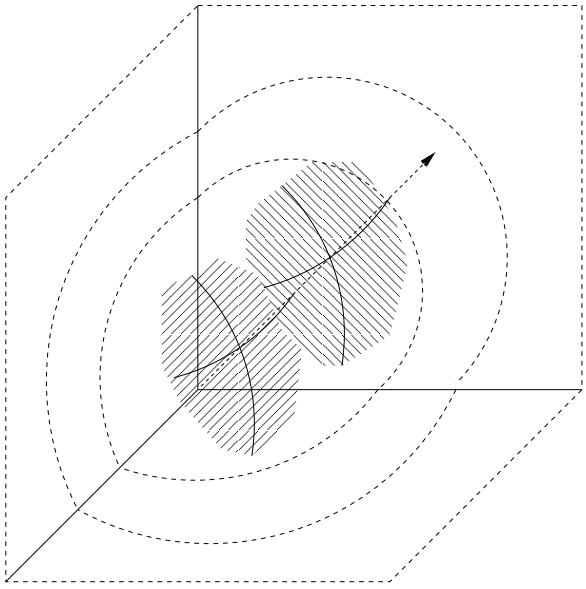

FIG. 8.2 – Volume d'intérêt dans le cas d'une option basket en dimension 3, exprimé dans les variables $x_i = \dfrac{S_i}{S_i^\star + S_i}$. En choisissant $S_i^\star = \frac{K}{d}$, ce volume est compris entre deux sphères concentriques de rayon respectif $\frac{1}{2}d^{\frac{1}{2}} - \epsilon$ et $\frac{1}{2}d^{\frac{1}{2}} + \epsilon$. Chacune de ces sphères est représentée sur le schéma par la surface hachurée et par l'intersection d'un parallèle et d'un méridien. Nous représentons également l'intersection des sphères avec les bords $x_i = 0$ du domaine.

Dans le paragraphe suivant, nous étudions un changement d'inconnue, valable pour toute fonction payoff bornée et dans $L^2\left(\mathbb{R}_+^d\right)$. La nouvelle inconnue, notée v, vérifie des conditions de Dirichlet homogènes, en particulier, $v\left(S_1,\ldots,S_d,t\right)=0$ si $S_i=0$ et $\lim_{S_i\to\infty} v=0$, pour tout t et quelque soit la valeur des autres s_k. Ce changement d'inconnue est uniquement motivé par l'utilisation d'une méthode de Sparse Grid. En effet, une fonction nulle aux bords du domaine est représentée sur une base plus petite.

8.2 Formulation adaptée aux méthodes de Sparse Grid

Le changement d'inconnue est présenté avant le changement de variables afin de simplifier les développements techniques.

8.2.1 Changement d'inconnue

La nouvelle inconnue est la fonction v définie par

$$v\left(S_1,\ldots,S_d,t\right)=S_1\ldots S_dP\left(S_1,\ldots,S_d,t\right).$$

Proposition 8.4 *Soient $\mathcal{Q}=\left(\mathbb{R}_+^\star\right)^d$, et P la solution de l'équation (8.5) alors la fonction v définie par $(S_1,\ldots,S_d,t)\to S_1\ldots S_dP\left(S_1,\ldots,S_d,t\right)$ est solution de*

$$\frac{\partial v}{\partial t}-\frac{1}{2}\sum_{i,j=1}^{d}\Xi_{i,j}S_iS_j\frac{\partial^2 v}{\partial S_i\partial S_j}-\sum_{i=1}^{d}\left(r-\sum_{j=1}^{d}\Xi_{i,j}\right)S_i\frac{\partial v}{\partial S_i} \tag{8.16}$$
$$+\left[(d+1)r-\sum_{i,j=1}^{d}\Xi_{i,j}\frac{1+\delta_i^j}{2}\right]v=0 \text{ sur } (0,T]\times\mathcal{Q}$$
$$v\left(S_1,\ldots,S_d,0\right)=S_1\ldots S_dh\left(S_1,\ldots,S_d\right) \text{ sur } \mathcal{Q}. \tag{8.17}$$

Par la suite, le produit $S_1\ldots S_d$ sera noté π et la fonction g est définie par πh.

Preuve La fonction v vérifie

$$\frac{\partial v}{\partial S_i}=\left[\frac{\partial P}{\partial S_i}+\frac{P}{S_i}\right]\pi \quad\Rightarrow\quad \frac{\partial P}{\partial S_i}=\frac{1}{\pi}\left[\frac{\partial v}{\partial S_i}-\frac{v}{S_i}\right],$$
$$\frac{\partial^2 v}{\partial S_i\partial S_j}=\left[\frac{\partial^2 P}{\partial S_i\partial S_j}+\frac{1}{S_j}\frac{\partial P}{\partial S_i}+\frac{1}{S_i}\frac{\partial P}{\partial S_j}-\left(1+\delta_i^j\right)\frac{v}{SiS_j}\right]\pi$$
$$\Rightarrow\quad \frac{\partial^2 P}{\partial S_i^2}=\frac{1}{\pi}\left[\frac{\partial^2 v}{\partial S_i\partial S_j}-\frac{1}{S_j}\frac{\partial v}{\partial S_i}-\frac{1}{S_i}\frac{\partial v}{\partial S_j}+\left(1+\delta_i^j\right)\frac{v}{S_iS_j}\right].$$

Nous en déduisons que

$$\frac{\partial v}{\partial t}-\frac{1}{2}\sum_{i,j=1}^{d}\Xi_{i,j}S_iS_j\left[\frac{\partial^2 v}{\partial S_i\partial S_j}-\frac{1}{S_j}\frac{\partial v}{\partial S_i}-\frac{1}{S_i}\frac{\partial v}{\partial S_j}+\frac{1+\delta_i^j}{S_iS_j}v\right]$$
$$-\sum_{i=1}^{d}rS_i\left[\frac{\partial v}{\partial S_i}-\frac{v}{S_i}\right]+rv=0,$$

ce qui implique (8.16). ■

8.2.2 Changement de variables

Soit $T_{s^\star}$ le changement de variable défini par :

$$T_{s^\star} : \mathbb{R}_+ \to [0,1], \qquad s \to \frac{s}{s+s^\star}, \tag{8.18}$$

$$T_{s^\star}^{-1} : [0,1] \to \mathbb{R}_+, \qquad x \to \frac{xs^\star}{1-x}, \tag{8.19}$$

où $s^\star$ est une constante positive.

Proposition 8.5 *Soient v la solution du problème de Cauchy (8.16) et $\left(T_{s_i^\star}\right)_{1\leq i\leq d}$ les transformations définies par (8.18). Soit $\widetilde{\mathcal{L}}$ L'opérateur différentiel dans les variables $x_i = T_{s_1^\star}(s_i)$, défini par*

$$\widetilde{\mathcal{L}} : f \to \frac{1}{2}\sum_{i,j=1}^{d} \Xi_{i,j}(1-x_i)x_i(1-x_j)x_j\frac{\partial^2 f}{\partial x_i \partial x_j} + \sum_{i}^{d}\left(r - \sum_{j=1}^{d}(1+\delta_i^j x_j)\Xi_{i,j}\right)x_i(1-x_i)\frac{\partial f}{\partial x_i}, \tag{8.20}$$

alors $u : (x_1,\ldots,x_d,t) \to v\left(T_{s_1^\star}^{-1}(x_1),\ldots,T_{s_d^\star}^{-1}(x_d),t\right)$ est solution du problème de Cauchy :

$$\begin{aligned} \frac{\partial u}{\partial t} - \widetilde{\mathcal{L}}u + \Upsilon u = 0 \qquad & (0,T]\times(0,1)^d \\ u(x_1,\ldots,x_d,0) = g\left(T_{s_1^\star}^{-1}(x_1),\ldots,T_{s_d^\star}^{-1}(x_d)\right) \qquad & \forall (0,1)^d, \end{aligned} \tag{8.21}$$

où $\Upsilon = r(d+1) - \sum_{i,j=1}^{d}\Xi_{i,j}\frac{1+\delta_i^j}{2}$.

Preuve Il suffit de remarquer que

$$S_i\frac{\partial v}{\partial S_i} = x_i(1-x_i)\frac{\partial u}{\partial x_i} \qquad \text{et} \qquad S_i^2\frac{\partial^2 v}{\partial S_i^2} = x_i^2(1-x_i)^2\frac{\partial u}{\partial x_i} - 2x_i^2(1-x_i)\frac{\partial u}{\partial x_i}.$$

■

Remarque 8.1 *Le problème est bien défini si la fonction, $x_1,\ldots,x_d \to g\left(T_{s_1^\star}^{-1}(x_1),\ldots,T_{s_d^\star}^{-1}(x_d)\right)$, appartient à $L^2\left(]0,1[^d\right)$. Cette condition est vérifiée pour toute fonction $g \in L^2\left(\mathbb{R}_+^d\right)$.*

Par la suite, la fonction composée $g \circ \left(T_{s_i^\star}\right)_{1\leq i\leq d}$ sera notée f.

Remarque 8.2 *En $x_i = 0$ et $x_i = 1$, l'équation dégénère et la dimension du problème diminue. Dans le cas de la dimension 1, l'équation dégénère en une équation différentielle ordinaire.*

8.2.3 Formulation variationnelle du problème de Cauchy (8.21)

L'opérateur $\widetilde{\mathcal{L}}$ défini par (8.20) s'écrit sous une forme divergence :

$$\begin{aligned}\widetilde{\mathcal{L}}u &= \frac{1}{2}\sum_{j=1}^{d}\frac{\partial}{\partial x_j}\left[\Xi_{i,j}(1-x_j)x_j\sum_{i=1}^{d}(1-x_i)x_i\frac{\partial u}{\partial x_i}\right] \\ &+\sum_{i=1}^{d}\left[r-\sum_{j=1}^{d}\Xi_{i,j}\left(1+\delta_i^j x_j+\frac{1+\delta_i^j}{2}(1-2x_j)\right)\right]x_i(1-x_i)\frac{\partial u}{\partial x_i} \\ &= \frac{1}{2}\sum_{j=1}^{d}\frac{\partial}{\partial x_j}\left[\Xi_{i,j}(1-x_j)x_j\sum_{i=1}^{d}(1-x_i)x_i\frac{\partial u}{\partial x_i}\right] \\ &+\sum_{i=1}^{d}\left[r-\frac{1}{2}\sum_{j=1}^{d}\Xi_{i,j}\left(3+\delta_i^j-2x_j\right)\right]x_i(1-x_i)\frac{\partial u}{\partial x_i}.\end{aligned} \tag{8.22}$$

En multipliant $-\widetilde{\mathcal{L}}w+\Upsilon w$ par une fonction test $v\in\mathcal{D}(\Omega)$ puis en intégrant par partie, nous obtenons la forme bilinéaire :

$$\begin{aligned}b(w,v) =& \frac{1}{2}\sum_{i,j=1}^{d}\Xi_{i,j}\int_\Omega(1-x_i)x_i\frac{\partial w}{\partial x_i}(1-x_j)x_j\frac{\partial v}{\partial x_j} \\ &-\sum_{i=1}^{d}\int_\Omega\left[r-\frac{1}{2}\sum_{j=1}^{d}\Xi_{i,j}\left(3+\delta_i^j-2x_j\right)\right]x_i(1-x_i)\frac{\partial w}{\partial x_i}v \\ &+\int_\Omega\left[r(d+1)-\sum_{i,j=1}^{d}\Xi_{i,j}\right]vw.\end{aligned} \tag{8.23}$$

Soit $\Omega=]0,1[^d$, nous introduisons l'espace de Hilbert

$$\mathcal{V}(\Omega)=\left\{v\in L^2(\Omega)\,\middle|\,\forall\ 1\le i\le d,\ (1-x_i)\,x_i\frac{\partial v}{\partial x_i}\in L^2(\Omega)\right\},$$

muni de la norme

$$\|v\|_{\mathcal{V}}=\left(\|v\|^2_{L^2(\Omega)}+\sum_{i=1}^{d}\left\|x_i(1-x_i)\frac{\partial v}{\partial x_i}\right\|^2_{L^2(\Omega)}\right)^{\frac{1}{2}}. \tag{8.24}$$

Nous pouvons montrer que l'espace $\mathcal{D}(\Omega)$ des fonctions $\mathcal{C}^\infty(\Omega)$ à support compact est dense dans $\mathcal{V}(\Omega)$.

Proposition 8.6 *Si la fonction u vérifie le problème de Cauchy (8.21) alors u est solution du problème* variationnel *:*
Trouver u telle que, pour tout v appartenant à $\mathcal{V}$ et pour presque tout t dans $]0,T]$,

$$\left\langle\frac{\partial u(t)}{\partial t},v\right\rangle_{L^2(\Omega)}+b(u(t),v)=0\quad et\quad u(x_1,\ldots,x_d,0)=f(x_1,\ldots,x_d). \tag{8.25}$$

où la forme bilinéaire b est définie de $\mathcal{V}\times\mathcal{V}\to\mathbb{R}$ par (8.23).

Rappelons que la matrice Ξ est définie positive, *i.e.* il existe une constante positive $\underline{\sigma}$ telle que, pour tout $q \in \mathbb{R}^d$,

$$\sum_{i=1}^{d}\sum_{j=1}^{d} \Xi_{i,j} q_i q_j \geq \underline{\sigma}^2 |q|^2. \tag{8.26}$$

Lemme 8.7 *La forme bilinéaire b est continue de $\mathcal{V} \times \mathcal{V}$* i.e. *il existe une constante $\overline{c}$ indépendante de t telle que, pour toute fonction $v, w \in \mathcal{V}$,*

$$b(v,w) \leq \overline{c} \|v\|_{\mathcal{V}} |w|_{\mathcal{V}}. \tag{8.27}$$

De plus, la forme bilinéaire vérifie une inégalité de Gårding : il existe deux constantes positives $\underline{c} > 0$ et $\lambda \geq 0$ telles que, pour toute fonction v dans $\mathcal{V}$,

$$b(v,v) \geq \underline{c} |v|^2_{\mathcal{V}} - \lambda \|v\|^2_{L^2(\Omega)}. \tag{8.28}$$

Preuve L'inégalité de Gårding est obtenue en remarquant que le terme d'ordre 1 est contrôlé puisqu'il s'écrit formellement

$$\left| \int_\Omega \mathcal{P}(x_1, \ldots, x_d) \frac{\partial v}{\partial x_i} v \right| \leq \frac{1}{2} \left| \int_\Omega \frac{\partial}{\partial x_i} \mathcal{P}(x_1, \ldots, x_d) v^2 \right|, \tag{8.29}$$

où $\mathcal{P}$ est un polynôme en x_i. La continuité de l'opérateur se démontre à l'aide d'arguments semblables. ■

Le théorème d'existence et d'unicité pour la formulation faible se déduit du lemme précédent.

Théorème 8.8 *Pour toute fonction $g \in L^2(\Omega)$, il existe un unique u dans $L^2(0,T;\mathcal{V}) \cap \mathcal{C}^0([0,T];L^2(\Omega))$, avec $\dfrac{\partial u}{\partial t} \in L^2\left(0,T;\mathcal{V}'\right)$ tel que, pour toute fonction régulière $\phi \in \mathcal{D}(0,T)$, pour toute fonction v dans $\mathcal{V}$,*

$$-\int_0^T \phi'(t) \int_Q u(t) v \, dt + \int_0^T \phi(t) b(u,v) dt = 0, \tag{8.30}$$

et

$$u(t=0) = f. \tag{8.31}$$

La solution de (8.30) dans le cas d'une option Put européen sur un panier de deux sous-jacents est représentée sur figure 8.3.

8.3 Résolution numérique

Considérons le schéma de discrétisation proposé par Schwab & al [PS04] pour les problèmes paraboliques linéaires :

- un schéma de Galerkin discontinu en temps (voir le § 2.6 et les références associées),
- un schéma de Galerkin sur une base obtenue par produit tensoriel sparse de bases d'ondelettes pour la discrétisation en espace (voir le § 2.4 et les références associées).

Le lecteur trouvera dans [PS04] une analyse de l'erreur de discrétisation de ce schéma dans le cas d'équations paraboliques, nous appliquons cette discrétisation à l'équation (8.25).

La répartition des pas de temps est conforme à la méthode proposée dans [PS04]. La distribution des pas de temps suit une loi de puissance définie au § 6.2.2.4 par (6.34). Le paramètre α de la discrétisation est égal au degrés de la méthode de Galerkin discontinue.

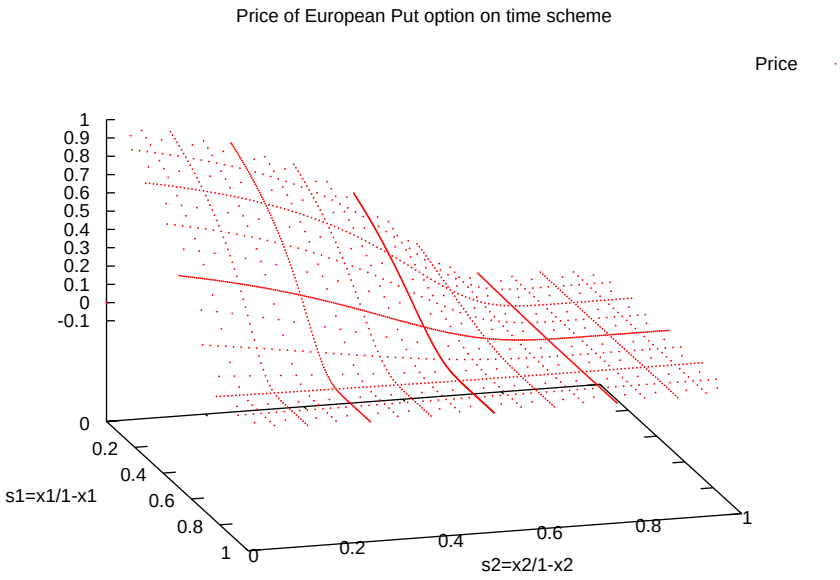

FIG. 8.3 – Prix d'une option Put européen en dimension 2 de strike $K = 1$ et de maturité $T = 1$.

8.3.1 Actifs parfaitement corrélés

Pour pouvoir comparer la solution calculée avec la solution exacte, nous nous plaçons dans le cas très spécial où tous les sous-jacents sont parfaitement corrélés, de telle sorte que le prix est donné par une formule semi-analytique. Nous testons donc notre procédure multi dimensionnelle sur un problème 1D.

8.3.1.1 Paramètres des expériences

Les expériences numériques sont réalisées sur l'évaluation d'un put européen sur un panier de d sous-jacents. Afin de simplifier l'analyse de l'erreur, nous considérons un taux d'intérêt nul, de sorte que la valeur à la monnaie est également la valeur à la monnaie forward. Les paramètres du contrat figurent dans le tableau 8.1.

TAB. 8.1 – Paramètres du put européen

Strike	$K = 100$
Poids	$\omega_i = 1/d$
Taux	$r = 0$
volatilité	$\sigma = 0.3$
Maturité	$T = 1$
Payoff	$\left(K - \sum_{i=1}^{d} \omega_i S_i\right)^+$

TAB. 8.2 – Prix à la monnaie $S_i = S_j$ en fonction de différents niveaux N_I, avec les paramètres $N_L = 7$ et $N_T = 300$.

d/N_I	1	2	3	4
7	11.898	12.157	12.282	12.094
8	11.906	12.088	11.987	11.409
9	11.907	11.933	11.712	11.269
10	11.908	11.922	11.748	11.746
11	11.908	11.893	11.799	11.800
12	11.908	11.887	11.811	11.891
13	11.908	11.885	11.863	11.938
14	11.908	11.884	11.893	11.971

8.3.1.2 Projection de la condition initiale

Ce premier paragraphe met en évidence l'importance du calcul du second membre au premier pas de temps. Ce calcul est obtenu en appliquant la méthode proposée au § 2.4.3.3. La fonction payoff est projetée sur une base obtenue par produit tensoriel sparse de bases d'ondelettes interpolantes définies au § 1.2.2.2. Nous calculons le produit scalaire de cette approximation avec les éléments de la base utilisée pour la méthode de Galerkin *i.e.* une base obtenue par produit tensoriel sparse de base d'ondelettes biorthogonales $(p, \tilde{p}) = (2, 2)$ (définie § 1.2.2.1). Dans ce qui suit, N_L sera le niveau de la base utilisée dans la méthode de Galerkin et N_I sera le niveau de la base utilisée pour l'interpolation de la fonction Payoff.

Les résultats donnés au tableau 8.2 illustrent l'importance du paramètre de discrétisation N_I. Le prix à la monnaie est de 11.9235. Ces résultats sont calculés à l'aide d'un schéma d'Euler implicite à N_T pas de temps. Une grille adaptée au payoff, obtenue dans le cas de la dimension 2, est reproduite sur la figure 8.4.

8.3.1.3 Erreur en norme $L^2\left((0,1)^d\right)$

Nous considérons divers paramètres de discrétisation sur des problèmes en dimension $(2, \ldots, 5)$. Les erreurs sur une pseudo norme L^2 sont données au tableau 8.3. L'ordre de la méthode de Galerkin discontinue pour la résolution en temps est notée DG.

Les résultats expérimentaux semblent être conformes aux prévisions théoriques. En particulier, nous observons qu'à niveau N_L fixé, l'erreur varie comme N_L^{d-1} quand la dimension d varie.

Quelques commentaires L'erreur sur ce type de produits européens se détermine par rapport à une mesure de volatilité implicite. Supposons que ce produit soit évalué à l'aide d'un sous-jacent fictif modélisant la dynamique du panier. En reprenant les arguments donnés au § 6.1.2.1, l'erreur liée à la méthode numérique est significative si elle entraine une variation de la volatilité implicite supérieure à 10^{-4}. En admettant que la volatilité du panier soit de l'ordre de 20%, l'erreur relative admissible sur la volatilité est de l'ordre de $5 \cdot 10^{-3}$. La linéarité du prix par rapport à la volatilité à la monnaie permet d'extrapoler cette erreur relative sur le prix. En conclusion, la méthode numérique est acceptable si l'erreur exprimée dans la pseudo norme L^2 est de l'ordre de 10^{-3}.

Les discrétisations données ici sont donc acceptables jusqu'à la dimension 3. Au delà,

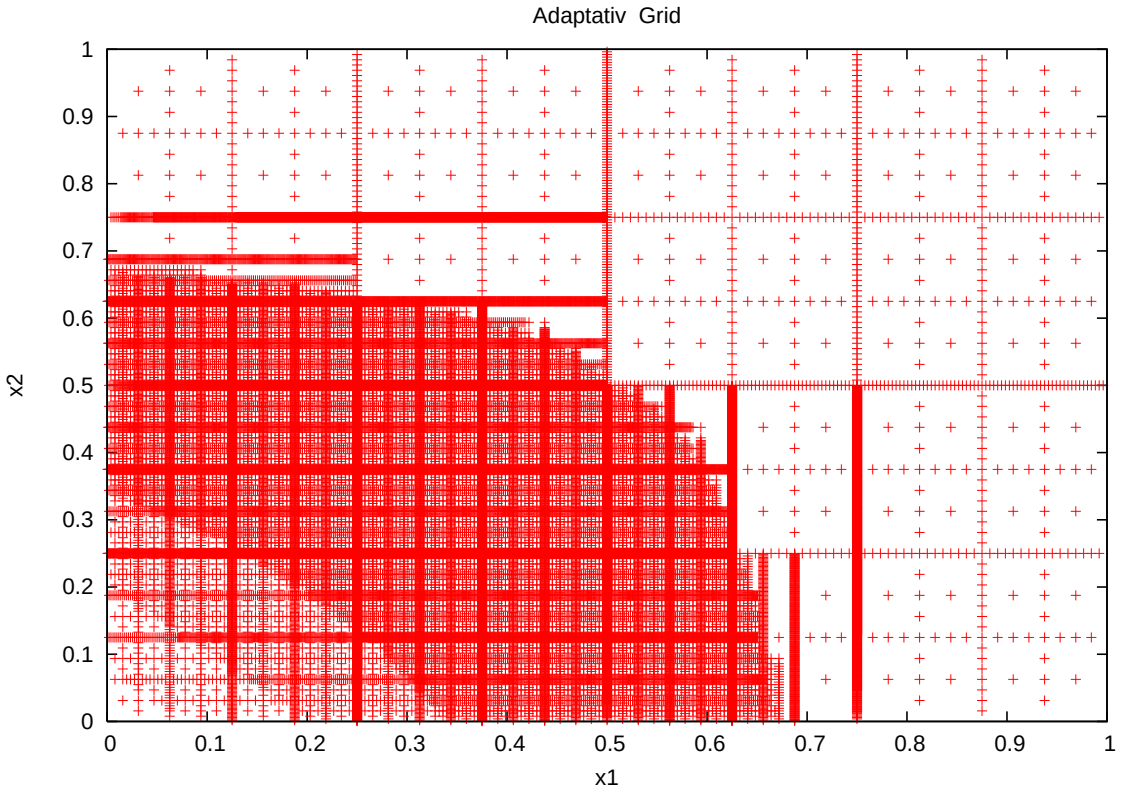

FIG. 8.4 – Grille d'interpolation pour la fonction Payoff d'un put européen, suivant les variables x_i.

TAB. 8.3 – Erreur relative par rapport à une pseudo norme L^2 calculée comme le carré des erreurs sur chacun des points de la grille sparse - avec le paramètre $N_I = 14$.

(DG, N_T)	$2^{-2n} n^{d-1}$	$(0, 100)$	$(1, 10)$	$(2, 5)$	$(2, 10)$
$Dim = 1$					
$N_L = 7$	$6.10 \cdot 10^{-5}$	$4.28 \cdot 10^{-4}$	$1.88 \cdot 10^{-4}$	$1.55 \cdot 10^{-4}$	$1.62 \cdot 10^{-4}$
$N_L = 8$	$1.52 \cdot 10^{-5}$	$4.36 \cdot 10^{-4}$	$8.54 \cdot 10^{-5}$	$1.29 \cdot 10^{-4}$	$6.12 \cdot 10^{-5}$
$N_L = 9$	$3.81 \cdot 10^{-6}$	$4.40 \cdot 10^{-4}$	$8.04 \cdot 10^{-5}$	$1.06 \cdot 10^{-4}$	$6.34 \cdot 10^{-5}$
$Dim = 2$					
$N_L = 7$	$4.27 \cdot 10^{-4}$	$1.64 \cdot 10^{-3}$	$1.25 \cdot 10^{-3}$	$1.27 \cdot 10^{-3}$	$1.29 \cdot 10^{-3}$
$N_L = 8$	$1.22 \cdot 10^{-4}$	$1.28 \cdot 10^{-3}$	$4.96 \cdot 10^{-4}$	$5.85 \cdot 10^{-4}$	$4.72 \cdot 10^{-4}$
$N_L = 9$	$3.43 \cdot 10^{-6}$	$1.23 \cdot 10^{-3}$	$2.56 \cdot 10^{-4}$	$9.30 \cdot 10^{-5}$	$1.53 \cdot 10^{-4}$
$Dim = 3$					
$N_L = 7$	$2.99 \cdot 10^{-3}$	$4.23 \cdot 10^{-3}$	$3.37 \cdot 10^{-3}$	$3.30 \cdot 10^{-3}$	$3.34 \cdot 10^{-3}$
$N_L = 8$	$9.76 \cdot 10^{-4}$	$2.97 \cdot 10^{-3}$	$1.35 \cdot 10^{-3}$	$1.38 \cdot 10^{-3}$	$1.40 \cdot 10^{-3}$
$Dim = 4$					
$N_L = 6$	$5.27 \cdot 10^{-2}$	$2.98 \cdot 10^{-2}$	$3.0 \cdot 10^{-2}$	$2.98 \cdot 10^{-2}$	$2.97 \cdot 10^{-2}$
$N_L = 7$	$2.09 \cdot 10^{-2}$	$1.40 \cdot 10^{-2}$	$1.32 \cdot 10^{-2}$	$1.32 \cdot 10^{-2}$	$1.33 \cdot 10^{-2}$
$Dim = 5$					
$N_L = 6$	$3.16 \cdot 10^{-1}$	$6.70 \cdot 10^{-2}$	$6.88 \cdot 10^{-2}$	$6.88 \cdot 10^{-2}$	$6.88 \cdot 10^{-2}$

il est nécessaire d'augmenter le niveau de discrétisation.

La tolérance du solveur itératif est fixé à 10^{-5} pour tenir compte de la remarque précédente. Les erreurs de cet ordre sont donc liées, en partie, à la méthode de résolution du système linéaire.

Le conditionnement de la matrice du système discrétisé temps et espace se dégrade rapidement avec le degrés DG de la méthode de Galerkin discontinue. Ce phénoméne explique les résultats à première vue en contradiction avec la théorie sur les discrétisations $(DG, N_T) = (2, 5)$ et $(DG, N_T) = (2, 10)$. Lorsque le schéma est convergé en espace (ce qui est le cas pour la dimension 1), le passage de 5 à 10 pas de temps permet de diminuer significativement l'erreur. A l'inverse, si le schéma n'est pas convergé en espace, le résultat peut se dégrader légèrement en augmentant de N_T. Le choix optimal de DG et N_T en fonction de N_L est discuté dans [PS04]. A condition de réduire le conditionnement du système couplé, les schémas de Galerkin discontinus permettent de réduire le temps de calcul de manière significative. La figure 8.5 (*resp.* 8.6) représente le profil de l'erreur sur l'ensemble du domaine dans le cas de la dimension 1(*resp.* 2).

Les résultats sont incomplets. Le choix de stocker les coefficients de la matrice de rigidité ne permet pas d'effectuer les expériences pour des dimensions ou des niveaux de discrétisation plus importants. Il est nécessaire d'adopter la structure sans stockage décrite au § 9.3.3. Cette méthode ne sera performante qu'à la condition de calculer rapidement le coefficient de la matrice de rigidité, en appliquant, par exemple la proposition 2.1.

Le temps de calcul peut diminuer à condition de réduire le nombre d'itérations du solveur itératif. Plusieurs pistes sont envisageables.

L'une d'elle consiste à trouver un meilleur préconditionneur pour le système qui couple la discrétisation en temps et en espace. Le nombre d'itérations augmente de manière significative avec l'ordre du schéma en temps. Nous souhaiterions conserver un nombre d'itérations proche de celui constaté pour le schéma d'Euler implicite donné par le tableau 8.4 et ainsi profiter pleinement du gain en nombre de pas de temps lié à l'utilisation de schémas de Galerkin discontinus.

La complexité du problème peut être diminuée en utilisant les techniques de compression données au § 2.4.3.5. Le gain est difficile à évaluer. En effet, il n'y a pas de résultats sur la compression de la matrice de masse. Le nombre de coefficients à calculer ne peut pas être réduit. Cependant, le terme de masse se calcule facilement puisqu'il est invariant par translation.

Enfin, la complexité diminuerait avec l'utilisation de stratégies adaptatives construites, par exemple, avec des estimateurs d'erreurs a posteriori.

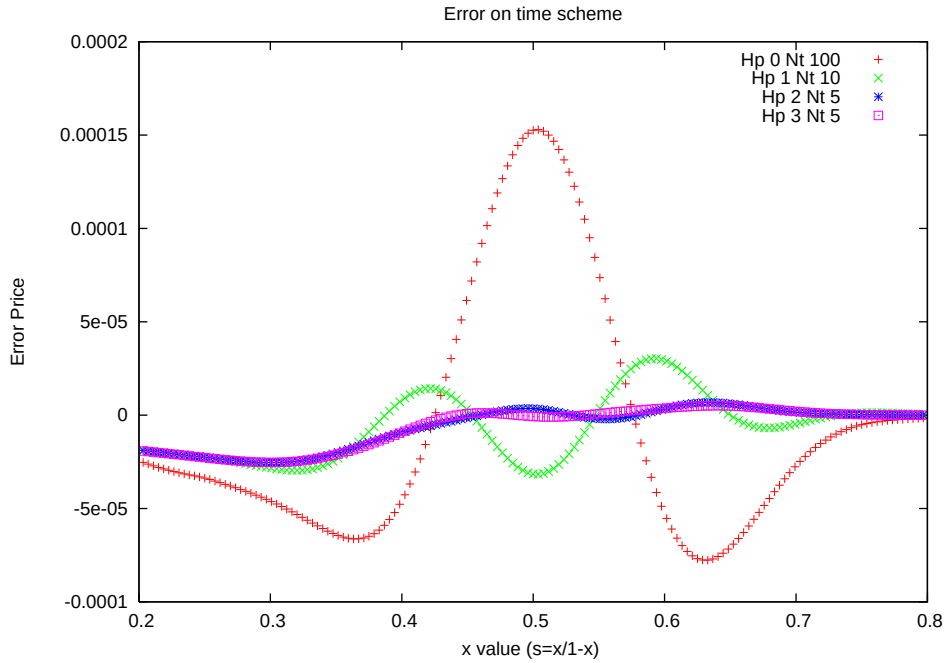

FIG. 8.5 – Erreur sur le prix d'une option Put européen de strike $K = 1$ et de maturité $T = 1$ pour différentes discrétisations. Les paramètres sont $N_L = 8, N_I = 15$ et les couples (DG, N_T) : $(0, 100)$ courbe rouge, $(1, 10)$ courbe verte, $(2, 5)$ courbe bleu et $(3, 5)$ courbe mauve.

TAB. 8.4 – Itération BICGSTAB ($\star$ indique que la structure du code en particulier le stockage de la matrice ne permet pas de mener l'expérience)

d/N_L	5	6	7	8	9	10
1	6	$6-7$	$6-7$	$6-7$	7	7
2	$9-10$	$12-13$	$15-16$	$15-17$	$16-17$	$16-18$
3	$13-14$	$17-20$	$24-25$	$26-29$	$\star$	$\star$
4	$17-18$	$23-25$	$27-29$	$\star$	$\star$	$\star$
5	$23-25$	$25-27$	$\star$	$\star$	$\star$	$\star$

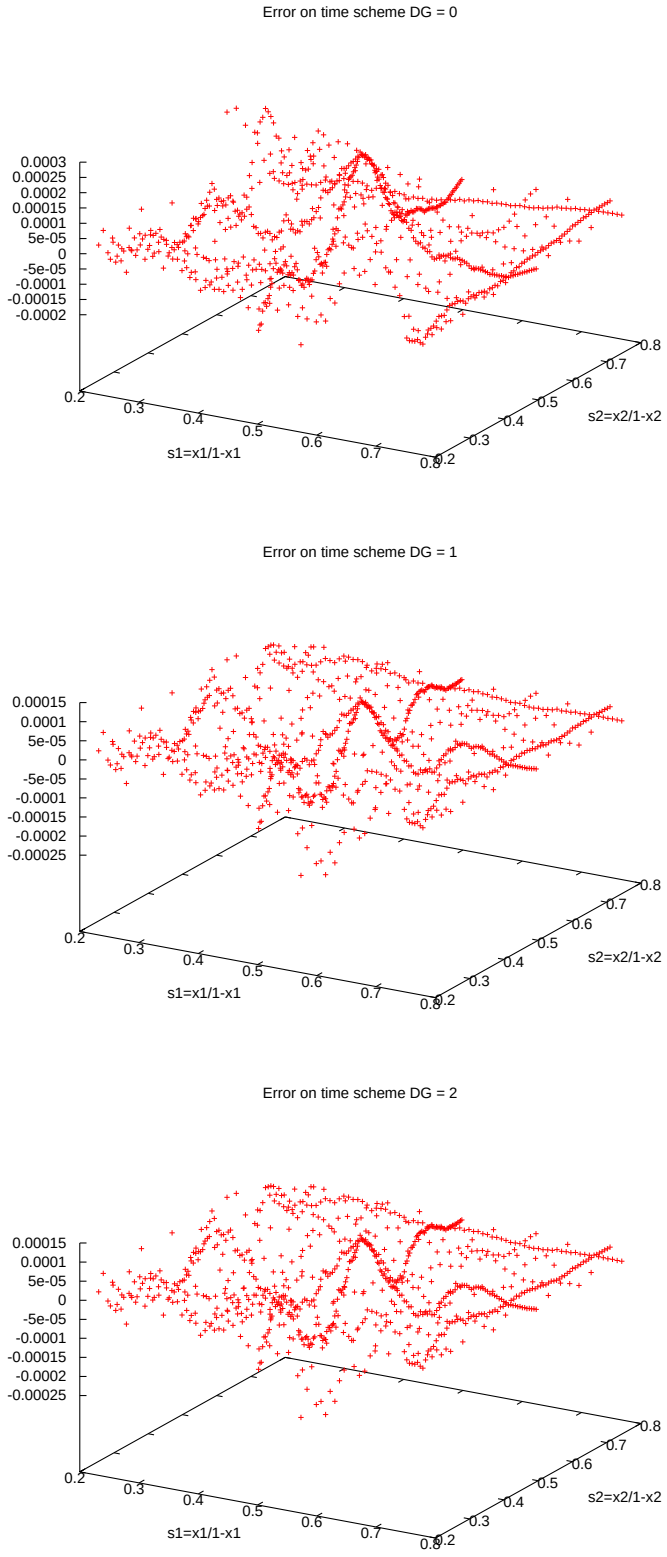

FIG. 8.6 – Prix et Erreur sur le prix d'une option Put européen en dimension 2 de strike $K = 1$ et de maturité $T = 1$ pour différentes discrétisations. Les paramètres sont $N_L = 8, N_I = 15$ et les couples (DG, N_T) : $(0, 100)$, $(1, 10)$, et $(2, 5)$.

8.3.2 Cas d'une matrice de corrélation non dégénérée

Les premiers tests réalisés dans le cas d'une matrice de corrélation non triviale sont conformes à notre hypothèse et les résultats obtenus au paragraphe précèdent semble se généraliser.

Nous proposons d'étudier deux cas pour lesquels la matrice de corrélation varient. Le vecteur de volatilité est $\sigma = (0.15, 0.20, 0.30, 0.25, 0.18)$, les autres paramètres sont donnés par le tableau 8.1 . Dans le premier cas, nous considérons une matrice de corrélation ρ qui vérifie $\rho_{i,j} = 0.4,\ i \neq j$ et 1 sinon. Les résultats partiels figuret dans le tableau 8.5. Dans le second cas, $\rho = Id$ et les résultats du tableau 8.6 sont acceptables en dimension 3 et 4 ; le passage à la dimension 5 nécessite d'augmenter le niveau de discrétisation.

(N_L, DG)	Prix MC	var	$(6,0)$	$(6,2)$	$(7,0)$	$(7,2)$	$(8,0)$	$(8,2)$
$d=2$	5.882	0.092	$\star$	$\star$	5.828	5.836	5.859	5.870
$d=3$	6.760	0.124	$\star$	$\star$	6.689	6.705	6.749	6.751
$d=4$	6.714	0.121	6.711	6.719	6.733	6.747	$\star$	$\star$
$d=5$	6.253	0.106	6.428	6.339	$\star$	$\star$	$\star$	$\star$

TAB. 8.5 – Prix à la monnaie d'un Put sur panier européen de maturité 1 an et de strike 100. La matrice de correlation ρ est définie par $\rho_{i,j} = 0.4,\ i \neq j$ et 1 sinon. $N_T = 100$ si $DG = 0$ et $N_T = 5$ si $DG = 2$.

	MC	Galerkin
$d = 3, N_L = 8$	5.1822	5.197
$d = 4, N_L = 7$	4.618	4.648
$d = 5, N_L = 6$	3.880	4.082

TAB. 8.6 – Prix à la monnaie d'un Put européen de maturité 1 an et de strike 100. La matrice de correlation ρ est l'identité. Le schéma d'Euler implicite ($DG = 0$) à 100 pas de temps est appliqué pour la discrétisation en temps.

Chapitre 9

Description du Code

Les deux architectures utilisées pour l'implémentation des algorithmes de résolution sur des grilles sparse sont présentées dans ce chapitre.

L'aspect général des deux codes est décrit dans la première partie. La deuxième partie est consacrée au code de la méthode de différences finies sur des grilles sparse. La dernière partie propose une implémentation de la méthode de Galerkin sur une base d'ondelettes sparse. L'architecture est construite autour d'une structure de stockage morse du produit tensoriel sparse. Le calcul des matrices de rigidité en dimension un est également détaillé.

9.1 Schéma général

Les schémas utilisés pour les problèmes paraboliques conduisent à la résolution à chaque pas de temps d'un problème elliptique.

9.1.1 Système linéaire

Nous sommes amenés à résoudre un système linéaire :

$$Ax = b, \tag{9.1}$$

où A est la matrice de l'opérateur elliptique. Les quelques précisions suivantes sur la forme de la matrice A justifient les choix effectués ensuite :

- En général, l'opérateur $\mathcal{L}$ associé au problème elliptique n'est pas symétrique ($\langle \mathcal{L}u, v\rangle \neq \langle u, \mathcal{L}v\rangle$). La matrice A n'est donc pas symétrique.
- La matrice A, bien que creuse, ne présente pas une structure particulière (voir Schiekofer [Sch98a] pour l'étude de la matrice de discrétisation d'une méthode de différences finies sparse). En particulier, la matrice n'est pas une matrice bande et la position des coefficients non nuls dépend fortement de l'indexation choisie pour les points de grille.

La seconde observation implique que les méthodes de résolution directes, factorisation LU suivie d'un algorithme de descente remontée, sont inappropriées. Le choix d'une méthode itérative s'impose. La première observation réduit le nombre de méthodes itératives à notre disposition. Nous considérons dans les applications numériques les méthodes GMRES (*Generalized Minimal Residual Method*) et BICGStab (stabilized biconjugate gradient). Ces deux algorithmes sont décrits dans [Saa96]. La méthode GMRES

est modifiée pour inclure les possibilités de redémarrage afin d'économiser les ressources mémoire. Le facteur de « restart » est de l'ordre de 10. Le choix du préconditionneur (à gauche, à droite ou appliqué de manière symétrique) est discuté aux pages 270-271 de [Saa96]. Nous utilisons un préconditionneur appliqué de manière symétrique pour la méthode de Galerkin et à gauche pour la méthode de différences finies.

Les deux codes présentés ci-dessous divergent sur le produit matrice - vecteur Ax. Dans le cas d'une méthode de différences finies, les opérateurs sont donnés par (2.69,. . . ,2.74). La technique consiste à composer des opérateurs unidimensionnels. La méthode de Galerkin suit une approche plus classique avec un véritable produit matrice-vecteur. La matrice A est stockée sous une forme creuse.

9.1.2 Représentation d'un point dans une subdivision dyadique

La discrétisation sur une base d'ondelettes (*i.e.* sur une base hiérarchique) implique la construction d'une grille dyadique. Une subdivision dyadique de $[0,1]$ de niveau n est constituée des points $x_k = \frac{k}{2^n}$ où $k = 0, \ldots, 2^n$. La grille est constituée de $2^n + 1$ points (*resp.*$2^n - 1$ sur l'ouvert $]0,1[$). Nous disposons de différentes représentations d'un point sur cette grille.

(i) La représentation classique, notée représentation nodale, consiste à indexer le point x par son indice de translation k.

(ii) La représentation d'échelle correspond à l'indexation des fonctions sur la base des fonctions chapeaux multi-niveaux. Elle correspond à la représentation des fonctions de base de l'espace V_ℓ. Cette indexation est caractérisée par un couple (ℓ, i) où ℓ représente le niveau de raffinement de la fonction et i l'indice de translation. Le point x correspond à la valeur $\dfrac{i}{2^\ell}$ avec $\ell \in \mathbb{N}$, $i = 0..2^\ell$. Cette représentation n'est pas unique. En effet, si i est pair, alors le couple $(\ell - 1, i/2)$ représente le même point. La représentation hiérarchique, notée (ℓ, i) , consiste à choisir le niveau de sorte que i soit impair.

(iii) La représentation en arbre (binaire), permet de stocker la représentation hiérarchique à l'aide d'un unique indice j. Les points de bords sont caractérisés par les valeurs $j = 0$ et $j = N$(N est l'entier maximal de la représentation), pour $x = 0$, et $x = 1$. Les autres points correspondent à la représentation sous forme d'arbre binaire $(1, 2, 3, 4, 5, \ldots) \rightarrow (0.5, 0.25, 0.75, 0.125, 0.375, \ldots)$.

(iv) La représentation en ondelette correspond à la représentation des fonctions de base de l'espace de détail. Le changement entre la représentation hiérarchique et la représentation en ondelette est mise en équation à la remarque 1.5. Notons que l'ondelette $\psi_{\ell,i}$ est centrée sur le point $x = \dfrac{2i+1}{2^{\ell+1}}$. Ceci provient de la définition des filtres d'ondelette qui implique un centrage de la fonction d'ondelette mère sur le point $\frac{1}{2}$ (dans le cas des ondelettes biorthogonales $n, \tilde{n}$ avec n et $\tilde{n}$ pairs). Nous en déduisons la relation de passage entre base hiérarchique (ℓ, i) et base d'ondelette $\left(\bar{\ell}, \bar{i}\right)$:

$$\bar{\ell} = \ell - 1, \quad \bar{i} = \frac{i-1}{2}. \tag{9.2}$$

Donnons à présent les relations de passage entre ces différentes caractérisations.

1. Passage de la représentation arbre à la représentation d'échelle : notons $j = Scale\,(\ell, i)$ si
$$\begin{cases} \ell = \log_2(j) + 1 \\ i = 1 + j - 2^{\log_2(j)} \end{cases} \tag{9.3}$$
La fonction $Ondelette : \bar{\ell} \times \bar{i} \to j$ se déduit de la composition de (9.2) et (9.3).
2. Passage de la représentation hiérarchique à la représentation nodale :
$$k = i2^{n-\ell}. \tag{9.4}$$
3. Passage de la représentation arbre à la représentation nodale :
$$k = (1 + j)\, 2^{n-1-\log_2(j)} - 2^{n-1}. \tag{9.5}$$

Pour obtenir les meilleurs performances, nous avons choisi de mettre en oeuvre la représentation par arbre.

Représentation dyadique de $[0,1]^d$ La notion de subdivision dyadique de $[0,1]$ est étendue $[0,1]^d$ en considérant un vecteur $\boldsymbol{j} = (j_1, \ldots, j_d)$ où j_k est l'indice de la représentation par arbre de la $k^{\text{ème}}$ coordonnée x_k. Nous nommerons **point** ce vecteur d'indice.

Nous aurons besoin de définir différentes fonctions de « niveau » de ce point : le niveau ℓ_k dans chacune de direction, le niveau global c.-à-d. le maximum ou la somme des ℓ_k.

9.2 Méthode de différences finies

Nous proposons ici une approche innovante, différente de celle présentée par exemple par Zumbusch [Zum00].

9.2.1 Grille sparse

La grille est constituée d'une liste de points correspondant chacun à une fonction de base. L'élément p de cette liste est associé à la $p^{\text{ème}}$ coordonnée du vecteur d'inconnues, de sorte que si une fonction u est représentée sur la base sparse, alors

$$u = \sum_{\boldsymbol{\ell} \in \mathcal{I}_n, \boldsymbol{\imath} \in \tau_{\boldsymbol{\ell}}} u_p \psi_{\boldsymbol{\ell},\boldsymbol{\imath}}, \quad \text{avec point}[p] = (j_1, \ldots, j_d) \text{ et } \forall\, k = 1, \ldots, d \quad j_k = \text{Scale}\,(\ell_k, i_k)\,. \tag{9.6}$$

La grille fournit aussi un ensemble de relations entre ces points. Nous distinguons deux types de relation : celles utilisées dans le produit matrice vecteur (à optimiser en priorité) et les autres, n'intervenant que dans la construction de la grille.

L'approche de Zumbusch consiste à définir ces relations sur les multi-indices $(j_1, \ldots, j_d)$ puis à disposer d'une méthode efficace de recherche des couples (point, indice) basée sur les hash-map et les « space-filling curve ».

Les informations pertinentes pour un point de la grille sont
- ses fils (au plus deux par dimension),
- ses pères (nous définissons deux pères, un direct et l'autre indirect pour accélérer les transformations nodal → hiérarchique),

– ses voisins : les points les plus proches (au plus deux par dimension).

Nous devrons remplir les tableaux **son**, **father** et **neighbour** qui dépendront de trois indices. Par exemple, $son\,[p, dim, lr]$ est l'indice du fils du point d'indice p dans la dimension dim et la direction lr. L'entier dim varie de 1 à d et l'entier lr vaut 0 (à gauche) ou 1 (à droite). Le changement de direction correspond à l'opération $lr \to 1 - lr$.

9.2.1.1 Définition des tableaux

(i) Si l'indice p a pour représentation dans l'arbre le multi-indice $(j_1, \ldots, j_d)$, *i.e.* $point[p] = (j_1, \ldots, j_d)$, alors $son\,[p, dim, lr]$ est représenté par le multi-indice $(j_1, \ldots, j_{dim-1}, 2j_{dim} + lr, j_{dim+1}, \ldots, j_d)$, si ce point est présent dans la grille.

(ii) La notion de père intervient dans la transformation nodal → hiérarchique et sa réciproque. Nous distinguons le père direct et le père indirect. Pour un point d'indice p dont la représentation est $(j_1, \ldots, j_d)$, la direction directe dans la dimension dim sera

$$direct = \begin{cases} 1 \text{ (droite)} & \text{si } j_{dim} \text{ est pair} \\ 0 \text{ (gauche)} & \text{sinon} \end{cases} . \tag{9.7}$$

Alors $father\,[p, dim, direct]$ est l'indice du point représenté par $(j_1, \ldots, j_{dim-1}, j_{dim}/2, j_{dim+1}, \ldots, j_d)$. Le père indirect dans la dimension dim est déterminé par la relation de récurrence

$$father[p, dim, 1 - direct] = father\,[father\,[p, dim, direct]\,, dim, 1 - direct]\,. \tag{9.8}$$

La figure 9.1 illustre cette définition.

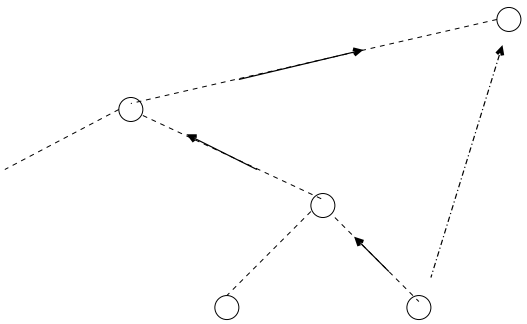

FIG. 9.1 – Définition de **father** : la relation de père direct est représentée par une flèche pleine. La relation de père indirect est indiquée par une flèche en pointillé. Celle-ci est obtenue en parcourant récursivement les ancêtres directs jusqu'à ce que la direction *direct* change

(iii) L'indice $neighbour\,[p, dim, lr]$ désigne le point de la grille qui minimise la distance au point d'indice p sous les contraintes suivantes :
- toutes ses coordonnées (sauf dans la dimension dim) sont égales à celle de p,
- sa coordonnée dans la dimension dim est supérieure (*resp.* inférieure) à celle de p si $lr = 1$ (*resp.* 0).

9.2.1.2 Construction des tableaux

Le tableau **son** permet de construire les deux autres. Un algorithme récursif pour la construction de **son** sera présenté ultérieurement.

(i) Construction du tableau **father** : un premier balayage du tableau **son** permet de construire les pères directs des points de la grille, d'après la règle de symétrie :

$$father\,[son[p, dim, 1 - direct], dim, direct] = p, \tag{9.9}$$

où la règle de construction de direct est donnée par (9.7). Comme la numérotation des points est telle que

$$father\,[p, dim, lr] < p, \quad \forall p,$$

nous pouvons construire, dans le même balayage, le tableau des pères indirects en utilisant (9.8).

(i) Construction du tableau **neighbour**. Elle peut être abordée de deux manières : soit à partir du tableau **son** uniquement (plus de calcul), soit après construction du tableau **father**. Nous aurons recours à la relation de symétrie

$$neighbour\,[neighbour[p, dim, lr], dim, 1 - lr] = p. \tag{9.10}$$

Nous définissons un ensemble utile pour la construction de **neighbour** :

$$Sans\_Descendant_{dim} = \left\{ \begin{array}{l} p \text{ tel qu'il n'existe pas de point } q \text{ dans la grille avec} \\ q = son\,[p, dim, lr]\,,\ lr = 0 \text{ ou } 1 \end{array} \right\}. \tag{9.11}$$

1 **Construction de neighbour à partir de son.** L'idée est que les voisins de p sont des descendants de p qui n'ont pas de descendants (voir figure 9.2)

if $p \notin Sans\_Descendant_{dim}$
(initialisation) $n = p$,
while $(n \notin Sans\_Descendant_{dim})$ loop
$n = son\,[n, dim, lr^{\star}]$,
end loop
$neighbour[p, dim, lr] = n$,
$neighbour[p, dim, 1 - lr] = p$, end if

avec $lr^{\star} = lr$ si $n = p$ (première itération de l'algorithme), et $lr^{\star} = 1 - lr$ sinon. Les voisins n'ayant pas de fils dans la dimension dim sont construits par la relation de symétrie (9.10) dans la dernière ligne de l'algorithme.

2 **Construction de neighbour à partir des tableaux son et father.** Le tableau est construit d'après la règle « un noeud n'ayant pas de fils a pour voisin ses pères ». L'algorithme consiste à parcourir la grille. Si le point p n'a pas de fils dans la dimension dim et dans la direction lr, alors son voisin (dim, lr) est son père (dim, lr).

if $(p \in Sans\_Descendant_{dim})$ then
$neighbour[p, dim, lr] = father[p, dim, lr]$,
$neighbour\,[p.father[dim, lr], dim, 1 - lr] = p$,
end if.

Remarque 9.1 *Les règles décrites restent valables dans le cas où l'arbre binaire n'est pas* complet*, c.-à-d. sur des grilles non-uniformes.*

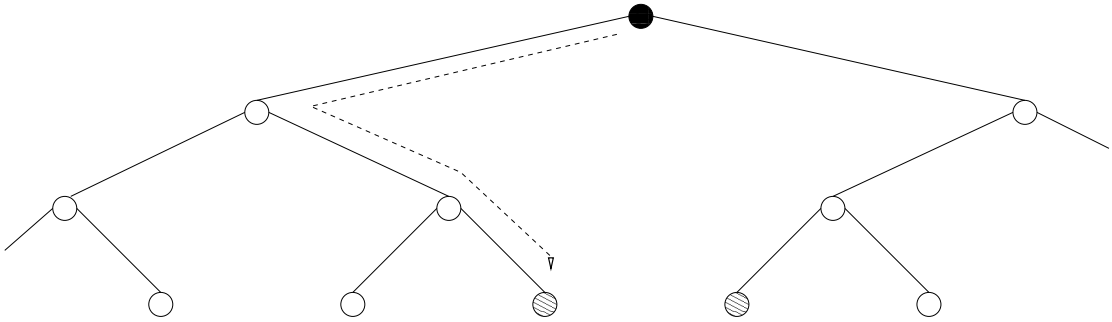

FIG. 9.2 – Construction du tableau **neigbour**, suivant la méthode 1. Le voisin à gauche ($lr = 0$) du noeud en noir est construit en descendant le graphe d'abord à gauche ($lr^\star = lr$) puis toujours à droite ($lr^\star = 1 - lr$) jusqu'à la feuille de l'arbre

9.2.2 Construction de la grille

La construction d'une grille sparse se fait naturellement par récurrence. L'algorithme proposé consiste à construire les points géométriques et à mettre à jour le tableau **son** après la création d'un point. La relation de récurrence correspond à la notion de somme d'espaces de détail :

$$\hat{V}_{\mathbf{n}} = \hat{V}_{\mathbf{n-1}} \oplus \hat{W}_{\mathbf{n}} = \bigoplus_{|k|_1 \leq n+d-1} \hat{W}_{\mathbf{k}}. \tag{9.12}$$

Les points de la grille Ω_n sont obtenus par la réunion de la grille Ω_{n-1} et de l'ensemble des points de Ω_{n-1} :

$$\Omega_n = \Omega_{n-1} \cup \left\{ p \middle| p = son\left[q, dim, lr\right], \quad q \in \Omega_{n-1},\ 1 \leq dim \leq d,\ 0 \leq lr \leq 1 \right\}. \tag{9.13}$$

Plus précisément, l'algorithme construit les grilles $\Omega_{n/(n-1)}$ définies par la récurrence

$$\Omega_{n/(n-1)} \left\{ p \middle| p = son\left[q, dim, lr\right], \quad q \in \Omega_{(n-1)/(n-2)},\ 1 \leq dim \leq d,\ 0 \leq lr \leq 1 \right\}, \tag{9.14}$$

et $\Omega_{0/(-1)} = \Omega_0$.

L'apparente simplicité de cette construction par récurrence ne met pas en évidence certaines difficultés liées à la structure du tableau **son**. En effet, bien que la structure soit proche de celle d'un arbre binaire, elle ne peut pas être considérée comme tel, puisque

$$\exists\ p, q \in \Omega_{(n-1)},\ p \neq q,\ \ 1 \leq d_1 \leq d_2 \leq d,\ lr \in \{0,1\} \quad son\left[p, d_1, lr\right] = son\left[q, d_2, lr\right]. \tag{9.15}$$

Prenons un exemple pour $d = 2$: le point $(2,2)$ peut être obtenu en calculant les fils des points $(1,2)$ et $(2,1)$. En revanche, lorsque nous orientons le graphe, aucune boucle n'est possible. La structure de base est donc celle d'un graphe orienté (voir la figure 9.3). Pour cette raison, il est nécessaire d'utiliser une structure ordonnée (une map, ou hash-map) qui permet de vérifier qu'un point n'a pas été ajouté précédemment.

Remarque 9.2 *Une importante littérature traite de la structure informatique pour la représentation de bases ou grilles hiérarchiques. Ces travaux sont regroupés autour du système DAGH (Dynamic Adaptative Grid Hierarchies). Celui-ci propose l'utilisation des Space Filling Curve en tant que fonction de répartition pour la table de hachage. Le lecteur trouvera dans [Zum03] une illustration de cette méthode pour des* Sparse Grid.

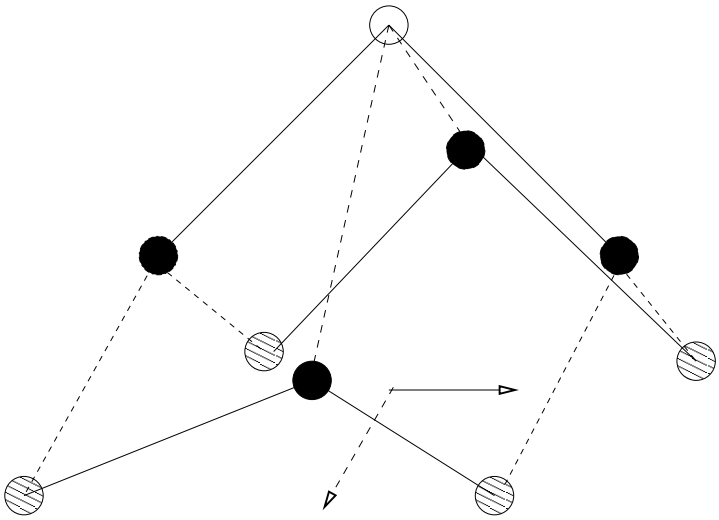

FIG. 9.3 – Multiplication des chemins - Partant de la racine du graphe, il est possible d'atteindre chacune des feuilles (noeuds hachurés) par au moins deux chemins : soit en suivant d'abord une direction dans la dimension x_1 (trait pointillé) puis une direction dans la dimension x_2 (trait plein) ou l'inverse.

9.2.3 Produit matrice-vecteur

Nous précisons dans ce paragraphe le « produit matrice-vecteur » sur deux exemples. Le premier exemple, bien que simple puisqu'il provient du problème de Poisson, permet de comprendre le découpage par dimension. Le second exemple reprend l'équation de valorisation du modèle à volatilité stochastique donné au chapitre 5 et, plus spécifiquement, la résolution numérique de (6.32). Certains termes de cette équation permettent de mettre en évidence des comportements spécifiques des tenseurs de discrétisation.

9.2.3.1 Le Laplacien

Reprenons les notations du paragraphe 2.3 et rappelons la discrétisation du Laplacien (2.69)

$$\frac{\partial^2}{\partial x_i^2} \approx \mathrm{T}_{(i)} \circ \mathrm{D}^2_{(i)} \circ \mathrm{T}^{-1}_{(i)}, \quad \Delta \approx \widehat{\Delta} = \sum_{i=1}^{d} \mathrm{T}_{(i)} \circ \mathrm{D}^2_{(i)} \circ \mathrm{T}^{-1}_{(i)}, \tag{9.16}$$

où $\mathrm{T}_{(i)}$ est la transformation nodal $\rightarrow$ hiérarchique dans la dimension i et $\mathrm{T}^{-1}_{(i)}$ la transformation réciproque.

Le tenseur $\mathrm{D}^2_{(dim)}$ donné par (2.51) correspond au pseudo code :

$$\mathrm{D}^2_{(dim)}(U)(p) = \frac{U(neighbour[p, dim, 1]) - 2U(p) + U(neighbour[p, dim, 0])}{(dist(p, neighbour[p, dim, 1], dim))^2}, \tag{9.17}$$

où U est le vecteur de la représentation de la fonction u sur la grille et *dist* une fonction qui calcule la distance.

La généralisation à une grille anisotrope est facile.

La transformation nodal → hiérarchique dans le dimension dim et sa réciproque sont également obtenues à partir de la fonction **father**

$$\mathrm{T}_{(dim)}\left(U\right)\left(p\right) = U(p) - \frac{1}{2}\left(U\left(father\left[p, dim, 1\right]\right) + U\left(father\left[p, dim\right]\right)\right), \tag{9.18}$$

$$\mathrm{T}^{-1}_{(dim)}\left(U\right)\left(p\right) = U(p) + \frac{1}{2}\left(\mathrm{T}^{-1}_{(dim)}\left(U\right)\left(father\left[p, dim, 1\right]\right) + \mathrm{T}^{-1}_{(dim)}(U)\left(father\left[p, dim, 0\right]\right)\right). \tag{9.19}$$

L'ordre de parcours des vecteurs est primordial pour cette dernière fonction : pour calculer la transformation au point p, il est nécessaire de l'avoir préalablement calculée aux points $father\left[p, dim, lor\right],\ lor = 0, 1$.

9.2.3.2 Modèle à volatilité stochastique

Nous donnons les fonctions correspondant aux termes qui posent problème. Pour alléger la présentation, considérons l'équation sur la variable $x = log(s)$, et notons m le nombre de facteurs du modèle à volatilité stochastique $m = d - 1$. Nous étudions les termes

- de diffusion sur x : $f\left(y_1, \ldots, y_m\right)^2 \dfrac{\partial^2 u}{\partial x^2} = \mathcal{D}_x(u)$,
- de corrélation x, y_i : $\displaystyle\sum_{i=1}^{m} \beta_i f\left(y_1, \ldots, y_m\right) \frac{\partial^2 u}{\partial x \partial y_i} = \sum_{i=1}^{m} \beta_i \mathcal{C}_{x,y_i}(u)$.

Les difficultés viennent de la complexité de l'algorithme servant à effectuer le produit de fonctions de grille (voir l'équation (2.74)).

En notant $F(p)$ $(resp.F^2(p))$ les fonctions qui renvoient à $f\left(y_1, \ldots, y_m\right)$ $(resp.f\left(y_1, \ldots, y_m\right)^2)$, le premier terme est calculé à l'aide de la composition suivante

$$\mathcal{D}_x(U)\left(p\right) = \mathrm{T}_{(y_1)} \ldots \mathrm{T}_{(y_m)}\ F^2(p)\ \mathrm{T}^{-1}_{(y_m)} \ldots \mathrm{T}^{-1}_{(y_1)}\ \mathrm{T}_{(x)}\mathrm{D}^2_{(x)}\mathrm{T}^{-1}_{(x)}\left(U\right)\left(p\right). \tag{9.20}$$

(9.20) est équivalente à

$$\mathcal{D}_x(U)\left(p\right) = \mathrm{T}_{(x)}\mathrm{D}^2_{(x)}\mathrm{T}^{-1}_{(x)}\ \mathrm{T}_{(y_1)} \ldots \mathrm{T}_{(y_m)} F^2(p)\ \mathrm{T}^{-1}_{(y_m)} \ldots \mathrm{T}^{-1}_{(y_1)}\left(U\right)\left(p\right). \tag{9.21}$$

Le terme de corrélation $\mathcal{C}_{x,y_i}$ nécessite également une opération de multiplication de deux fonctions de grille. L'écriture la plus générale de cette opérateur est donnée par

$$\mathcal{C}_{x,y_i}(U)\left(p\right) = \mathrm{T}_{(y_1)} \ldots \mathrm{T}_{(y_m)} F(p)\ \mathrm{T}^{-1}_{(y_m)} \ldots \mathrm{T}^{-1}_{(y_1)} \mathrm{T}_{(y_i)}\mathrm{D}_{(y_i)}\mathrm{T}^{-1}_{(y_i)}\mathrm{T}_{(x)}\mathrm{D}_{(x)}\mathrm{T}^{-1}_{(x)}\ \left(U\right)\left(p\right). \tag{9.22}$$

Il est possible d'économiser deux transformations $\mathrm{T}_{(y_i)}$ et $\mathrm{T}^{-1}_{(y_i)}$:

$$\mathcal{C}_{x,y_i}(U)\left(p\right) = \mathrm{T}_{(y_i)}\mathrm{T}_{(y_1)} \ldots \mathrm{T}_{(y_m)} F(p) \mathrm{T}^{-1}_{(y_1)} \ldots \mathrm{T}^{-1}_{(y_m)} \mathrm{D}_{(y_i)}\mathrm{T}^{-1}_{(y_i)}\mathrm{T}_{(x)}\mathrm{D}_{(x)}\mathrm{T}^{-1}_{(x)}\ \left(U\right)\left(p\right). \tag{9.23}$$

Si f est à variables séparées, il est possible de calculer ce terme de la façon suivante :

$$\mathcal{C}_{x,y_i}(U)\left(p\right) = \mathrm{T}_{(y_1)} F_1(p) \mathrm{T}^{-1}_{(y_1)} \ldots \mathrm{T}_{(y_m)} F_m(p) \mathrm{T}^{-1}_{(y_m)} \mathrm{T}_{(y_i)} F_i(p) \mathrm{D}_{(y_i)}\mathrm{T}^{-1}_{(y_i)}\mathrm{T}_{(x)}\mathrm{D}_{(x)}\mathrm{T}^{-1}_{(x)}\ \left(U\right)\left(p\right). \tag{9.24}$$

Enfin dans le cas particulier $f_i(y_i) = \exp\left(c_i y_i\right)$, en écrivant le terme de corrélation sous la forme divergence :

$$s\sum_{i=1}^{m}\beta_i f\left(y_1,\ldots,y_m\right)\frac{\partial^2 u}{\partial x\partial y_i}=\sum_{i=1}^{m}\beta_i\frac{\partial^2}{\partial x\partial y_i}\left(f\left(y\right)u\right)-\sum_{i=1}^{m}\beta_i c_i\frac{\partial}{\partial x}\left(f\left(y\right)u\right),$$

il est possible de pré-calculer point par point les produits $f(y)u$ et $f(y)^2u$ avant d'appliquer des opérateurs de différences finies.

Remarque 9.3 *La formulation (9.23) peut être la seule admissible, par exemple dans le cas où la fonction de volatilité $f(y)$ est définie par :*

$$f(y_1,\ldots,y_m)=min\left(\exp\left(\frac{1}{2}\sum_{i=1}^{m}y_i\right),C\right).$$

Cette modification, apparemment légère, de la volatilité accroît considérablement la complexité du calcul.

Estimation de la complexité de la discrétisation de (5.13) A partir des schémas de discrétisation établis ci-dessus, nous évaluons le nombre de parcours du vecteur pour la multiplication matrice-vecteur. Le nombre de parcours du vecteur d'inconnues est de l'ordre de :

pré-calcul : m transformations hiérarchique $\rightarrow$ nodal, et $2m$ transformations nodal $\rightarrow$ hiérarchique. Plus précisément, le passage hiérarchique $\rightarrow$ nodal de u dans toutes les dimensions y_i, et la transformation réciproque de $f(y)u$ et $f(y)^2u$.

Différences finies en x : 3 transformations hiérarchique $\rightarrow$ nodal sur u, $f(y)u$ et $f(y)^2u$. 2 transformations réciproques, la première sur la somme des termes de dérivation par rapport à x uniquement et la seconde sur le terme de corrélation, noté $\mathcal{C}_x(u)$.

Différences finies en y_k : sur chacune des dimensions y_k, il est nécessaire d'effectuer $k+1$ transformations hiérarchique $\rightarrow$ nodal. Elles sont effectuées sur u, $\mathcal{C}_x(u)$ et les $k-1$ dérivées croisées sur les variables y_i y_k, $1\leq i<k$. Nous effectuons deux transformations réciproques : la première sur les termes dépendant de y_k uniquement, de y_k, x et de y_k, y_i pour $i<k$; la seconde sur le terme $\dfrac{\partial u}{\partial y_k}$. Ce terme permet de calculer ensuite les termes de dérivés croisées $\frac{\partial^2 u}{\partial y_k\partial y_j}$, $j>k$. Si $k=n$, cette dernière transformation n'est pas nécessaire.

Au total nous obtenons

$$\begin{aligned} h\ &\rightarrow\ n: \quad m+3+\sum_{k=1}^{m}(k+1)=\frac{(m+3)(m+2)}{2}=\frac{(d+2)(d+1)}{2},\\ n\ &\rightarrow\ h: \quad 2m+2+(2m-1)=4m+1=4d-3. \end{aligned} \tag{9.25}$$

Le nombre de parcours du vecteur est de l'ordre de $C(d)=\dfrac{(d+2)(d+1)}{2}+4d-3$, ce qui pour $d=4$ donne 28 parcours du vecteur d'inconnues.

Nous comparons cette complexité théorique avec celle d'une méthode ADI sur une grille pleine dans une dimension de référence.

Dans le cas d'une méthode ADI en dimension d_{ref}, nous supposons que deux parcours du vecteur par dimension sont suffisants pour calculer la solution de l'équation $Ax=b$.

Nous supposons que le nombre d'itérations M de GMRES est constant et considérons les deux cas $M = 10$ et $M = 20$, valeurs qui sont observées pour les résolutions des systèmes linéaires.

Suivant ces hypothèses, il est possible de comparer la complexité théorique de l'algorithme avec une méthode d'ADI en dimension 3. Le tableau 9.1 donne le nombre de points dans chacune des dimensions de la grille isotrope d'une méthode d'ADI en dimension 3 qui aurait la même complexité théorique que la méthode de discrétisation sur une grille sparse. Ce tableau est obtenu en appliquant la formule $\left(\frac{C(d)M2^n n^{d-1}}{2d_{ref}}\right)^{1/d_{ref}}$, avec $d_{ref} = 3$. Il nous donne une idée assez précise du temps de calcul de la méthode de différences finies sparse en fonction de la dimension et du niveau de discrétisation. Par exemple, en dimension 4 avec une grille de niveau de discrétisation 7, la méthode sparse a une complexité du même ordre qu'une méthode ADI en dimension 3 sur une grille à $127^3 \approx 2e^6$ points. Le tableau 9.2 illustre l'intérêt des techniques de Sparse Grid. Nous y indiquons le nombre de point de grille d'une méthode ADI dans la même dimension $d_{ref} = d$.

Tab. 9.1 – Complexité théorique de la méthode de différences finies pour le modèle à volatilité stochastique exprimée par rapport au nombre de points d'une grille utilisée dans une méthode ADI en dimension 3, $M = 10$ (gauche) et $M = 20$ (droite).

Dim/Niv	3	4	5	6
7	58	127	269	560
8	80	183	404	881
9	109	259	596	1352

Dim/Niv	3	4	5	6
7	74	160	339	705
8	102	230	510	1111
9	138	326	752	1703

Tab. 9.2 – Complexité théorique de la méthode de différences finies sur le modèle à volatilité stochastique en dimension d exprimée par rapport au nombre de points d'une grille utilisée dans une méthode ADI dans la même dimension, $M = 10$ (gauche) et $M = 20$ (droite).

Dim/Niv	4	5	6
7	35	26	21
8	46	33	26
9	60	42	33

Dim/Niv	4	5	6
7	42	30	24
8	55	38	30
9	71	48	37

9.3 Méthode de Galerkin sur une base d'ondelettes

Nous présentons dans ce paragraphe le code de résolution numérique par une méthode de Galerkin sur une base d'ondelettes. Dans une première partie, nous détaillons la structure de données pour les matrices de rigidité d'un opérateur sur $[0, 1]$. Nous introduisons ensuite la même structure pour l'opérateur tensoriel.

9.3.1 Construction des matrices de rigidité sur $[0,1]$

9.3.1.1 Stockage morse de la matrice de rigidité

Considérons la forme bilinéaire

$$a(u,v) = \int_\Omega f(x) u^{(n_1)}(x) v^{(n_2)}(x) dx, \tag{9.26}$$

où f correspond à un coefficient de l'équation différentielle et n_1, n_2 sont les ordres de dérivation.

La base d'ondelettes est représentée à l'aide de la structure d'arbre binaire. Plus précisément, l'indice γ est associé à l'ondelette $\psi_{\ell,\imath}$, avec $\gamma = \text{Ondelette}\,(\ell,\imath)$ (voir (9.2) et (9.3)).

La structure adoptée est celle d'un stockage morse. Seuls les coefficients non nuls des matrices de rigidité associées à chacun des opérateurs sont stockés, correspondant à des couples d'ondelettes dont les supports sont non-disjoints. Construisons l'ensemble des couples d'ondelettes indexées par $\gamma \leftrightarrow (\ell,\imath)$ et $\overline{\gamma} \leftrightarrow \left(\bar{\ell},\bar{\imath}\right)$ tels que

$$\text{supp}\,(\psi_{\ell,\imath}) \cap \text{supp}\left(\psi_{\bar{\ell},\bar{\imath}}\right) \neq \emptyset.$$

Cet ensemble est stocké sous la forme d'un vecteur m de couple $(\gamma,\overline{\gamma})$. Ce stockage des emplacements des coefficients a priori non-nuls de la matrice de rigidité s'apparente à un stockage morse. Nous notons k l'indice du couple $(\gamma,\overline{\gamma})$ dans le vecteur m, *i.e.* $m[k] \rightarrow (\gamma,\overline{\gamma})$.

Pour chaque opérateur différentiel, nous calculons le coefficient de la matrice de rigidité associé à ce couple $(\gamma,\overline{\gamma})$. En conséquence, nous rangeons les coefficients non nuls de la matrice dans un vecteur v tel que :

$$v[k] \rightarrow a\left(\psi_{\ell,\imath}, \psi_{\bar{\ell},\bar{\imath}}\right). \tag{9.27}$$

Nous aurons besoin de la fonction $(\gamma,\overline{\gamma}) \rightarrow k$. Celle-ci correspond à la recherche dans le tableau m de l'indice du couple $(\gamma,\overline{\gamma})$. On appelle $A_{\gamma,\ell}$ l'ensemble des indices $\overline{\gamma}$ des ondelettes dont le niveau est ℓ et dont le support intersecte celui de l'ondelette γ. Pour accélérer la recherche dans le tableau m, nous proposons d'associer à chaque indice γ et à chaque niveau ℓ une structure de type map qui associe à chaque indice $\overline{\gamma}$ dans $A_{\gamma,\ell}$ l'indice k du couple $(\gamma,\overline{\gamma})$ dans le tableau m.

Nous obtenons aussi un tableau U à deux indices (γ,ℓ) dont le champ est de type map. Pour un couple $(\gamma,\overline{\gamma})$, la recherche s'écrit

$$k = U\left[\gamma, \log_2(\overline{\gamma})\right].Find(\overline{\gamma}).$$

La taille de la map $U[\gamma,\ell]$ est égale à la taille des filtres de la base d'ondelettes utilisée, par exemple 3 ou 5.

Afin d'accélérer encore la recherche et d'économiser les ressources mémoires, nous notons

$$\widetilde{A}_{\gamma,\ell} = \left\{\overline{\gamma} \in A_{\gamma,\ell} | \overline{\gamma} \leq \gamma\right\},$$

et nous associons à chaque indice γ et niveau ℓ inférieur ou égal à celui de γ, une structure de type map qui associe à chaque indice $\overline{\gamma}$ dans $\widetilde{A}_{\gamma,\ell}$ un couple d'entier correspondant à l'indice de $(\gamma,\overline{\gamma})$ dans m et à celui de $(\overline{\gamma},\gamma)$ dans m.

9.3.1.2 Calcul des matrices de rigidité

Le code utilisé pour le calcul des coefficients de la matrice de rigidité d'un opérateur de la forme (9.27) suit la méthode proposée au paragraphe 2.4.3.2. Il nécessite une fonction qui renvoie $a\left(\varphi_{\ell,\imath},\varphi_{\ell,\imath+k}\right)$, où $\varphi_{\ell,\imath}$ est la translatée/dilatée de la fonction chapeau. L'annexe C donne ces fonctions pour certaines formes bilinéaires a. Dans les cas où l'obtention d'une forme analytique de $a\left(\varphi_{\ell,\imath},\varphi_{\ell,\imath+k}\right)$ s'avère délicate, le logiciel MAPLE permet d'obtenir le code C de cette fonction.

9.3.2 Produit tensoriel sparse sur $[0,1]^d$

Soit a la forme bilinéaire associée au problème aux limites que nous souhaitons discrétiser sur une base d'ondelettes sparse, nous supposons qu'il existe une décomposition de la forme :

$$a\left(t;u,v\right)=\sum_{n}\alpha_n(t)a_n\left(u,v\right),$$

où a_n ne dépend pas de t. Nous supposons également que l'opérateur différentiel associé à a_i est à coefficients à variables séparées, voir (2.160,2.161). Les matrices de rigidité associées aux formes bilinéaires a_n sont stockées sous la forme d'un vecteur $\boldsymbol{v}_n$ de taille le nombre de paires de fonctions de base à supports non-disjoints. Étudions cette forme de stockage morse de la matrice de rigidité.

Nous construisons une table contenant les paires de fonctions de base à supports non-disjoints. Cette table est stockée sous la forme d'un vecteur noté $\boldsymbol{m}$. Elle fait correspondre à un indice $\boldsymbol{k}$ un couple d'indices $(\boldsymbol{\gamma},\overline{\boldsymbol{\gamma}})$. Le couple d'indices appartient à la table si, pour tout $dim=1,\ldots,d$,

$$\exists\ell,\ \overline{\gamma}_{dim}\in A_{\gamma_{dim},\ell}, \tag{9.28}$$

où $point\left[\boldsymbol{\gamma}\right]=(\gamma_1,\ldots,\gamma_d)$ et $point\left[\overline{\boldsymbol{\gamma}}\right]=(\overline{\gamma}_1,\ldots,\overline{\gamma}_d)$.

Nous construisons un second vecteur $\boldsymbol{t}$ dont le champ est une liste d'indices. Cette table renvoie, pour un chaque indice $\boldsymbol{k}$, la liste des indices $k_{dim},\ 1\leq dim\leq d$ pour les matrices 1D correspondantes. Nous avons

$$\boldsymbol{m}[\boldsymbol{k}]=(\boldsymbol{\gamma},\overline{\boldsymbol{\gamma}})\Leftrightarrow\quad\forall\ 1\leq dim\leq d,\quad m\left[k_{dim}\right]=(\gamma_{dim},\overline{\gamma}_{dim})\,.$$

La construction du vecteur $\boldsymbol{m}$ est une étape coûteuse. Nous avons adopté une méthode naïve qui consiste à parcourir pour chaque indice $\boldsymbol{\gamma}$ tous les indices $\overline{\boldsymbol{\gamma}}\geq\boldsymbol{\gamma}$ et à ajouter le couple à la table si les supports sont non-disjoints. Le tableau 2.6, qui montre que la matrice est quasiment pleine, justifie cette première approche.

L'opération de multiplication matrice-vecteur $a_i\left(u,\psi_{\boldsymbol{\ell},\boldsymbol{\imath}}\right)=\left(f,\psi_{\boldsymbol{\ell},\boldsymbol{\imath}}\right),\ \forall\boldsymbol{\ell}\in I_n^0$ est implémentée sous la forme suivante

$$\begin{array}{l}\text{pour }\boldsymbol{k}<\text{taille de }M\\ \quad F\left[\boldsymbol{m}[\boldsymbol{k}](2)\right]+=\boldsymbol{v}_n\left[\boldsymbol{k}\right]U\left[\boldsymbol{m}[\boldsymbol{k}](1)\right].\end{array} \tag{9.29}$$

Le coefficient $\boldsymbol{v}_n\left[\boldsymbol{k}\right]$ est calculé comme produit de coefficients de matrice de rigidité en dimension 1. A partir du vecteur $\boldsymbol{t}$ et des tableaux m et v (9.27) en dimension 1, nous calculons le coefficient par

$$\boldsymbol{v}_n\left[\boldsymbol{k}\right]=\prod_{dim=1}^{d}v_{dim}\left[m_{dim}\left[\boldsymbol{t}\left[\boldsymbol{k}\right](dim)\right]\right].$$

9.3.3 Solution sans stockage

Supposons qu'il soit possible de trouver une fonction analytique ou tabulée qui permette de calculer « rapidement » le coefficient de la matrice de rigidité. Cette fonction est notée

$$\mathcal{A} : (\boldsymbol{\ell}, \boldsymbol{\imath}), \left(\overline{\boldsymbol{\ell}}, \overline{\boldsymbol{\imath}}\right) = \boldsymbol{\gamma}, \overline{\boldsymbol{\gamma}} \rightarrow \mathcal{A}(\boldsymbol{\gamma}, \overline{\boldsymbol{\gamma}}).$$

Une telle fonction peut être obtenue à partir de la proposition 2.1.

Dans ce cas, il n'est pas nécessaire de construire les vecteurs $\boldsymbol{t}$ et $\boldsymbol{v}_n$. Il suffit de construire la table $\boldsymbol{m}$ du produit tensoriel sparse, à partir de (9.28). L'opération de multiplication matrice-vecteur $a_i(u, \psi_{\boldsymbol{\ell},\boldsymbol{\imath}}) = (f, \psi_{\boldsymbol{\ell},\boldsymbol{\imath}}),\ \forall \boldsymbol{\ell} \in I_n^0$ est implémentée sous la forme suivante

$$\begin{array}{l} \text{pour } \boldsymbol{k} < \text{taille de } M \\ \quad F[\boldsymbol{m}[\boldsymbol{k}](2)] += \mathcal{A}(\boldsymbol{m}[\boldsymbol{k}](1), \boldsymbol{m}[\boldsymbol{k}](2))\, U[\boldsymbol{m}[\boldsymbol{k}](1)]. \end{array} \tag{9.30}$$

Cette solution, même si elle s'avère être plus coûteuse en temps de calcul, peut devenir indispensable dans la mesure les ressources mémoire emaactuellement disponibles sur les ordinateurs ne suffisent pas pour les matrices associées au problème.

9.3.4 Perspective, parallélisation

La fin de ce chapitre nous permet de faire quelques remarques sur la parallélisation de cette méthode, qui reste un problème ouvert pour les architectures à mémoire distribuée. Les travaux de Zumbush [Zum03] semblent pouvoir être repris dans le cadre de cette méthode.

Sur une architecture à mémoire partagée, il est possible d'aborder d'une manière simpliste la parallélisation. Il suffit de trier le vecteur $\boldsymbol{m}$ par rapport à la seconde coordonnée et de le répartir sur p processeurs de sorte que les vecteurs $\left(\boldsymbol{m}^i\right)_{1 \leq i \leq p}$ vérifient :

$$i \neq j \Rightarrow \forall \boldsymbol{k}^i < size\left(\boldsymbol{m}^i\right),\ \boldsymbol{k}_j < size\left(\boldsymbol{m}_j\right)\ \ \boldsymbol{m}_i[\boldsymbol{k}_i](2) \neq \boldsymbol{m}_j[\boldsymbol{k}_j](2).$$

Cette méthode conduit à un algorithme parallèle quasi-optimal.

Troisième partie

A posteriori error estimates for parabolic inequalities

Chapitre 10

A posteriori error estimates for parabolic variational inequalities

The first part of this chapter has been accepted for publication in the Journal of Scientific Computing. It's a joint work with Yves Achdou & Frédéric Hecht. This work concerned with a posteriori error estimates in the energy norm for the numerical solutions of parabolic obstacle problems allowing for space/time mesh adaptive refinement. These estimates are based on a posteriori error indicators which can be computed from the solution of the discrete problem. Good error indicators should have two properties : *reliability* and *efficiency*. Reliability means that the error between the solutions of the discrete and continuous problems (which is of course not available) can be bounded from above by the error indicators. Efficiency means that the error indicator provides a lower estimate for the error, or in other words that the previously mentioned upper bounds for the error are sharp. If the error indicators provide a local lower bound for the local error then they can be used for designing an adaptive mesh refinement/coarsening strategy.
The present work can be seen as an effort to combine some existing strategies for elliptic obstacle problems on the one hand and parabolic equations on the other hand :

Elliptic obstacle problems : this work owes a great deal to the existing literature, in particular to the nice works by Chen and Nochetto [CN00], Veeser[Vee01] and Nochetto, Siebert and Veeser [NSV03, NSV05]. In particular, the definition of a multiplier for the discrete obstacle problem will be central, and for this, we will use the ideas contained in [NSV03, NSV05]. Note that there were previous works on multilevel adaptive methods for elliptic variational inequalities e.g. by Hoppe and Kornhuber [HK94] and by Kornhuber [Kor96]. A posteriori error estimates for a penalty approach were presented in [Joh92].

Parabolic equations : we were inspired by the study of a posteriori error estimates for parabolic equations by Bergam, Bernardi and Mghazli [BBM05] on Euler implicit time schemes combined with finite element spatial discretization, where the errors coming from the time and spatial discretization were estimated separately. We consider two families of error indicators, both of residual type. The first family is global with respect to the space variable and local with respect to time : it gives relevant information in order to refine the grid in the time variable. The second family is local with respect to both space and time variables, and provides an efficient tool for mesh adaption in the space variable at each time step. As it will be seen below, some difficulties (some of them already present in [BBM05]) will come from the fact that the spatial mesh may depend on time.

Note that this approach is different from the one of Eriksson, Johnson see e.g. [EJ91, EJ95], where space-time finite element methods were used.

Finally, we refer to the book by Verfürth [Ver96] for a review on a posteriori error estimates and mesh adaptivity.

The paper is organized as follows : the continuous problem is presented below. The discrete problem is proposed in Section 10.2. In particular, the important concepts of discrete full contact zone and multiplier for the discrete problems are presented in § 10.2.2.2. In Section 10.3, we consider the case when the obstacle function belongs to all the finite element spaces (they are several such spaces because the spatial mesh evolves in time) ; we propose error indicators and prove their reliability and efficiency. The case when the obstacle function does not belong to the discrete spaces is studied in Section 10.4. Numerical tests are presented in Section 10.5. Finally, we present (with no proof) the indicators for the slightly different variational inequality obtained in the context of the pricing of an American option on a two assets basket using the model of Black and Scholes. Related numerical results are given as well.

10.1 The obstacle problem

For simplicity, we consider two−dimensional problems, but what follows can be generalized.

We first introduce the continuous obstacle problem. Let Ω be a polygonal domain in $\mathbb{R}^2$. Let $\chi \in H^1(\Omega) \cap \mathcal{C}^0(\overline{\Omega})$ be a lower obstacle such that $\chi \leq 0$ on $\partial\Omega$. We call $\mathbb{K}$ the closed and convex subset of $H_0^1(\Omega)$:

$$\mathbb{K} = \{v \in H_0^1(\Omega),\ v \geq \chi \text{ a.e. in } \Omega\}, \tag{10.1}$$

and we introduce $\mathcal{K}$:

$$\mathcal{K} = \{v \in L^2(0,T; H_0^1(\Omega)) \text{ s.t. } v(t) \in \mathbb{K} \text{ for a.a. } t \in (0,T)\}. \tag{10.2}$$

For the data $f \in L^2(\Omega)$ and $u_0 \in L^2(\Omega)$, we consider the parabolic obstacle problem :

find $u \in \mathcal{K} \cap C^0([0,T]; L^2(\Omega))$, such that $\frac{\partial u}{\partial t} \in L^2(0,T; H^{-1}(\Omega))$ and $u(t=0) = u_0$ and satisfying for a.a. $t \in (0,T)$,

$$\left\langle \frac{\partial u}{\partial t}(t), v - u(t) \right\rangle + a(u(t), v - u(t)) \geq \int_\Omega f(v - u(t)), \qquad \forall v \in \mathbb{K}, \tag{10.3}$$

where $\langle , \rangle$ is the duality pairing between $H^{-1}(\Omega)$ and $H_0^1(\Omega)$ and a is the bilinear form

$$a(w,v) = \int_\Omega \nabla v \cdot \nabla w.$$

We will also use the notations $(u,v) = \int_\Omega u(x)v(x)dx$ for the inner product in $L^2(\Omega)$, $\|u\|_{L^2(\Omega)} = \sqrt{(u,u)}$, $\|u\|_{H^1(\Omega)} = \left(\int_\Omega |\nabla u|^2 + \int_\Omega u^2\right)^{1/2}$ and $|u|_{H^1(\Omega)} = \left(\int_\Omega |\nabla u|^2\right)^{1/2}$. For $\phi \in H^{-1}(\Omega)$, the notation $\|\phi\|_{H^{-1}(\Omega)}$ will stand for

$$\|\phi\|_{H^{-1}(\Omega)} = \sup_{0 \neq v \in H_0^1(\Omega)} \frac{\langle \phi, v \rangle}{|v|_{H^1(\Omega)}}.$$

From the theory of variational inequalities, we know that there exists a unique $\mu \in L^2(0,T;H^{-1}(\Omega))$, such that $\mu \geq 0$, $\int_0^T \langle \mu(t), u(t)-\chi\rangle dt = 0$ and for a.a. $t \in (0,T)$,

$$\left\langle \frac{\partial u}{\partial t}(t), v \right\rangle + a(u(t),v) = (f,v) + \langle \mu(t), v\rangle, \qquad \forall v \in H_0^1(\Omega). \tag{10.4}$$

One can view μ as a multiplier for the constraint $u \geq \chi$. We have $\mu = -\Delta\chi - f$ in the sense of distributions in the interior of the contact set $\Lambda = \{u = \chi\}$ and $\mu = 0$ in the noncontact set $\{u > \chi\}$.

10.2 The Discrete Problem

10.2.1 The Time Semi-Discrete Problem

We introduce a partition of the interval $[0,T]$ into subintervals $[t_n, t_{n+1}]$, $0 \leq n \leq N-1$, such that $0 = t_0 < t_1 < \cdots < t_N = T$. Let $\Delta t_n = t_{n+1} - t_n$, $n = 0, \ldots, N-1$ be the time steps ; let Δt be the maximal time step, i.e. $\Delta t = \max_{0\leq n<N} \Delta t_n$. We also define the regularity parameter $\rho_{\Delta t}$:

$$\rho_{\Delta t} = \max_{1\leq n<N} \frac{\Delta t_n}{\Delta t_{n-1}}. \tag{10.5}$$

For a continuous function f on $[0,T]$, we introduce the notation $f^n = f(t_n)$, $n = 0, \ldots, N$. The semi-discrete problem arising from an implicit Euler scheme is :
find $(u^n)_{0\leq n\leq N}$ such that $u^0 = u_0$ and that for all $n \in \{1, \ldots, N\}$,

$$\begin{aligned} &u^n \in \mathbb{K}, \\ &\forall v \in \mathbb{K}, \quad \frac{1}{\Delta t_{n-1}} \left(u^n - u^{n-1}, v - u^n\right) + a(u^n, v-u^n) \geq \int_\Omega f(v-u^n). \end{aligned} \tag{10.6}$$

From the theory of variational inequalities, we know that for all $n \in \{1, \ldots, N\}$, there exists a unique distribution $\mu^n \in H^{-1}(\Omega)$ such that

$$\begin{aligned} \frac{1}{\Delta t_{n-1}} \left(u^n - u^{n-1}, v\right) + a(u^n, v) &= (f,v) + \langle \mu^n, v\rangle, \quad \forall v \in H_0^1(\Omega), \\ \langle \mu^n, u^n - \chi\rangle &= 0, \end{aligned} \tag{10.7}$$

and it is clear that $\mu^n \geq 0$. We call $u_{\Delta t}$ the function in $\mathcal{C}^0([0,T];L^2(\Omega))$ which is affine w.r.t. t on the intervals $[t_{n-1}, t_n]$, and such that $u_{\Delta t}(t_n) = u^n$ for $n = 0, \ldots, N$. We call $\mu_{\Delta t}$ the function in $L^2(0,T;H^{-1}(\Omega))$ which is constant w.r.t. t on the intervals $(t_{n-1}, t_n]$, and such that $\mu_{\Delta t}|_{(t_{n-1},t_n]} = \mu^n$.

10.2.2 The Fully Discrete Problem

10.2.2.1 Triangular Finite Elements

We now describe the full discretization of (10.3). For each n, $0 \leq n \leq N$, let $(\mathcal{T}_{n,h})_h$ be a family of triangular meshes of Ω with the classical shape regularity property, see [Cia78, Cia91] : for a given element $\omega \in \mathcal{T}_{n,h}$, let h_ω be the diameter of ω and ρ_ω be the maximal radius of a ball contained in ω ; there exists a positive constant γ such that for

all n, h, for all ω in $\mathcal{T}_{n,h}$, $h_\omega/\rho_\omega \leq \gamma$.
For each n, h, we define the discrete spaces by

$$V_{n,h} = \left\{ v_h \in H^1(\Omega),\ \forall \omega \in \mathcal{T}_{n,h}, v_{h|\omega} \in \mathcal{P}_1 \right\}, \qquad V^0_{n,h} = V_{n,h} \cap H^1_0(\Omega). \tag{10.8}$$

Assumption 10.1 *The meshes $\mathcal{T}_{n,h}$ for different values of n are not independent : indeed, each triangulation $\mathcal{T}_{n,h}$ is derived from $\mathcal{T}_{n-1,h}$ by cutting some elements of $\mathcal{T}_{n-1,h}$ into smaller triangles or on the contrary by gluing together elements of $\mathcal{T}_{n-1,h}$. As a consequence, the quantities (w^{n-1}_h, v^n_h) can be computed without errors if $w^{n-1}_h \in V_{n-1,h}$ and $v^n_h \in V_{n,h}$.*

Assumption 10.2 *We assume that f is such that for all n, h, $\int_\Omega f v_h$ can be computed exactly for all $v_h \in V_{n,h}$.*

Remark 10.0.1 *We make Assumption 10.2 only for simplicity. If it is not satisfied, we have to take additional quadrature errors into account.*

Let $\chi^n_h \in V_{n,h}$ be an approximation of χ such that $\chi^n_h \leq 0$ on $\partial\Omega$. Let us define the closed and non empty convex subset of $V^0_{n,h}$:

$$\mathbb{K}_{n,h} = \{ v_h \in V^0_{n,h} \text{ such that } v_h \geq \chi^n_h \}. \tag{10.9}$$

Assuming for simplicity that $u_0 \in V_{0,h}$, the fully discrete problem reads :
find $(u^n_h)_{0\leq n\leq N}$, such that $u^0_h = u_0$ and that for all $n \in \{1, \ldots, N\}$,

$$\begin{aligned} &u^n_h \in \mathbb{K}_{n,h}, \\ &\forall v_h \in \mathbb{K}_{n,h}, \quad \frac{1}{\Delta t_{n-1}} \left(u^n_h - u^{n-1}_h, v_h - u^n_h \right) + a(u^n_h, v_h - u^n_h) \geq \int_\Omega f(v_h - u^n_h). \end{aligned} \tag{10.10}$$

We call $u_{h,\Delta t}$ the function which is affine on each interval $[t_{n-1}, t_n]$, and such that $u_{h,\Delta t}(t_n) = u^n_h$ for $n = 0, \ldots, N$. We need to define a discrete analogue of μ^n for $n = 1, \ldots, N$.

10.2.2.2 Discrete Full Contact Zone and Multiplier

Here, we introduce the full contact zone : it is obtained by modifying the definition given by Nochetto et al in the case of an elliptic contact problem, see [NSV05]. We first introduce some notations :
Let $\mathcal{N}_{n,h}$ be the set of the nodes of $\mathcal{T}_{n,h}$, and let $\mathcal{N}^0_{n,h}$ be the set of interior nodes.
The nodal basis functions of $V_{n,h}$ are called $(\phi_z)_{z\in\mathcal{N}_{n,h}}$. They satisfy $\sum_{z\in\mathcal{N}_{n,h}} \phi_z = 1$. The nodal basis functions of $V^0_{n,h}$ are $(\phi_z)_{z\in\mathcal{N}^0_{n,h}}$. The Lagrange interpolation of a continuous function v on $V_{n,h}$ is $I_{n,h}v = \sum_{z\in\mathcal{N}_{n,h}} v(z)\phi_z$.
For $z \in \mathcal{N}_{n,h}$, let ω_z be the support of ϕ_z, which is the union of all the triangles in $\mathcal{T}_{n,h}$ sharing z as a vertex. Let h_z be the diameter of ω_z. For $z \in \mathcal{N}^0_{n,h}$, we also set $\rho_z = \max\{r > 0 : B(z, r) \subset \omega_z\}$,.
Let $\mathcal{E}_{n,h}$ be the set of the edges of $\mathcal{T}_{n,h}$ and $\mathcal{E}^0_{n,h}$ be the set of all interior edges. We denote the union of all the interelements boundaries of $\mathcal{T}_{n,h}$ by $\Gamma_{n,h}$: $\Gamma_{n,h} = \cup_{S\in\mathcal{E}^0_{n,h}} S$.
Let J^n_h be the jump of the normal derivative of u^n_h across the interelements boundary :

if $S \subset \Gamma_{n,h}$ is the common side of the two triangles κ^- and κ^+ in $\mathcal{T}_{n,h}$, then $J_h^n|_S = (\nabla u_h^n|_{\kappa_+} - \nabla u_h^n|_{\kappa_-}) \cdot \boldsymbol{n}$, where $\boldsymbol{n}$ is the unit normal vector to S pointing from κ_- to κ_+.
We also set $\rho_S = \max\{r > 0 : B(x_S, r) \subset S \cup \kappa_+ \cup \kappa_-\}$, where x_S is the midpoint of S.
For $z \in \mathcal{N}_{n,h}$, let γ_z be the union of all the interelement boundaries lying in ω_z : $\gamma_z = \Gamma_{n,h} \cap \omega_z$.
For $n = 1, \ldots, N$, we define the set $\mathcal{C}_{n,h}$ of the full contact nodes at t_n by :

$$\mathcal{C}_{n,h} = \left\{ z \in \mathcal{N}_{n,h} \text{ s.t. } \left| \begin{array}{l} u_h^n = \chi_h^n \text{ and } u_h^n - u_h^{n-1} \geq \Delta t_{n-1} f \text{ in } \omega_z, \\ J_h^n \leq 0 \text{ on } \gamma_z \end{array} \right. \right\}. \tag{10.11}$$

We define the discrete full contact zone $\Omega_{n,h}^0$ at t_n by :

$$\Omega_{n,h}^0 = \left\{ x \in \Omega \text{ s.t. } \sum_{z \in \mathcal{C}_{n,h}} \phi_z(x) = 1 \right\}, \tag{10.12}$$

and its complement $\Omega_{n,h}^+ = \Omega \backslash \Omega_{n,h}^0$. We set $\Gamma_{n,h}^0 = \Omega_{n,h}^0 \cap \Gamma_{n,h}$ and $\Gamma_{n,h}^+ = \Omega_{n,h}^+ \cap \Gamma_{n,h}$. The following observation will be useful :

$$z \in \mathcal{N}_{n,h} \backslash \mathcal{C}_{n,h} \Rightarrow \omega_z \subset \overline{\Omega_{n,h}^+} \text{ and } \gamma_z \subset \overline{\Gamma_{n,h}^+}. \tag{10.13}$$

We will use an interpolation operator $\Pi_{n,h} : L^1(\Omega) \to V_{n,h}^0$ in the family introduced by Chen and Nochetto (see [CN00],[NSV03]). These operators are positivity preserving and second order accurate. We rapidly describe the construction of $\Pi_{n,h}$ since this will be important in § 10.3.2 : given $z \in \mathcal{N}_{n,h}^0$ and $\phi \in L^1(\Omega)$, we choose $B(z)$ to be a ball contained in ω_z and define the nodal value

$$(\Pi_{n,h}\phi)(z) = \frac{1}{|B(z)|} \int_{B(z)} \phi. \tag{10.14}$$

The radius of $B(z)$ must be comparable to the diameter of ω_z for $\Pi_{n,h}$ to have the suitable stability and accuracy properties (see[CN00]). In § 10.3.2, we will investigate the choice of the radius.
Following Nochetto et al [NSV05] in the elliptic case, we define the discrete multiplier $\mu_h^n \in H^{-1}(\Omega)$ as follows : $\forall \varphi \in H_0^1(\Omega)$,

$$\begin{aligned} \langle \mu_h^n, \varphi \rangle &= \sum_{z \in \mathcal{N}_{n,h}} \langle \mu_h^n, \varphi \phi_z \rangle, \\ \langle \mu_h^n, \varphi \phi_z \rangle &= \left\{ \begin{array}{l} \displaystyle\int_{\Omega_{n,h}^0} \left(\frac{u_h^n - u_h^{n-1}}{\Delta t_{n-1}} - f\right) \varphi \phi_z - \int_{\Gamma_{n,h}^0} J_h^n \varphi \phi_z \\ \displaystyle + \int_{\Omega_{n,h}^+} (\Pi_{n,h}\varphi)(z) \left(\frac{u_h^n - u_h^{n-1}}{\Delta t_{n-1}} - f\right) \phi_z - \int_{\Gamma_{n,h}^+} (\Pi_{n,h}\varphi)(z) J_h^n \phi_z \end{array} \right. \end{aligned} \tag{10.15}$$

From (10.13), note that if $z \in \mathcal{N}_{n,h} \backslash \mathcal{C}_{n,h}$, then

$$\langle \mu_h^n, \varphi \phi_z \rangle = (\Pi_{n,h}\varphi)(z)\, m_z, \quad \text{where } m_z = \int_\Omega \left(\frac{u_h^n - u_h^{n-1}}{\Delta t_{n-1}} - f\right) \phi_z - \int_{\Gamma_{n,h}} J_h^n \phi_z. \tag{10.16}$$

Lemme 10.1 *For any $z \in \mathcal{N}_{n,h}$, for any $\varphi \in H_0^1(\Omega)$ such that $\varphi \geq 0$, we have $\langle \mu_h^n, \varphi\phi_z \rangle \geq 0$.*

Proof From the properties of $\Pi_{n,h}$, we know that $\Pi_{n,h}(\varphi) \geq 0$ and that $\Pi_{n,h}(\varphi) = 0$ on $\partial\Omega$.
Consider first $z \in (\mathcal{N}_{n,h} \backslash \mathcal{C}_{n,h}) \cap \partial\Omega$: it is clear that $\Pi_{n,h}(\varphi)(z) = 0$ and the conclusion follows from (10.16).
Take $z \in \mathcal{N}_{n,h}^0 \backslash \mathcal{C}_{n,h}$: it is enough to verify that m_z in (10.16) is nonnegative. Taking $u_h^n + \phi_z$ as a test function in the variational inequality satisfied by u_h^n, (see (10.10)) yields

$$\left(\frac{u_h^n - u_h^{n-1}}{\Delta t_{n-1}} - f, \phi_z \right) + a(u_h^n, \phi_z) \geq 0,$$

which combined with (10.16) gives the desired result.
Finally, take $z \in \mathcal{C}_{n,h}$: from the definition of $\mathcal{C}_{n,h}$,

$$\frac{u_h^n - u_h^{n-1}}{\Delta t_{n-1}} - f \geq 0 \quad \text{in } \omega_z, \quad \text{and} \quad J_h^n \leq 0 \quad \text{on } \gamma_z.$$

From this, we see that each of the four terms in the right hand side of (10.15) are nonnegative. This concludes the proof. ■

We will use the notation $\lesssim$ to indicate that there may arise constants in the estimates which depend only Ω, T and the shape regularity parameter γ common to all $\mathcal{T}_{n,h}$.

10.3 The case when $\chi \in V_{n,h}$. A Posteriori Error Estimates

In this part, we make the following assumption

Assumption 10.3 *We assume that $\chi \in V_{n,h}$, for all n, h.*

Assumption 10.3 says that χ is piecewise linear and that the meshes $\mathcal{T}_{n,h}$ are all chosen in such a way that the restriction of χ to any triangle of $\mathcal{T}_{n,h}$ is linear. This allows for choosing

$$\chi_h^n = \chi, \quad \forall n \in \{1, \ldots, N\}. \tag{10.17}$$

Clearly, $\mathbb{K}_{n,h} \subset \mathbb{K}$.

10.3.1 Reliability : Global Upper Bounds

We begin by evaluating the error between the solutions u of the continuous problem and $u_{\Delta t}$ of the semi-discrete problem.

Lemme 10.2 *For any $n \in \{1, \ldots, N\}$,*

$$\left(\begin{array}{l} \|u(t_n) - u^n\|_{L^2(\Omega)}^2 + \displaystyle\int_0^{t_n} |u_{\Delta t}(\tau) - u(\tau)|_{H^1(\Omega)}^2 d\tau \\ + \displaystyle\int_0^{t_n} \|\frac{\partial u_{\Delta t}}{\partial t}(\tau) - \frac{\partial u}{\partial t}(\tau) + \mu_{\Delta t}(\tau) - \mu(\tau)\|_{H^{-1}(\Omega)}^2 \end{array} \right) \leq 5 \sum_{p=0}^{n-1} \Delta t_p |u^p - u^{p+1}|_{H^1(\Omega)}^2. \tag{10.18}$$

Proof We introduce the Galerkin linear functional defined on $H_0^1(\Omega)$:

$$\langle \mathcal{G}_t, v\rangle = \left\langle \frac{\partial u_{\Delta t}}{\partial t}(t), v \right\rangle + a(u_{\Delta t}(t), v) - \langle \mu_{\Delta t}(t), v\rangle - (f, v), \quad \forall v \in H_0^1(\Omega). \tag{10.19}$$

From (10.7), we see that for $t \in (t_p, t_{p+1})$,

$$\langle \mathcal{G}_t, v\rangle = a(u_{\Delta t} - u^{p+1}, v) = -\frac{t_{p+1} - t}{\Delta t_p} a(u^{p+1} - u^p, v),$$

which yields that for $t \in (t_p, t_{p+1})$,

$$\|\mathcal{G}_t\|_{H^{-1}(\Omega)} \leq |u^{p+1} - u^p|_{H^1(\Omega)}. \tag{10.20}$$

We see from (10.4) that for a.a. $t \in (0, T)$,

$$\left\langle \frac{\partial u_{\Delta t}}{\partial t}(t) - \frac{\partial u}{\partial t}(t), v \right\rangle + a(u_{\Delta t}(t) - u(t), v) - \langle \mu_{\Delta t}(t) - \mu(t), v\rangle = \langle \mathcal{G}_t, v\rangle, \quad \forall v \in H_0^1(\Omega). \tag{10.21}$$

Taking $v(t) = u_{\Delta t}(t) - u(t)$ and integrating between t_p and t_{p+1} yields

$$\begin{aligned} &\frac{1}{2}\left(\|u^{p+1} - u(t_{p+1})\|^2_{L^2(\Omega)} - \|u^p - u(t_p)\|^2_{L^2(\Omega)}\right) \\ &+ \int_{t_p}^{t_{p+1}} |u_{\Delta t}(t) - u(t)|^2_{H^1(\Omega)} dt - \int_{t_p}^{t_{p+1}} \langle \mu_{\Delta t}(t) - \mu(t), u_{\Delta t}(t) - u(t)\rangle dt \\ &= \int_{t_p}^{t_{p+1}} \langle \mathcal{G}_t, u_{\Delta t}(t) - u(t)\rangle dt. \end{aligned} \tag{10.22}$$

It is easy to check from (10.4) and (10.7) that

$$\langle \mu_{\Delta t}(t) - \mu(t), u_{\Delta t}(t) - u(t)\rangle \geq 0.$$

Summing (10.22) and using the previous observation, we obtain that

$$\|u(t_n) - u^n\|^2_{L^2(\Omega)} + \int_0^{t_n} |u_{\Delta t}(t) - u(t)|^2_{H^1(\Omega)} dt \leq \int_0^{t_n} \|\mathcal{G}_t\|^2_{H^{-1}(\Omega)} dt. \tag{10.23}$$

On the other hand, (10.21) yields that, for a.a. $t \in (0, T)$,

$$\|\frac{\partial u_{\Delta t}}{\partial t}(t) - \frac{\partial u}{\partial t}(t) + \mu_{\Delta t}(t) - \mu(t)\|^2_{H^{-1}(\Omega)} \leq 2\|\mathcal{G}_t\|^2_{H^{-1}(\Omega)} + 2|u_{\Delta t}(t) - u(t)|^2_{H^1(\Omega)}. \tag{10.24}$$

Integrating (10.24) with respect to t,

$$\begin{aligned} &\int_0^{t_n} \|\frac{\partial u_{\Delta t}}{\partial t}(\tau) - \frac{\partial u}{\partial t}(\tau) + \mu_{\Delta t}(\tau) - \mu(\tau)\|^2_{H^{-1}(\Omega)} d\tau \\ &\leq 2\int_0^{t_n} \left(\|\mathcal{G}_\tau\|^2_{H^{-1}(\Omega)} + |u_{\Delta t}(\tau) - u(\tau)|^2_{H^1(\Omega)}\right) d\tau \\ &\leq 4\int_0^{t_n} \|\mathcal{G}_\tau\|^2_{H^{-1}(\Omega)} d\tau, \end{aligned} \tag{10.25}$$

where the last estimate comes from (10.23). We obtain (10.18) by combining (10.20), (10.23) and (10.25). ■

We now introduce the discrete Galerkin functional :

$$\langle \mathcal{G}_h^n, v\rangle = \frac{1}{\Delta t_{n-1}}\left(u_h^n - u_h^{n-1}, v\right) + a(u_h^n, v) - \langle \mu_h^n, v\rangle - (f, v), \quad \forall v \in H_0^1(\Omega). \tag{10.26}$$

Lemme 10.3

$$\begin{aligned}&\left(\begin{array}{l}\frac{3}{\Delta t_{n-1}}\|u_h^n - u^n\|^2_{L^2(\Omega)} + |u_h^n - u^n|^2_{H^1(\Omega)} + \frac{3}{\Delta t_{n-1}}\|u_h^n - u_h^{n-1} - u^n + u^{n-1}\|^2_{L^2(\Omega)}\\ +\|\frac{1}{\Delta t_{n-1}}\left(u_h^n - u_h^{n-1} - u^n + u^{n-1}\right) - \mu_h^n + \mu^n\|^2_{H^{-1}(\Omega)}\end{array}\right)\\ \le& 5\|\mathcal{G}_h^n\|^2_{H^{-1}(\Omega)} + \frac{3}{\Delta t_{n-1}}\|u_h^{n-1} - u^{n-1}\|^2_{L^2(\Omega)} + 6\left\langle \mu_h^n - \mu^n, u_h^n - u^n\right\rangle.\end{aligned} \tag{10.27}$$

Proof Clearly,

$$\langle \mathcal{G}_h^n, v\rangle = \frac{1}{\Delta t_{n-1}}\left(u_h^n - u_h^{n-1} - u^n + u^{n-1}, v\right) + a(u_h^n - u^n, v) - \langle \mu_h^n - \mu^n, v\rangle. \tag{10.28}$$

This yields that

$$\|\mathcal{G}_h^n\|_{H^{-1}(\Omega)} \le \|\frac{1}{\Delta t_{n-1}}\left(u_h^n - u_h^{n-1} - u^n + u^{n-1}\right) - \mu_h^n + \mu^n\|_{H^{-1}(\Omega)} + |u_h^n - u^n|_{H^1(\Omega)}. \tag{10.29}$$

From the identity

$$|u_h^n - u^n|^2_{H^1(\Omega)} = \langle \mathcal{G}_h^n, u_h^n - u^n\rangle - \left\langle \frac{1}{\Delta t_{n-1}}\left(u_h^n - u_h^{n-1} - u^n + u^{n-1}\right) - \mu_h^n + \mu^n, u_h^n - u^n\right\rangle,$$

we obtain that

$$|u_h^n - u^n|^2_{H^1(\Omega)} \le \|\mathcal{G}_h^n\|^2_{H^{-1}(\Omega)} - 2\left\langle \frac{1}{\Delta t_{n-1}}\left(u_h^n - u_h^{n-1} - u^n + u^{n-1}\right) - \mu_h^n + \mu^n, u_h^n - u^n\right\rangle. \tag{10.30}$$

On the other hand, (10.28) implies

$$\begin{aligned}&\|\frac{1}{\Delta t_{n-1}}\left(u_h^n - u_h^{n-1} - u^n + u^{n-1}\right) - \mu_h^n + \mu^n\|^2_{H^{-1}(\Omega)}\\ \le& 2\|\mathcal{G}_h^n\|^2_{H^{-1}(\Omega)} + 2|u_h^n - u^n|^2_{H^1(\Omega)}\\ \le& 4\|\mathcal{G}_h^n\|^2_{H^{-1}(\Omega)} - 4\left\langle \frac{1}{\Delta t_{n-1}}\left(u_h^n - u_h^{n-1} - u^n + u^{n-1}\right) - \mu_h^n + \mu^n, u_h^n - u^n\right\rangle\end{aligned}$$

where the last line comes from (10.30). Therefore, from this and (10.30),

$$\begin{aligned}&|u_h^n - u^n|^2_{H^1(\Omega)} + \|\frac{1}{\Delta t_{n-1}}\left(u_h^n - u_h^{n-1} - u^n + u^{n-1}\right) - \mu_h^n + \mu^n\|^2_{H^{-1}(\Omega)}\\ \le& 5\|\mathcal{G}_h^n\|^2_{H^{-1}(\Omega)} - 6\left\langle \frac{1}{\Delta t_{n-1}}\left(u_h^n - u_h^{n-1} - u^n + u^{n-1}\right) - \mu_h^n + \mu^n, u_h^n - u^n\right\rangle.\end{aligned} \tag{10.31}$$

From (10.31) and the well known identity

$$2\left(w^n - w^{n-1}, w^n\right) = \|w^n\|^2_{L^2(\Omega)} + \|w^n - w^{n-1}\|^2_{L^2(\Omega)} - \|w^{n-1}\|^2_{L^2(\Omega)}$$

applied to $w^n = u_h^n - u^n$ and $w^{n-1} = u^{n-1} - u_h^{n-1}$, we obtain (10.27). ■

Corollary 10.1 *For $n = 1, \dots, N$,*

$$3\|u_h^n - u^n\|^2_{L^2(\Omega)} + 3\sum_{p=1}^{n} \|u_h^p - u_h^{p-1} - u^p + u^{p-1}\|^2_{L^2(\Omega)} + \sum_{p=1}^{n} \Delta t_{p-1} |u_h^p - u^p|^2_{H^1(\Omega)} \quad (10.32)$$
$$+ \sum_{p=1}^{n} \Delta t_{p-1} \|\frac{1}{\Delta t_{p-1}} \left(u_h^p - u_h^{p-1} - u^p + u^{p-1}\right) - \mu_h^p + \mu^p\|^2_{H^{-1}(\Omega)}$$
$$\leq 5\sum_{p=1}^{n} \Delta t_{p-1} \|\mathcal{G}_h^p\|^2_{H^{-1}(\Omega)} + 6\sum_{p=1}^{n} \Delta t_{p-1} \left\langle \mu_h^p - \mu^p, u_h^p - u^p \right\rangle.$$

Lemme 10.4

$$\langle \mathcal{G}_h^n, v\rangle = \int_{\Omega_{n,h}^+} (v - \Pi_{n,h} v)(\frac{u_h^n - u_h^{n-1}}{\Delta t_{n-1}} - f) - \int_{\Gamma_{n,h}^+} (v - \Pi_{n,h} v) J_h^n. \quad (10.33)$$

Proof Since the family $(\phi_z)_{z \in \mathcal{N}_{n,h}}$ is a partition of unity,

$$\langle \mathcal{G}_h^n, v\rangle = \sum_{z \in \mathcal{N}_{n,h}} \left(\frac{1}{\Delta t_{n-1}} \left(u_h^n - u_h^{n-1}, v\phi_z\right) + a(u_h^n, v\phi_z) - \langle \mu_h^n, v\phi_z\rangle - (f, v\phi_z) \right).$$

From this and the definition of μ_h^n, see (10.15),

$$\langle \mathcal{G}_h^n, v\rangle = \sum_{z \in \mathcal{N}_{n,h}} \left(\begin{array}{l} a(u_h^n, v\phi_z) + \int_\Omega (\frac{u_h^n - u_h^{n-1}}{\Delta t_{n-1}} - f) v\phi_z \\ - \int_{\Omega_{n,h}^0} (\frac{u_h^n - u_h^{n-1}}{\Delta t_{n-1}} - f) v\phi_z + \int_{\Gamma_{n,h}^0} J_h^n v\phi_z \\ - \int_{\Omega_{n,h}^+} (\Pi_{n,h} v)(z) (\frac{u_h^n - u_h^{n-1}}{\Delta t_{n-1}} - f)\phi_z + \int_{\Gamma_{n,h}^+} (\Pi_{n,h} v)(z) J_h^n \phi_z \end{array} \right)$$
$$= \sum_{z \in \mathcal{N}_{n,h}} \left(\int_{\Omega_{n,h}^+} (v - (\Pi_{n,h} v)(z))(\frac{u_h^n - u_h^{n-1}}{\Delta t_{n-1}} - f)\phi_z - \int_{\Gamma_{n,h}^+} (v - (\Pi_{n,h} v)(z)) J_h^n \phi_z \right)$$
$$= \int_{\Omega_{n,h}^+} (v - \Pi_{n,h} v)(\frac{u_h^n - u_h^{n-1}}{\Delta t_{n-1}} - f) - \int_{\Gamma_{n,h}^+} (v - \Pi_{n,h} v) J_h^n,$$

where we have used the fact that $-\int_{\Gamma_{n,h}} J_h^n v\phi_z = a(u_h^n, v\phi_z)$ and where the last identity comes from the fact that $\sum_{z \in \mathcal{N}_{n,h}} \phi_z = 1$ and that $\Pi_{n,h} v = \sum_{z \in \mathcal{N}_{n,h}} (\Pi_{n,h} v)(z)\phi_z$. ■

Lemme 10.5

$$\|\mathcal{G}_h^n\|^2_{H^{-1}(\Omega)} \lesssim \sum_{z \in \mathcal{N}_{n,h}} \left(h_z^2 \|\frac{u_h^n - u_h^{n-1}}{\Delta t_{n-1}} - f\|^2_{L^2(\omega_z \cap \Omega_{n,h}^+)} + h_z \|J_h^n\|^2_{L^2(\gamma_z \cap \Gamma_{n,h}^+)} \right). \quad (10.34)$$

Proof From (10.33), we obtain that $\|\mathcal{G}_h^n\|_{H^{-1}(\Omega)} \leq I + II$, where

$$I = \sup_{0 \neq v \in H_0^1(\Omega)} \frac{\left| \int_{\Omega_{n,h}^+} (v - \Pi_{n,h} v)(\frac{u_h^n - u_h^{n-1}}{\Delta t_{n-1}} - f) \right|}{|v|_{H_0^1(\Omega)}} \quad \text{and } II = \sup_{0 \neq v \in H_0^1(\Omega)} \frac{\left| \int_{\Gamma_{n,h}^+} (v - \Pi_{n,h} v) J_h^n \right|}{|v|_{H_0^1(\Omega)}}.$$

Let us estimate I and II separately. For bounding I, we see from the properties of $\Pi_{n,h}$ that

$$\left|\int_{\Omega^+_{n,h}} (v-\Pi_{n,h}v)(\frac{u_h^n-u_h^{n-1}}{\Delta t_{n-1}}-f)\right| = \left|\sum_{z\in\mathcal{N}_{n,h}}\int_{\Omega^+_{n,h}\cap\omega_z}(v-\Pi_{n,h}v)(\frac{u_h^n-u_h^{n-1}}{\Delta t_{n-1}}-f)\phi_z\right|$$

$$\lesssim \sum_{z\in\mathcal{N}_{n,h}} \|\frac{u_h^n-u_h^{n-1}}{\Delta t_{n-1}}-f\|_{L^2(\omega_z\cap\Omega^+_{n,h})} h_z \|\nabla v\|_{L^2(\omega_z\cap\Omega^+_{n,h})}$$

$$\leq \left(\sum_{z\in\mathcal{N}_{n,h}} h_z^2 \|\frac{u_h^n-u_h^{n-1}}{\Delta t_{n-1}}-f\|^2_{L^2(\omega_z\cap\Omega^+_{n,h})}\right)^{1/2} \left(\sum_{z\in\mathcal{N}_{n,h}} \|\nabla v\|^2_{L^2(\omega_z\cap\Omega^+_{n,h})}\right)^{1/2}$$

$$\lesssim \left(\sum_{z\in\mathcal{N}_{n,h}} h_z^2 \|\frac{u_h^n-u_h^{n-1}}{\Delta t_{n-1}}-f\|^2_{L^2(\omega_z\cap\Omega^+_{n,h})}\right)^{1/2} |v|_{H^1_0(\Omega)}.$$

The same argument can be used for bounding II. Indeed,

$$\left|\int_{\Gamma^+_{n,h}} (v-\Pi_{n,h}v)J_h^n\right| = \left|\sum_{z\in\mathcal{N}_{n,h}}\int_{\Gamma^+_{n,h}\cap\gamma_z}(v-\Pi_{n,h}v)J_h^n\right|$$

$$\lesssim \sum_{z\in\mathcal{N}_{n,h}} h_z \|\nabla v\|_{L^2(\omega_z\cap\Omega^+_{n,h})} \|J_h^n\|_{L^2(\gamma_z\cap\Gamma^+_{n,h})}$$

$$\lesssim \left(\sum_{z\in\mathcal{N}_{n,h}} h_z \|J_h^n\|^2_{L^2(\gamma_z\cap\Gamma^+_{n,h})}\right)^{1/2} |v|_{H^1_0(\Omega)}$$

■

Remark 10.1.1 *All the results that have been proved so far, namely Lemma 10.3, Corollary 10.1, Lemmas 10.4 and 10.5, do not rely on Assumption 10.3 and (10.17), and hold for any choice of functions χ_h^n, $n = 1, \dots, N$.*

In view of (10.27), we are left with finding an upper bound for $\langle \mu_h^n - \mu^n, u_h^n - u^n\rangle$. Here Assumption 10.3 plays an important role :

Lemme 10.6 *Under Assumption 10.3 and with (10.17),*

$$\langle \mu_h^n - \mu^n, u_h^n - u^n\rangle$$

$$\lesssim \sum_{\substack{z\in\mathcal{N}^0_{n,h}\backslash\mathcal{C}_{n,h}\\ u_h^n(z)=\chi(z)}} \left(h_z(\|\widetilde{J_h^n}\|^2_{L^2(\gamma_z\cap\Gamma^+_{n,h})} + \|J_h^n\|^2_{L^2(\gamma_z\cap\Gamma^+_{n,h})}) + h_z^2 \left\|\frac{u_h^n-u_h^{n-1}}{\Delta t_{n-1}}-f\right\|^2_{L^2(\omega_z\cap\Omega^+_{n,h})}\right), \tag{10.35}$$

where $\widetilde{J_h^n}$ is the jump of the normal derivative of $u_h^n - \chi$ across the interelements boundary : if $S \subset \Gamma_{n,h}$ is the common side of the two triangles κ^- and κ^+ in $\mathcal{T}_{n,h}$, then $\widetilde{J_h^n}|_S = (\nabla(u_h^n-\chi)|_{\kappa_+} - \nabla(u_h^n-\chi)|_{\kappa_-})\cdot \boldsymbol{n}$, where $\boldsymbol{n}$ is the unit normal vector to S pointing from κ_- to κ_+.

Proof From the definition of μ^n, $-\langle \mu^n, u_h^n - u^n\rangle \leq 0$ because $\mathbb{K}_{n,h} \subset \mathbb{K}$ from Assumption 10.3 and (10.17).
We now consider the remaining term, i.e. $\langle \mu_h^n, u_h^n - u^n\rangle = \sum_{z\in\mathcal{N}_{n,h}} \langle \mu_h^n, \phi_z(u_h^n - u^n)\rangle$.
If $z \in \mathcal{C}_{n,h}$, then $u_h^n = \chi$ in ω_z and $\langle \mu_h^n, \phi_z(u_h^n - u^n)\rangle = \langle \mu_h^n, \phi_z(\chi - u^n)\rangle$. Thus $\langle \mu_h^n, \phi_z(u_h^n - u^n)\rangle \leq 0$, from Lemma 10.1 and the fact that $u^n \geq u_h^n$ in ω_z (because $u^n \in \mathbb{K}$).
Therefore

$$\langle \mu_h^n, u_h^n - u^n\rangle \leq \sum_{z\in\mathcal{N}_{n,h}\backslash\mathcal{C}_{n,h}} \langle \mu_h^n, \phi_z(u_h^n - u^n)\rangle = \sum_{z\in\mathcal{N}_{n,h}\backslash\mathcal{C}_{n,h}} m_z \Pi_{n,h}(u_h^n - u^n)(z),$$

where the last identity comes from (10.16). Moreover, from (10.13), the quantities m_z in the last sum are

$$m_z = \int_{\Omega_{n,h}^+\cap\omega_z} (\frac{u_h^n - u_h^{n-1}}{\Delta t_{n-1}} - f)\phi_z - \int_{\Gamma_{n,h}^+\cap\gamma_z} J_h^n \phi_z.$$

We first consider the contribution of the nodes $z \in \mathcal{N}_{n,h}\backslash\mathcal{C}_{n,h}$ such that $u_h^n(z) > \chi(z)$. If $z \notin \partial\Omega$, taking $u_h^n \pm \delta\phi_z$ with δ small enough as a test function in (10.10) yields that $\frac{1}{\Delta t_{n-1}}\left(u_h^n - u_h^{n-1}, \phi_z\right) + a(u_h^n, \phi_z) = \int_\Omega f\phi_z$, or in other words $m_z = 0$. On the other hand, if $z \in \partial\Omega$, $\Pi_{n,h}(u_h^n - u^n)(z) = 0$. We have proven that

$$\langle \mu_h^n, u_h^n - u^n\rangle \leq \sum_{z\in\mathcal{N}_{n,h}\backslash\mathcal{C}_{n,h}} m_z \Pi_{n,h}(u_h^n - u^n)(z) = \sum_{\substack{z\in\mathcal{N}_{n,h}^0\backslash\mathcal{C}_{n,h} \\ u_h^n(z)=\chi(z)}} m_z \Pi_{n,h}(u_h^n - u^n)(z).$$

Since $\Pi_{n,h}$ preserves positivity, we know that $\Pi_{n,h}(\chi - u^n) \leq 0$. Thus $\Pi_{n,h}(u_h^n - u^n) \leq \Pi_{n,h}(u_h^n - \chi)$, and since $m_z \geq 0$ for all $z \in \mathcal{N}_{n,h}^0\backslash\mathcal{C}_{n,h}$, we have that

$$\langle \mu_h^n, u_h^n - u^n\rangle \leq \sum_{\substack{z\in\mathcal{N}_{n,h}^0\backslash\mathcal{C}_{n,h} \\ u_h^n(z)=\chi(z)}} m_z \Pi_{n,h}(u_h^n - \chi)(z).$$

By using the fact that $u_h^n(z) = \chi(z)$ and Lemma 3.3 in [CN00] , we see that

$$|\Pi_{n,h}(u_h^n - \chi)(z)| \lesssim h_z^{1/2} \|\widetilde{J_h^n}\|_{L^2(\gamma_z)} = h_z^{1/2} \|\widetilde{J_h^n}\|_{L^2(\gamma_z\cap\Gamma_{n,h}^+)}. \tag{10.36}$$

On the other hand,

$$|m_z| = \left| \int_{\omega_z} (\frac{u_h^n - u_h^{n-1}}{\Delta t_{n-1}} - f)\phi_z - \int_{\gamma_z} J_h^n \phi_z \right| \tag{10.37}$$

$$\lesssim h_z \left\| \frac{u_h^n - u_h^{n-1}}{\Delta t_{n-1}} - f \right\|_{L^2(\omega_z\cap\Omega_{n,h}^+)} + h_z^{1/2} \|J_h^n\|_{L^2(\gamma_z\cap\Gamma_{n,h}^+)} \tag{10.38}$$

Combining this and (10.36) yields (10.35). ■

Proposition 10.7 *Under Assumption 10.3 and with (10.17), we have the a posteriori error estimate for the error between u^n and u_h^n, $n = 1, \dots, N$:*

$$\|u_h^n - u^n\|^2_{L^2(\Omega)} + \sum_{p=1}^{n} \Delta t_{p-1} |u_h^p - u^p|^2_{H^1(\Omega)} \tag{10.39}$$
$$+ \sum_{p=1}^{n} \Delta t_{p-1} \left\| \frac{1}{\Delta t_{p-1}} \left(u_h^p - u_h^{p-1} - u^p + u^{p-1} \right) - \mu_h^p + \mu^p \right\|^2_{H^{-1}(\Omega)} \lesssim \sum_{p=1}^{n} \Delta t_{p-1} \sum_{z \in \mathcal{N}_{p,h}} \eta_{p,z}^2,$$

where

$$\eta_{p,z}^2 = h_z^2 \left\| \frac{u_h^p - u_h^{p-1}}{\Delta t_{p-1}} - f \right\|^2_{L^2(\omega_z \cap \Omega_{p,h}^+)} + h_z \|J_h^p\|^2_{L^2(\gamma_z \cap \Gamma_{p,h}^+)} \tag{10.40}$$
$$+ h_z 1_{\left\{ \begin{array}{l} z \in \mathcal{N}_{p,h}^0 \backslash \mathcal{C}_{p,h} \\ u_h^p(z) = \chi(z) \end{array} \right\}} \|\widetilde{J_h^p}\|^2_{L^2(\gamma_z \cap \Gamma_{p,h}^+)}.$$

Proof The result is a direct consequence of Corollary 10.1 and Lemmas 10.5 and 10.6. ■

Théorème 10.8 (reliability) *Under Assumption 10.3 and with (10.17), we have the a posteriori error estimate for the error between u and $u_{h,\Delta t}$: for any $n \in \{1, \dots, N\}$,*

$$\|u(t_n) - u_h^n\|^2_{L^2(\Omega)} + \int_0^{t_n} |u_{h,\Delta t}(\tau) - u(\tau)|^2_{H^1(\Omega)} d\tau \tag{10.41}$$
$$+ \int_0^{t_n} \left\| \frac{\partial u_{h,\Delta t}}{\partial t}(\tau) - \frac{\partial u}{\partial t}(\tau) + \mu_{h,\Delta t}(\tau) - \mu(\tau) \right\|^2_{H^{-1}(\Omega)} \lesssim \sum_{p=1}^{n} \Delta t_{p-1} \left(\eta_p^2 + \sum_{z \in \mathcal{N}_{p,h}} \eta_{p,z}^2 \right),$$

where

$$\eta_p^2 = |u_h^p - u_h^{p-1}|^2_{H^1(\Omega)}, \tag{10.42}$$

and $\eta_{p,z}$ is given by (10.40).

Proof The result is a direct consequence of Lemma 10.2 and Proposition 10.7. ■

Theorem 10.8 tells us that the local indicators η_p and $\eta_{p,z}$ are reliable because they can be used for bounding the error between the solutions of the continuous and discrete problems. Note that these indicators are respectively local w.r.t. time and space-time and that they can actually be computed because they only depend on the discrete solutions. If the estimate (10.41) is optimal, then it can be used in a mesh refinement/coarsening strategy. The paragraph below is devoted to showing that (10.41) is in some sense optimal.

10.3.2 Efficiency : Local Lower Bounds

We begin with giving an upper bound for η_p, see (10.42).

Lemme 10.9 *For any $p \in \{1, \dots, N\}$,*

$$\Delta t_{p-1}\eta_p^2 \lesssim \Delta t_{p-1}\left(|u_h^p - u^p|^2_{H^1(\Omega)} + |u^{p-1} - u_h^{p-1}|^2_{H^1(\Omega)}\right) + \int_{t_{p-1}}^{t_p} |u_{\Delta t}(\tau) - u(\tau)|^2_{H^1(\Omega)} + \int_{t_{p-1}}^{t_p} \left\| \frac{\partial u_{\Delta t}}{\partial t}(\tau) - \frac{\partial u}{\partial t}(\tau) + \mu_{\Delta t}(\tau) - \mu(\tau) \right\|^2_{H^{-1}(\Omega)}. \tag{10.43}$$

Proof

$$\begin{aligned} \Delta t_{p-1}\eta_p^2 =& \Delta t_{p-1}|u_h^p - u_h^{p-1}|^2_{H^1(\Omega)} \\ \leq& 3\Delta t_{p-1}\left(|u_h^p - u^p|^2_{H^1(\Omega)} + |u^p - u^{p-1}|^2_{H^1(\Omega)} + |u^{p-1} - u_h^{p-1}|^2_{H^1(\Omega)}.\right) \end{aligned} \tag{10.44}$$

Let us find a bound for $\Delta t_{p-1}|u^p - u^{p-1}|^2_{H^1(\Omega)}$:

$$\begin{aligned} &\Delta t_{p-1}|u^p - u^{p-1}|^2_{H^1(\Omega)} = \Delta t_{p-1} a(u^p - u^{p-1}, u^p - u^{p-1}) \\ =&2\int_{t_{p-1}}^{t_p} a(u_{\Delta t} - u^p, u^{p-1} - u^p) \\ =&2\int_{t_{p-1}}^{t_p} a(u_{\Delta t} - u, u^{p-1} - u^p) + 2\int_{t_{p-1}}^{t_p} a(u - u^p, u^{p-1} - u^p) \\ =&2\int_{t_{p-1}}^{t_p} a(u_{\Delta t} - u, u^{p-1} - u^p) - 2\int_{t_{p-1}}^{t_p} \left\langle \frac{\partial u}{\partial t} - \mu - f, u^{p-1} - u^p \right\rangle \\ &+ 2\int_{t_{p-1}}^{t_p} \left\langle \frac{\partial u_{\Delta t}}{\partial t} - \mu_{\Delta t} - f, u^{p-1} - u^p \right\rangle \\ \leq&2|u^p - u^{p-1}|^2_{H^1(\Omega)} \int_{t_{p-1}}^{t_p} \left(|u_{\Delta t} - u|_{H^1(\Omega)} + \left\| \frac{\partial u}{\partial t} - \mu - \frac{\partial u_{\Delta t}}{\partial t} + \mu_{\Delta t} \right\|_{H^{-1}(\Omega)} \right), \end{aligned}$$

and the desired result follows easily. ■

For finding local upper bounds for the error indicators $\eta_{p,z}$, $p = 1, \dots, N$ $z \in \mathcal{N}^0_{p,h}$, we need a further assumption which has also been made by Bernardi et al [BBM05] :

Assumption 10.4 *For $n = 1, \dots, N$, there exists a regular family of triangulations $(\mathcal{T}^*_{n,h})_h$ such that for all h and n each triangle of $\mathcal{T}_{n,h}$ and of $\mathcal{T}_{n-1,h}$ is the union of at most s triangles of $\mathcal{T}^*_{n,h}$, (where s is bounded independently of h and n).*

Assumption 10.4 says that the meshes corresponding to two successive time steps must not differ too much.

As a consequence of Assumption 10.4, there exists a number $\delta_0 \in (0, 1)$ independent of n and h such that, for all triangle κ of $\mathcal{T}_{n,h}$, if $\widehat{\kappa}$ is the retraction of κ with factor δ_0 with respect to the barycenter of κ, then $\widehat{\kappa}$ intersects all the triangles $\kappa^* \in \mathcal{T}^*_{n,h}$ such that $\kappa^* \subset \kappa$, and $\widehat{\kappa} \cap \kappa^*$ contains a ball whose diameter is greater than ch_κ for a fixed constant c. For all edge $S \subset \Gamma_{n,h}$, (S is the side common to $\kappa_{S,1}$ and $\kappa_{S,2}$, x_S is the barycenter of S), we introduce $\widetilde{\kappa}_{S,i}$, the retraction of $\kappa_{S,i}$ with factor δ_0 with respect to x_S.

Moreover, there exists a constant $\delta_1 \in (0, 1)$ independent of n and h such that for all $z \in \mathcal{N}_h$,

– $B(z, \delta_1 \rho_z) \cap \widehat{\kappa} = \emptyset$, for all $\kappa \in \mathcal{T}_{n,h}$,
– $B(z, \delta_1 \rho_z) \cap \widetilde{\kappa}_{S,i} = \emptyset$, for all $S \in \mathcal{E}^0_{n,h}$ and $(\widetilde{\kappa}_{S,i})_{i=1,2}$ constructed as above.

We choose $B(z) = B(z, \delta_1 \rho_z)$ in (10.14).
For each element $\kappa \in \mathcal{T}_{n,h}$, let b_κ be the bubble function such that

$$b_\kappa(x) = \begin{cases} 0 & \text{if } x \notin \widehat{\kappa}, \\ \prod_{i=1}^3 \widehat{\lambda}_i(x) & \text{if } x \in \widehat{\kappa}, \end{cases}$$

where $\widehat{\lambda}_i$ are the barycentric coordinates related to $\widehat{\kappa}$.
This construction implies that for all function $\phi \in L^1(\Omega)$ and for all $\kappa \in \mathcal{T}_{n,h}$,

$$\Pi_{n,h}(\phi b_\kappa) = 0. \tag{10.45}$$

Similarly, for each $S \in \mathcal{E}^0_{n,h}$, ($S$ is the side common to κ and κ', x_S is the barycenter of S), let b_S be the bubble function such that

$$b_S(x) = \begin{cases} 0 & \text{if } x \notin \widetilde{\kappa}_{S,1} \cup \widetilde{\kappa}_{S,2}, \\ \prod_{k=1}^2 \widetilde{\lambda}^2_{1,k} \widetilde{\lambda}^2_{2,k} & \text{if } x \in \widetilde{\kappa}_{S,1} \cup \widetilde{\kappa}_{S,2}, \end{cases}$$

where $\widetilde{\lambda}_{i,k}$ are the barycentric coordinates related to $\widetilde{\kappa}_{S,i}$, with a numbering such that $\widetilde{\lambda}_{i,3}$ corresponds to the node not contained in S.
Here also, for all function $\phi \in L^1(\Omega)$ and for all $S \in \mathcal{E}^0_{n,h}$,

$$\Pi_{n,h}(\phi b_S) = 0. \tag{10.46}$$

Following the work of Verfürth (Lemma 3.3. in [Ver96]), there exists a constant c independent of n and h such that

– For all $\kappa \in \mathcal{T}_{n,h}$,
  – for all $\phi_\kappa \in \mathcal{P}_1(\kappa)$,

$$\|b_\kappa \phi_\kappa\|_{L^2(\kappa)} \leq \|\phi_\kappa\|_{L^2(\kappa)} \leq c \|\sqrt{b_\kappa} \phi_\kappa\|_{L^2(\kappa)}, \tag{10.47}$$

$$\|b_\kappa \phi_\kappa\|_{H^1(\kappa)} \leq c h_\kappa^{-1} \|\phi_\kappa\|_{L^2(\kappa)}. \tag{10.48}$$

  – for all $\kappa^* \in \mathcal{T}^*_{n,h}$ such that $\kappa^* \subset \kappa$, for all $\phi_{\kappa^*} \in \mathcal{P}_1(\kappa^*)$,

$$\|b_\kappa \phi_{\kappa^*}\|_{L^2(\kappa^*)} \leq \|\phi_{\kappa^*}\|_{L^2(\kappa^*)} \leq c \|\sqrt{b_\kappa} \phi_{\kappa^*}\|_{L^2(\kappa^*)}, \tag{10.49}$$

$$\|b_\kappa \phi_{\kappa^*}\|_{H^1(\kappa^*)} \leq c h_\kappa^{-1} \|\phi_{\kappa^*}\|_{L^2(\kappa^*)}. \tag{10.50}$$

– For all $S \in \mathcal{E}^0_{n,h}$, ($S$ is the side common to $\kappa_{S,1}$ and $\kappa_{S,2}$, x_S is the barycenter of S),

$$c h_S \leq \int_S b_S, \tag{10.51}$$

$$\|b_S\|_{H^1(\kappa_{S,i})} \leq c h_{\kappa_{S,i}}^{-1} \|b_S\|_{L^2(\kappa_{S,i})} \lesssim 1, \qquad i = 1, 2, \tag{10.52}$$

  where h_S is the diameter of S.

Lemme 10.10 *Under Assumption 10.4, for all $z \in \mathcal{N}^0_{p,h}$, $p = 1, \ldots, N$,*

$$h_z^2 \left\| \frac{u_h^p - u_h^{p-1}}{\Delta t_{p-1}} - f \right\|^2_{L^2(\omega_z \cap \Omega^+_{p,h})} \lesssim \|\mathcal{G}_h^p\|^2_{H^{-1}(\omega_z \cap \Omega^+_{p,h})} + h_z^2 \sum_{\kappa \in \mathcal{T}_{p,h}, \kappa \subset \omega_z \cap \Omega^+_{p,h}} \|f - \bar{f}_\kappa\|^2_{L^2(\kappa)}. \tag{10.53}$$

where $\bar{f}_\kappa$ is the average of f in κ.

Proof

$$h_z^2\left\|\frac{u_h^p-u_h^{p-1}}{\Delta t_{p-1}}-f\right\|^2_{L^2(\omega_z\cap\Omega^+_{p,h})}=h_z^2\sum_{\kappa\in\mathcal{T}_{p,h},\kappa\subset\omega_z\cap\Omega^+_{p,h}}\left\|\frac{u_h^p-u_h^{p-1}}{\Delta t_{p-1}}-f\right\|^2_{L^2(\kappa)}$$

$$\leq 2h_z^2\sum_{\kappa\in\mathcal{T}_{p,h},\kappa\subset\omega_z\cap\Omega^+_{p,h}}\sum_{\kappa^*\in\mathcal{T}^*_{p,h},\kappa^*\subset\kappa}\left\|\frac{u_h^p-u_h^{p-1}}{\Delta t_{p-1}}-\bar{f}_\kappa\right\|^2_{L^2(\kappa^*)}+2h_z^2\sum_{\kappa\in\mathcal{T}_{p,h},\kappa\subset\omega_z\cap\Omega^+_{p,h}}\|f-\bar{f}_\kappa\|^2_{L^2(\kappa)}.$$

Let us focus on $\|\frac{u_h^p-u_h^{p-1}}{\Delta t_{p-1}}-\bar{f}_\kappa\|^2_{L^2(\kappa^*)}$ From (10.49), we know that

$$\left\|\frac{u_h^p-u_h^{p-1}}{\Delta t_{p-1}}-\bar{f}_\kappa\right\|^2_{L^2(\kappa^*)}\lesssim\int_{\kappa^*}\left|\frac{u_h^p-u_h^{p-1}}{\Delta t_{p-1}}-\bar{f}_\kappa\right|^2 b_\kappa$$

$$\lesssim\int_{\kappa^*}\left|\frac{u_h^p-u_h^{p-1}}{\Delta t_{p-1}}-f\right|^2 b_\kappa+\int_{\kappa^*}|f-\bar{f}_\kappa|^2.$$

The first term is bounded as follows :

$$\int_{\kappa^*}\left|\frac{u_h^p-u_h^{p-1}}{\Delta t_{p-1}}-f\right|^2 b_\kappa\lesssim\int_{\kappa^*}(\frac{u_h^p-u_h^{p-1}}{\Delta t_{p-1}}-f)(\frac{u_h^p-u_h^{p-1}}{\Delta t_{p-1}}-\bar{f}_\kappa)b_\kappa$$

$$+\|f-\bar{f}_\kappa\|_{L^2(\kappa^*)}\|\frac{u_h^p-u_h^{p-1}}{\Delta t_{p-1}}-f\|_{L^2(\kappa^*)}.$$

Focusing on the first term, we know that

$$\sum_{\kappa^*\in\mathcal{T}^*_{p,h},\kappa^*\subset\kappa}\int_{\kappa^*}(\frac{u_h^p-u_h^{p-1}}{\Delta t_{p-1}}-f)(\frac{u_h^p-u_h^{p-1}}{\Delta t_{p-1}}-\bar{f}_\kappa)b_\kappa=-a(u_h^p,(\frac{u_h^p-u_h^{p-1}}{\Delta t_{p-1}}-\bar{f}_\kappa)b_\kappa)$$

$$+\left\langle\mu_h^p,(\frac{u_h^p-u_h^{p-1}}{\Delta t_{p-1}}-\bar{f}_\kappa)b_\kappa\right\rangle+\left\langle\mathcal{G}_h^p,(\frac{u_h^p-u_h^{p-1}}{\Delta t_{p-1}}-\bar{f}_\kappa)b_\kappa\right\rangle.$$

But from the fact that $\kappa\subset\Omega^+_{p,h}$ and from (10.15), (10.45), we see that $\left\langle\mu_h^p,(\frac{u_h^p-u_h^{p-1}}{\Delta t_{p-1}}-\bar{f}_\kappa)b_\kappa\right\rangle=0$. Integrating by part, we also see that $a(u_h^p,(\frac{u_h^p-u_h^{p-1}}{\Delta t_{p-1}}-\bar{f}_\kappa)b_\kappa)=0$. Therefore,

$$\sum_{\kappa^*\in\mathcal{T}^*_{p,h},\kappa^*\subset\kappa}\int_{\kappa^*}(\frac{u_h^p-u_h^{p-1}}{\Delta t_{p-1}}-f)(\frac{u_h^p-u_h^{p-1}}{\Delta t_{p-1}}-\bar{f}_\kappa)b_\kappa=\left\langle\mathcal{G}_h^p,(\frac{u_h^p-u_h^{p-1}}{\Delta t_{p-1}}-\bar{f}_\kappa)b_\kappa\right\rangle$$

$$\leq\|\mathcal{G}_h^p\|_{H^{-1}(\kappa)}\left|(\frac{u_h^p-u_h^{p-1}}{\Delta t_{p-1}}-\bar{f}_\kappa)b_\kappa\right|_{H^1_0(\kappa)}$$

$$\leq\|\mathcal{G}_h^p\|_{H^{-1}(\kappa)}\left(\sum_{\kappa^*\in\mathcal{T}^*_{p,h},\kappa^*\subset\kappa}\left|(\frac{u_h^p-u_h^{p-1}}{\Delta t_{p-1}}-\bar{f}_\kappa)b_\kappa\right|^2_{H^1(\kappa^*)}\right)^{1/2}$$

$$\lesssim h_\kappa^{-1}\|\mathcal{G}_h^p\|_{H^{-1}(\kappa)}\left\|\frac{u_h^p-u_h^{p-1}}{\Delta t_{p-1}}-\bar{f}_\kappa\right\|_{L^2(\kappa)},$$

where we have used (10.50). The conclusion follows easily by combining the above estimates, and noticing that $h_\kappa \simeq h_z$. ■

Lemme 10.11 *Under Assumption 10.4, for all $z \in \mathcal{N}^0_{p,h}$, $p = 1, \dots, N$,*

$$h_z \|J_h^p\|^2_{L^2(\gamma_z \cap \Gamma^+_{p,h})} \lesssim \|\mathcal{G}_h^p\|^2_{H^{-1}(\omega_z \cap \Omega^+_{p,h})} + h_z^2 \sum_{\kappa \in \mathcal{T}_{p,h}, \kappa \subset \omega_z \cap \Omega^+_{p,h}} \|f - \bar{f}_\kappa\|^2_{L^2(\kappa)}. \tag{10.54}$$

Proof

$$h_z \|J_h^p\|^2_{L^2(\gamma_z \cap \Gamma^+_{p,h})} = h_z \sum_{S \in \mathcal{E}^0_{p,h}, S \subset \gamma_z \cap \Gamma^+_{p,h}} \|J_h^p\|^2_{L^2(S)} \lesssim h_z \sum_{S \in \mathcal{E}^0_{p,h}, S \subset \gamma_z \cap \Gamma^+_{p,h}} \left\| J_h^p b_S^{\frac{1}{2}} \right\|^2_{L^2(S)},$$

because J_h^p is constant on each edge S. We focus on a single edge S such that $S \subset \gamma_z \cap \Gamma^+_{p,h}$, S is the side common to $\kappa_{S,1}$ and $\kappa_{S,2}$. Integrating by part and using the fact that $J_h^p|_S$ is constant,

$$\|J_h^p b_S^{\frac{1}{2}}\|^2_{L^2(S)} = J_h^p|_S \, a(u_h^p, b_S) = J_h^p|_S \left(- \int_\Omega (\frac{u_h^p - u_h^{p-1}}{\Delta t_{p-1}} - f) b_S + \left\langle \mu_h^p, b_S \right\rangle + \left\langle \mathcal{G}_h^p, b_S \right\rangle \right)$$

From (10.46), we see that $\left\langle \mu_h^p, b_S \right\rangle = 0$. It is also clear from (10.52) that

$$\left| \left\langle \mathcal{G}_h^p, b_S \right\rangle \right| \lesssim \|\mathcal{G}_h^p\|_{H^{-1}(\kappa_{S,1} \cup \kappa_{S,2})} \lesssim \|\mathcal{G}_h^p\|_{H^{-1}(\omega_z \cap \Omega^+_{p,h})}$$

and that

$$\begin{aligned} \left| \int_\Omega (\frac{u_h^p - u_h^{p-1}}{\Delta t_{p-1}} - f) b_S \right| &\leq \left\| \frac{u_h^p - u_h^{p-1}}{\Delta t_{p-1}} - f \|_{L^2(\omega_z \cap \Omega^+_{p,h})} \| b_S \right\|_{L^2(\omega_z \cap \Omega^+_{p,h})} \\ &\lesssim h_z \left\| \frac{u_h^p - u_h^{p-1}}{\Delta t_{p-1}} - f \right\|_{L^2(\omega_z \cap \Omega^+_{p,h})}. \end{aligned}$$

Finally, using (10.53) and combining the estimates above, we see that

$$h_z \|J_h^p\|^2_{L^2(S)} \lesssim h_z \left| J_h^p|_S \right| \left(\|\mathcal{G}_h^p\|^2_{H^{-1}(\omega_z \cap \Omega^+_{p,h})} + h_z^2 \sum_{\kappa \in \mathcal{T}_{p,h}, \kappa \subset \omega_z \cap \Omega^+_{p,h}} \|f - \bar{f}_\kappa\|^2_{L^2(\kappa)} \right)^{1/2},$$

which implies that

$$h_z^{1/2} \|J_h^p\|_{L^2(S)} \lesssim \left(\|\mathcal{G}_h^p\|^2_{H^{-1}(\omega_z \cap \Omega^+_{p,h})} + h_z^2 \sum_{\kappa \in \mathcal{T}_{p,h}, \kappa \subset \omega_z \cap \Omega^+_{p,h}} \|f - \bar{f}_\kappa\|^2_{L^2(\kappa)} \right)^{1/2},$$

and the desired result follows easily. ■

As a corollary of Lemma 10.11, we have the following :

Corollary 10.2 *For all $z \in \mathcal{N}^0_{p,h}\backslash\mathcal{C}_{p,h}$ such that $u^p_h(z) = \chi(z)$,*

$$h_z\|\widetilde{J^p_h}\|^2_{L^2(\gamma_z\cap\Gamma^+_{p,h})} \lesssim \|\mathcal{G}^p_h\|^2_{H^{-1}(\omega_z\cap\Omega^+_{p,h})} + h_z^2 \sum_{\kappa\in\mathcal{T}_{p,h},\kappa\subset\omega_z\cap\Omega^+_{p,h}} \|f-\bar{f}_\kappa\|^2_{L^2(\kappa)} + h_z\|[\frac{\partial\chi}{\partial n}]\|^2_{L^2(\gamma_z\cap\Gamma^+_{p,h})}, \tag{10.55}$$

where if S is the edge shared by κ_- and κ_+ with $S \subset \omega_z\cap\Omega^+_{p,h}$, $[\frac{\partial\chi}{\partial n}]|_S = (\nabla\chi|_{\kappa_+} - \nabla\chi|_{\kappa_-})\cdot\boldsymbol{n}$ and $\boldsymbol{n}$ is the unit normal vector to S pointing from κ_- to κ_+.

Conclusion on the efficiency In Lemmas 10.10, 10.11, and Corollary 10.2, the indicators are bounded by terms depending on the data f and on local norms of $\mathcal{G}^p_h$. On the other hand, from (10.28), we see that

$$\|\mathcal{G}^p_h\|_{H^{-1}(\omega_z\cap\Omega^+_{p,h})} \le \|\frac{1}{\Delta t_{p-1}}\left(u^p_h - u^{p-1}_h - u^p + u^{p-1}\right) - \mu^p_h + \mu^p\|_{H^{-1}(\omega_z\cap\Omega^+_{p,h})} + |u^p_h - u^p|_{H^1(\omega_z\cap\Omega^+_{p,h})}. \tag{10.56}$$

Therefore, from Lemmas 10.10, 10.11, and Corollary 10.2, obtain the following :

Corollary 10.3 *Under Assumption 10.4,*

- *for all $z \in \mathcal{N}^0_{p,h}$, $p = 1,\dots,N$,*

$$h_z^2\left\|\frac{u^p_h - u^{p-1}_h}{\Delta t_{p-1}} - f\right\|^2_{L^2(\omega_z\cap\Omega^+_{p,h})} \lesssim \left\|\frac{1}{\Delta t_{p-1}}\left(u^p_h - u^{p-1}_h - u^p + u^{p-1}\right) - \mu^p_h + \mu^p\right\|^2_{H^{-1}(\omega_z\cap\Omega^+_{p,h})} + |u^p_h - u^p|^2_{H^1(\omega_z\cap\Omega^+_{p,h})} + h_z^2 \sum_{\kappa\in\mathcal{T}_{p,h},\kappa\subset\omega_z\cap\Omega^+_{p,h}} \|f-\bar{f}_\kappa\|^2_{L^2(\kappa)}, \tag{10.57}$$

- *for all $z \in \mathcal{N}^0_{p,h}$, $p = 1,\dots,N$,*

$$h_z\|J^p_h\|^2_{L^2(\gamma_z\cap\Gamma^+_{p,h})} \lesssim \left(\begin{array}{l} \left\|\frac{1}{\Delta t_{p-1}}\left(u^p_h - u^{p-1}_h - u^p + u^{p-1}\right) - \mu^p_h + \mu^p\right\|^2_{H^{-1}(\omega_z\cap\Omega^+_{p,h})} \\ +|u^p_h - u^p|^2_{H^1(\omega_z\cap\Omega^+_{p,h})} + h_z^2 \sum_{\kappa\in\mathcal{T}_{p,h},\kappa\subset\omega_z\cap\Omega^+_{p,h}} \|f-\bar{f}_\kappa\|^2_{L^2(\kappa)}, \end{array}\right) \tag{10.58}$$

- *for all $z \in \mathcal{N}^0_{p,h}\backslash\mathcal{C}_{p,h}$ such that $u^p_h(z) = \chi(z)$,*

$$h_z\|\widetilde{J^p_h}\|^2_{L^2(\gamma_z\cap\Gamma^+_{p,h})} \lesssim \left(\begin{array}{l} \left\|\frac{1}{\Delta t_{p-1}}\left(u^p_h - u^{p-1}_h - u^p + u^{p-1}\right) - \mu^p_h + \mu^p\right\|^2_{H^{-1}(\omega_z\cap\Omega^+_{p,h})} \\ +|u^p_h - u^p|^2_{H^1(\omega_z\cap\Omega^+_{p,h})} \\ +h_z^2 \sum_{\kappa\in\mathcal{T}_{p,h},\kappa\subset\omega_z\cap\Omega^+_{p,h}} \|f-\bar{f}_\kappa\|^2_{L^2(\kappa)} + h_z\|[\frac{\partial\chi}{\partial n}]\|^2_{L^2(\gamma_z\cap\Gamma^+_{p,h})} \end{array}\right), \tag{10.59}$$

where $\bar{f}_\kappa$ is the average of f in κ.

Therefore, we can bound the indicators by local norms of the error between the solutions of the discrete and semi-discrete problems. On the other hand, we clearly have

$$|u_h^p - u^p|_{H^1(\omega_z \cap \Omega_{p,h}^+)} \leq |u_h^p - u(t_p)|_{H^1(\omega_z \cap \Omega_{p,h}^+)} + |u(t_p) - u^p|_{H^1(\omega_z \cap \Omega_{p,h}^+)}$$

and

$$\begin{aligned}
&\left\| u_h^p - u_h^{p-1} - u^p + u^{p-1} - \Delta t_{p-1}(\mu_h^p + \mu^p) \right\|_{H^{-1}(\omega_z \cap \Omega_{p,h}^+)} \\
&\leq \left(\begin{array}{l} \left\| u_h^p - u_h^{p-1} - \Delta t_{p-1}\mu_h^p - \displaystyle\int_{t_{p-1}}^{t_p} \left(\frac{\partial u}{\partial t}(\tau) + \mu(\tau) \right) d\tau \right\|_{H^{-1}(\omega_z \cap \Omega_{p,h}^+)} \\ + \displaystyle\int_{t_{p-1}}^{t_p} \left\| \frac{\partial u_{\Delta t}}{\partial t}(\tau) - \frac{\partial u}{\partial t}(\tau) + \mu_{\Delta t}(\tau) - \mu(\tau) \right\|_{H^{-1}(\omega_z \cap \Omega_{p,h}^+)} d\tau \end{array} \right),
\end{aligned}$$

so the indicators can be bounded by local norms of the errors between the solutions of the continuous and discrete/semi-discrete problems. In this respect, we can say that the error indicators are efficient.

Remark 10.3.1 *We have found optimal estimates for the error in some energy norms. It is possible to study other kinds of error, for example the pointwise error. This has been done in [NSV03] for an elliptic obstacle problem. Up to our knowledge, a posteriori pointwise error estimates for parabolic obstacle problems have not been studied yet.*

Remark 10.3.2 *In § 10.5, we will propose a mesh adaption strategy based on the estimates above ; given the error indicators, it will construct a new mesh which hopefully reduces the error. The question to know if the adapted mesh actually reduces the error is a very difficult one and we will not tackle it. Another open and difficult question is to study the convergence of the global adaption strategy : the answer to this question has been given in [BDD04] for an elliptic equation and a special method.*

10.4 The Case when $\chi \notin V_{n,h}$

Let us repeat the analysis above without Assumption 10.3. As seen in Remark 10.1.1, Lemma 10.3, Corollary 10.1, Lemmas 10.4 and 10.5 hold. Bounding $\langle \mu_h^n - \mu^n, u_h^n - u^n \rangle$ requires new arguments.

10.4.1 The Case when $\chi_h^n \geq \chi$, for $n = 1, \ldots, N$

Let us assume that for $n = 1, \ldots, N$, $\chi_h^n \geq \chi$. Therefore $\mathbb{K}_{n,h} \subset \mathbb{K}$ and the discretization is conforming. Constructing $\chi_h^n \in V_{N,h}$ such that $\chi_h^n \geq \chi$ can be done by modifying $I_{n,h}\chi$ locally. One can for example choose $\chi_h^n(z) = \chi(z) + \max_{x \in \bar{\omega}_z}(\chi(x) - I_{n,h}\chi(x))$ for $z \in \mathcal{N}_{n,h}$ such that $\bar{\omega}_z \cap \partial\Omega = \emptyset$. For the remaining vertices, a more careful construction has to be used so that $\chi_h^n \leq 0$ on $\partial\Omega$.

Lemme 10.12 *If $\chi_h^n \geq \chi$ for all n, h, then*

$$\begin{aligned}
&\langle \mu_h^n - \mu^n, u_h^n - u^n \rangle \\
&\lesssim \sum_{\substack{z \in \mathcal{N}_{n,h}^0 \setminus \mathcal{C}_{n,h} \\ u_h^n(z) = \chi_h^n(z)}} \left(h_z (\|\widetilde{J_h^n}\|^2_{L^2(\gamma_z \cap \Gamma_{n,h}^+)} + \|J_h^n\|^2_{L^2(\gamma_z \cap \Gamma_{n,h}^+)}) + h_z^2 \|\frac{u_h^n - u_h^{n-1}}{\Delta t_{n-1}} - f\|^2_{L^2(\omega_z \cap \Omega_{n,h}^+)} \right) \\
&\quad + \sum_{\substack{z \in \mathcal{N}_{n,h} \\ u_h^n(z) = \chi_h^n(z)}} \langle \mu_h^n, \phi_z(\chi_h^n - \chi) \rangle \,,
\end{aligned} \tag{10.60}$$

where $\widetilde{J_h^n}$ is the jump of the normal derivative of $u_h^n - \chi_h^n$ across the interelements boundary : if $S \subset \Gamma_{n,h}$ is the common side of the two triangles κ^- and κ^+ in $\mathcal{T}_{n,h}$, then $\widetilde{J_h^n}|_S = (\nabla(u_h^n - \chi_h^n)|_{\kappa_+} - \nabla(u_h^n - \chi_h^n)|_{\kappa_-}) \cdot \boldsymbol{n}$, where $\boldsymbol{n}$ is the unit normal vector to S pointing from κ_- to κ_+.

Proof We first see that $\langle \mu_h^n - \mu^n, u_h^n - u^n \rangle \leq \langle \mu_h^n, u_h^n - u^n \rangle$ because $\mathbb{K}_{n,h} \subset \mathbb{K}$. Then,

$$\langle \mu_h^n, u_h^n - u^n \rangle \leq \langle \mu_h^n, u_h^n - \chi \rangle = \sum_{z \in \mathcal{C}_{n,h}} \langle \mu_h^n, \phi_z(u_h^n - \chi) \rangle + \sum_{\substack{z \in \mathcal{N}_{n,h}^0 \setminus \mathcal{C}_{n,h} \\ u_h^n(z) = \chi_h^n(z)}} \langle \mu_h^n, \phi_z(u_h^n - \chi) \rangle$$

$$\leq \sum_{z \in \mathcal{C}_{n,h}} \langle \mu_h^n, \phi_z(\chi_h^n - \chi) \rangle + \sum_{\substack{z \in \mathcal{N}_{n,h}^0 \setminus \mathcal{C}_{n,h} \\ u_h^n(z) = \chi_h^n(z)}} \langle \mu_h^n, \phi_z(u_h^n - \chi_h^n) \rangle + \sum_{\substack{z \in \mathcal{N}_{n,h}^0 \setminus \mathcal{C}_{n,h}, \\ u_h^n(z) = \chi_h^n(z)}} \langle \mu_h^n, \phi_z(\chi_h^n - \chi) \rangle$$

$$= \sum_{\substack{z \in \mathcal{N}_{n,h} \\ u_h^n(z) = \chi_h^n(z)}} \langle \mu_h^n, \phi_z(\chi_h^n - \chi) \rangle + \sum_{\substack{z \in \mathcal{N}_{n,h}^0 \setminus \mathcal{C}_{n,h} \\ u_h^n(z) = \chi_h^n(z)}} \langle \mu_h^n, \phi_z(u_h^n - \chi_h^n) \rangle$$

The second term can be estimated by

$$\sum_{\substack{z \in \mathcal{N}_{n,h}^0 \setminus \mathcal{C}_{n,h} \\ u_h^n(z) = \chi_h^n(z)}} \left(h_z (\|\widetilde{J_h^n}\|^2_{L^2(\gamma_z \cap \Gamma_{n,h}^+)} + \|J_h^n\|^2_{L^2(\gamma_z \cap \Gamma_{n,h}^+)}) + h_z^2 \left\| \frac{u_h^n - u_h^{n-1}}{\Delta t_{n-1}} - f \right\|^2_{L^2(\omega_z \cap \Omega_{n,h}^+)} \right),$$

exactly as in the proof of Lemma 10.6. ■

Proposition 10.13 *Assuming that $\chi_h^n \geq \chi$ for all n, h, we have the a posteriori error estimate for the error between u^n and u_h^n, $n = 1, \ldots, N$:*

$$\begin{aligned}
&\left(\begin{array}{l} \|u_h^n - u^n\|^2_{L^2(\Omega)} + \sum_{p=1}^n \Delta t_{p-1} |u_h^p - u^p|^2_{H^1(\Omega)} \\ + \sum_{p=1}^n \Delta t_{p-1} \left\| \frac{1}{\Delta t_{p-1}} \left(u_h^p - u_h^{p-1} - u^p + u^{p-1} \right) - \mu_h^p + \mu^p \right\|^2_{H^{-1}(\Omega)} \end{array} \right) \\
&\lesssim \sum_{p=1}^n \Delta t_{p-1} \left(\sum_{z \in \mathcal{N}_{p,h}} \eta_{p,z}^2 + \sum_{\substack{z \in \mathcal{N}_{p,h} \\ u_h^p(z) = \chi_h^p(z)}} \langle \mu_h^p, \phi_z(\chi_h^p - \chi) \rangle \right),
\end{aligned} \tag{10.61}$$

where

$$\eta_{p,z}^2 = h_z^2 \left\| \frac{u_h^p - u_h^{p-1}}{\Delta t_{p-1}} - f \right\|^2_{L^2(\omega_z \cap \Omega^+_{p,h})} + h_z \|J_h^p\|^2_{L^2(\gamma_z \cap \Gamma^+_{p,h})} + h_z 1_{\left\{ \begin{array}{l} z \in \mathcal{N}^0_{p,h} \backslash \mathcal{C}_{p,h} \\ u_h^p(z) = \chi_h^p(z) \end{array} \right\}} \|\widetilde{J_h^p}\|^2_{L^2(\gamma_z \cap \Gamma^+_{p,h})} \tag{10.62}$$

Remark 10.3.3 *From Proposition 10.7 one can find an upper bound for the error between the solutions of the continuous and fully discrete problem.*

Remark 10.3.4 *Under Assumption 10.4, by using the definition of $\mathcal{G}_h^p$ and Lemmas 10.10 and 10.11, it is possible to prove that for $z \in \mathcal{N}^0_{p,h} \backslash \mathcal{C}_{p,h}$ such that $u_h^p(z) = \chi_h^p(z)$,*

$$\begin{aligned} \left| \langle \mu_h^p, \phi_z (u_h^p - \chi_h^p) \rangle \right| \lesssim \|\mathcal{G}_h^p\|^2_{H^{-1}(\omega_z \cap \Omega^+_{p,h})} + h_z^2 \sum_{\kappa \in \mathcal{T}_{p,h}, \kappa \subset \omega_z \cap \Omega^+_{p,h}} \|f - \bar{f}_\kappa\|^2_{L^2(\kappa)} \\ + \|\frac{1}{h_z}(\chi_h^p - \chi)\|^2_{L^2(\omega_z)} + \frac{1}{h_z}\|(\chi_h^p - \chi)\|^2_{L^2(\gamma_z)} + |\chi_h^p - \chi|^2_{H^1(\omega_z)}. \end{aligned}$$

This and Lemmas 10.10, 10.11, and Corollary 10.2 yield an upper bound for the error indicators in (10.61).

10.4.2 What Can be Said in the General Case ?

Here, we do not assume that $\chi_h^n \geq \chi$. In view of (10.27), we must find an upper bound for $\langle \mu_h^n - \mu^n, u_h^n - u^n \rangle$. We present an estimate that we have found, although it is not satisfactory because it requires a condition linking the time step and the measure of the set where $\chi > \chi_h^n$. In our opinion, this shows that it is better to choose $\chi_h^n \geq \chi$ when possible.
Let us introduce $u_h^{n*}(x) = \max(u_h^n(x), \chi(x))$. Clearly $u_h^{n*} \in \mathbb{K}$, thus

$$\begin{aligned} & - \langle \mu^n, u_h^n - u^n \rangle \\ = & - \langle \mu^n, u_h^n - u_h^{n*} \rangle - \langle \mu^n, u_h^{n*} - u^n \rangle \\ \leq & - \langle \mu^n, u_h^n - u_h^{n*} \rangle \\ = & \langle \mu^n, (\chi - u_h^n)_+ \rangle \\ = & \langle \mu_h^n, (\chi - u_h^n)_+ \rangle + \langle \mu^n - \mu_h^n, (\chi - u_h^n)_+ \rangle \\ = & \langle \mu_h^n, (\chi - u_h^n)_+ \rangle + \langle \mathcal{G}_h^n, (\chi - u_h^n)_+ \rangle - a(u_h^n - u^n, (\chi - u_h^n)_+) \\ & - \frac{1}{\Delta t_{n-1}} \langle u_h^n - u^n - u_h^{n-1} + u^{n-1}, (\chi - u_h^n)_+ \rangle. \end{aligned}$$

We use the following estimates

$$\begin{aligned} | \langle \mathcal{G}_h^n, (\chi - u_h^n)_+ \rangle | & \leq \|\mathcal{G}_h^n\|^2_{H^{-1}(\Omega)} + \tfrac{1}{4} |(\chi - u_h^n)_+|^2_{H_0^1(\Omega)}, \\ |a(u_h^n - u^n, (\chi - u_h^n)_+)| & \leq \epsilon |u_h^n - u^n|^2_{H_0^1(\Omega)} + \tfrac{1}{4\epsilon} |(\chi - u_h^n)_+|^2_{H_0^1(\Omega)}, \end{aligned}$$

for all positive number ϵ.
Similarly, for $\epsilon, \eta > 0$

$$
\begin{aligned}
&|\langle u_h^n - u^n - u_h^{n-1} + u^{n-1}, (\chi - u_h^n)_+\rangle| \\
\leq &\frac{\epsilon}{\Delta t_{n-1}} \|1_{\{\chi > \chi_h^n\}}(u_h^n - u^n - u_h^{n-1} + u^{n-1})\|^2_{L^{1+\eta}(\Omega)} + \frac{\Delta t_{n-1}}{4\epsilon} \|(\chi - u_h^n)_+\|^2_{L^{\frac{1+\eta}{\eta}}(\Omega)} \\
\leq &\frac{\epsilon}{\Delta t_{n-1}} \|1_{\{\chi > \chi_h^n\}}\|^2_{L^{\frac{2(1+\eta)}{1-\eta}}(\Omega)} \|(u_h^n - u^n - u_h^{n-1} + u^{n-1})\|^2_{L^2(\Omega)} + \frac{C\Delta t_{n-1}}{4\epsilon} |(\chi - u_h^n)_+|^2_{H^1(\Omega)},
\end{aligned}
$$

where we have used a Hölder inequality combined with Sobolev injections in dimension 2. Note that

$$
\|1_{\{\chi > \chi_h^n\}}\|^2_{L^{\frac{2(1+\eta)}{1-\eta}}(\Omega)} = \mathrm{meas}(\{\chi > \chi_h^n\})^{\frac{1-\eta}{1+\eta}}.
$$

We deduce from Lemma 10.3 that

$$
\begin{aligned}
&\left(\begin{array}{l}
\frac{3}{\Delta t_{n-1}} \|u_h^n - u^n\|^2_{L^2(\Omega)} + (1 - 6\epsilon)|u_h^n - u^n|^2_{H^1(\Omega)} \\
+ \left(\frac{3}{\Delta t_{n-1}} - \frac{6\epsilon}{\Delta t_{n-1}^2} \|1_{\{\chi > \chi_h^n\}}\|^2_{L^{\frac{2(1+\eta)}{1-\eta}}(\Omega)} \right) \|u_h^n - u_h^{n-1} - u^n + u^{n-1}\|^2_{L^2(\Omega)} \\
+ \| \frac{1}{\Delta t_{n-1}} \left(u_h^n - u_h^{n-1} - u^n + u^{n-1} \right) - \mu_h^n + \mu^n \|^2_{H^{-1}(\Omega)}
\end{array} \right) \\
\leq &11 \|\mathcal{G}_h^n\|^2_{H^{-1}(\Omega)} + \frac{3}{\Delta t_{n-1}} \|u_h^{n-1} - u^{n-1}\|^2_{L^2(\Omega)} + 6 \langle \mu_h^n, u_h^n - u^n + (\chi - u_h^n)_+ \rangle \\
&+ \left(\frac{3(1+\epsilon)}{2\epsilon} + \frac{3C}{2\epsilon} \right) |(\chi - u_h^n)_+|^2_{H_0^1(\Omega)}.
\end{aligned}
\tag{10.63}
$$

We choose any number ϵ such that $0 < \epsilon < 1/6$ so that $1 - 6\epsilon > 0$. Let us focus on $\langle \mu_h^n, u_h^n - u^n + (\chi - u_h^n)_+ \rangle$. We have

$$
\begin{aligned}
&\langle \mu_h^n, u_h^n - u^n + (\chi - u_h^n)_+ \rangle \leq \langle \mu_h^n, u_h^n - \chi \rangle + \langle \mu_h^n, (\chi - u_h^n)_+ \rangle \\
&= \langle \mu_h^n, (u_h^n - \chi)_+ \rangle \\
&= \sum_{z \in \mathcal{C}_{n,h}} \langle \mu_h^n, \phi_z (u_h^n - \chi)_+ \rangle + \sum_{\substack{z \in \mathcal{N}_{n,h}^0 \backslash \mathcal{C}_{n,h}, \\ u_h^n(z) = \chi_h^n(z)}} \langle \mu_h^n, \phi_z (u_h^n - \chi)_+ \rangle \\
&\leq \sum_{z \in \mathcal{C}_{n,h}} \langle \mu_h^n, \phi_z (\chi_h^n - \chi)_+ \rangle + \sum_{\substack{z \in \mathcal{N}_{n,h}^0 \backslash \mathcal{C}_{n,h}, \\ u_h^n(z) = \chi_h^n(z)}} \langle \mu_h^n, \phi_z (u_h^n - \chi_h^n)_+ \rangle \\
&\quad + \sum_{\substack{z \in \mathcal{N}_{n,h}^0 \backslash \mathcal{C}_{n,h}, \\ u_h^n(z) = \chi_h^n(z)}} \langle \mu_h^n, \phi_z (\chi_h^n - \chi)_+ \rangle \\
&= \sum_{\substack{z \in \mathcal{N}_{n,h}, \\ u_h^n(z) = \chi_h^n(z)}} \langle \mu_h^n, \phi_z (\chi_h^n - \chi)_+ \rangle + \sum_{\substack{z \in \mathcal{N}_{n,h}^0 \backslash \mathcal{C}_{n,h}, \\ u_h^n(z) = \chi_h^n(z)}} \langle \mu_h^n, \phi_z (u_h^n - \chi_h^n)_+ \rangle .
\end{aligned}
$$

The second term can be estimated exactly as in the proof of Lemma 10.6, by

$$
\sum_{\substack{z \in \mathcal{N}_{n,h}^0 \backslash \mathcal{C}_{n,h} \\ u_h^n(z) = \chi_h^n(z)}} \left[h_z (\|\widetilde{J_h^n}\|^2_{L^2(\gamma_z \cap \Gamma_{n,h}^+)} + \|J_h^n\|^2_{L^2(\gamma_z \cap \Gamma_{n,h}^+)}) + h_z^2 \left\| \frac{u_h^n - u_h^{n-1}}{\Delta t_{n-1}} - f \right\|^2_{L^2(\omega_z \cap \Omega_{n,h}^+)} \right],
$$

Using Lemma 10.5 and collecting all the estimates above, we obtain that

$$\begin{pmatrix} \|u_h^n - u^n\|^2_{L^2(\Omega)} + (1-6\epsilon)\sum_{p=1}^n \Delta t_{p-1}|u_h^p - u^p|^2_{H^1(\Omega)} \\ + \sum_{p=1}^n \left(3 - \frac{6\epsilon}{\Delta t_{p-1}}\mathrm{meas}(\{\chi > \chi_h^p\})^{\frac{1-\eta}{1+\eta}}\right)\|u_h^p - u_h^{p-1} - u^p + u^{p-1}\|^2_{L^2(\Omega)} \\ + \sum_{p=1}^n \Delta t_{p-1}\left\|\frac{1}{\Delta t_{p-1}}\left(u_h^p - u_h^{p-1} - u^p + u^{p-1}\right) - \mu_h^p + \mu^p\right\|^2_{H^{-1}(\Omega)} \end{pmatrix} \tag{10.64}$$

$$\lesssim \sum_{p=1}^n \Delta t_{p-1}\left(\sum_{z\in\mathcal{N}_{p,h}} \eta^2_{p,z} + |(\chi - u_h^p)_+|^2_{H_0^1(\Omega)} + \left\langle \mu_h^p, (\chi_h^p - \chi)_+\right\rangle\right),$$

where

$$\eta^2_{p,z} = h_z^2\left\|\frac{u_h^p - u_h^{p-1}}{\Delta t_{p-1}} - f\right\|^2_{L^2(\omega_z\cap\Omega^+_{p,h})} + h_z\|J_h^p\|^2_{L^2(\gamma_z\cap\Gamma^+_{p,h})} \tag{10.65}$$

$$+ h_z 1_{\left\{\begin{array}{l} z\in\mathcal{N}^0_{p,h}\backslash\mathcal{C}_{p,h} \\ u_h^p(z) = \chi_h^p(z)\end{array}\right\}}\|\widetilde{J_h^p}\|^2_{L^2(\gamma_z\cap\Gamma^+_{p,h})}.$$

Conclusion on the general case Without the assumption $\chi_h^n \geq \chi$, $1 \leq n \leq N$, we have obtained the a posteriori error estimate (10.64) which is useful only if there is a constant c such that

$$\mathrm{meas}(\{\chi > \chi_h^p\})^{\frac{1-\eta}{1+\eta}} \leq c\Delta t_{p-1}, \quad \forall 1 \leq p \leq n. \tag{10.66}$$

Indeed, in this case, one can choose ϵ, $0 < \epsilon < 1/6$ such that all the terms in (10.64) are positive. Condition (10.66) relates the time step and the measure of the set where $\chi > \chi_h^p$. Therefore, (10.64) is not fully satisfactory. In our opinion, this shows that it is better to choose $\chi_h^n \geq \chi$ when possible.

10.5 Numerical Results

10.5.1 A Piecewise Affine Obstacle

We take $T = 0.1$, $\Omega = (-1,1)^2$, $\chi(x) = \max(0.5 - |x_1| - |x_2|, 0)$, see Figure 10.1, $u_0 = \chi$ and $f(x) = -4$. The function χ is clearly piecewise affine. The method used to solve problem (10.10) is a semi-smooth Newton method studied by Hintermüller, Ito and Kunish [HIK02] and Ito and Kunish[IK03].

The tests have been done using the free software FreeFem++ [PH] and a C++ code for mesh intersections.

In this experiment, several successive time/space meshes adaptions based on the error indicators will be carried out. We need another index to distinguish the successive mesh refinements : the spatial meshes are now named $\mathcal{T}^{n,p}$ where n corresponds to the time step and p to the adaption level (we have dropped the index h, because it is not useful here). In the following tests, two successive mesh adaptions are made, so p varies from 0 to 2. During the first time stepping, all the meshes $\mathcal{T}^{n,0}$ are the same, namely $\mathcal{T}^0$. For $p \geq 1$, $\mathcal{T}^{n,p}$ is constructed by refining a mesh $\mathcal{T}^{m,p-1}$. Therefore, all the meshes $\mathcal{T}^{n,p}$ will be refinements of $\mathcal{T}^0$. Moreover $\mathcal{T}^0$ is a 20×20 uniform mesh constructed in such a way that χ belongs to the corresponding finite element space. From the previous two remarks,

it is clear that $\chi \in V_h^{n,p}$ for all n and p. The mesh $\mathcal{T}^0$ is represented on the top/left part of Figure 10.2.

Remark 10.3.5 *Mesh coarsening has not been implemented yet in our computer code which has been written only to test the indicators. It should certainly be done, because mesh coarsening is an important ingredient of adaptivity for evolutionary problems. Note that it is quite possible to design a strategy based on the error indicators proposed above and including mesh coarsening.*

The grid in the time variable is refined as well. The strategy is to divide the time steps corresponding to the larger values of the error indicator $\Delta t_{n-1}\eta_n^2$. Three successively adapted grids in the time variable are represented on the top/right part of Figure 10.2. We see that the grid is refined near $t = 0$. This could be expected, because the solution is singular at $t = 0$. The graph of the error indicator $\Delta t_{n-1}\eta_n^2$ as a function of time for the three levels of adaption is plotted in the left part of Figure 10.5. We see that the adaption strategy has the effect of decreasing the variations of the error indicator.
The spatial meshes corresponding to the first level of refinement and the dates $t = 0$ and $t = 0.1$ are respectively plotted in the bottom/left and bottom/right parts of Figure 10.2. The meshes corresponding to the second level of refinement and the dates $t = 0$ and $t = 0.1$ are respectively plotted in the left and right parts of Figure 10.3. We see that the meshes differ from $t = 0$ to $t = 0.1$. The mesh is not refined near $x = 0$ although the function is singular there, because this region is contained in the full contact zone (and because $\chi \in V_h^{n,p}$for all n, p). The mesh points are concentrated in the non contact zone where the obstacle is singular.
The strategy to refine the mesh $\mathcal{T}_h^{n,p}$ is to

- interpolate the error indicators $\eta_{n,z}^2$,
- refine the triangles where this function is large, so has to lessen the variation of the indicator.

The graph of the Hilbertian sum of the error indicators $\Delta t_{n-1}\eta_{n_z}^2$ as a function of time for the three levels of adaption is plotted in the right part of Figure 10.5. We see that the adaption strategy is very efficient for decreasing $\Delta t_{n-1}\sum_{z\in\mathcal{N}_{n,h}}\eta_{n,z}^2$.
The contours of the function $u_{\Delta t,h} - \chi$ at $t = 0.1$ are plotted in Figure 10.4, for the three successive adaption levels. We see that this function exhibits singularities and that there are two contact regions, near 0 and near the boundary $\partial\Omega$. The boundary of the contact region is plotted in Figure 10.6 for the last two adaption levels. This free boundary is singular near the axes.

10.5.2 A Non Piecewise Affine Obstacle

The only difference with the previous experiment is that we take $\chi(x) = \max(-0.1 + 0.6\exp(()-10*(x_1^2+x_2^2)), 0.5-\sqrt{x_1^2+x_2^2}, 0)$, see Figure 10.7. Three successively adapted grids in the time variable are represented on the top/right part of Figure 10.8. The graph of the error indicator $\Delta t_{n-1}\eta_n^2$ as a function of time for the three levels of adaption is plotted in the left part of Figure 10.11.
The spatial meshes corresponding to the first level of refinement and the dates $t = 0$ and $t = 0.1$ are respectively plotted in the bottom/left and bottom/right parts of Figure 10.8. The meshes corresponding to the second level of refinement and the dates $t = 0$ and $t = 0.1$ are respectively plotted in the left and right parts of Figure 10.9. We see that the

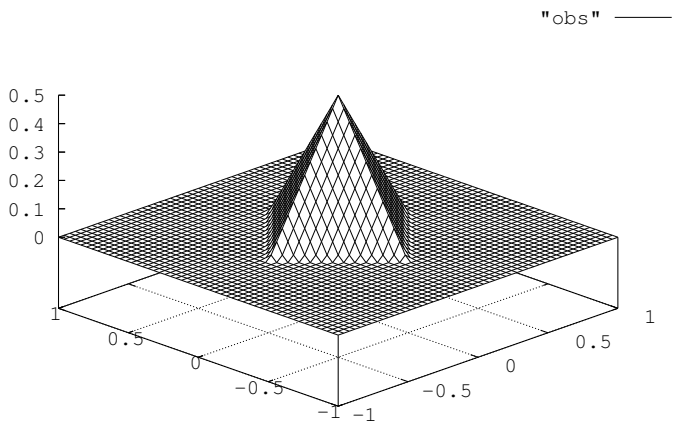

FIG. 10.1 – The function χ

meshes differ from $t = 0$ to $t = 0.1$. The mesh is refined near $x = 0$ although this region is contained in the full contact zone because $\chi \notin V_h^{n,p}$.
The graph of the Hilbertian sum of the error indicators $\Delta t_{n-1}\eta_{n_z}^2$ as a function of time for the three levels of adaption is plotted in the right part of Figure 10.11.
The contours of the function $u_{\Delta t,h} - \chi$ at $t = 0.1$ are plotted in Figure 10.10, for the three successive adaption levels. The boundary of the contact region is plotted in Figure 10.12 for the last two adaption levels. There are three connected contact regions, a ball centered at the origin, a concentric ring, and the complement of a ball in Ω.

10.6 An Application in Finance

10.6.1 The discrete method and the error indicators

We consider an American put option on a basket containing two assets, with maturity T. Let $x_{i,\tau}$ $i = 1, 2$, be the prices of the assets at time τ. We assume that these prices satisfy the stochastic differential equations

$$dx_{i,\tau} = x_{i,\tau}\left(\mu_i d\tau + \sigma_i dW_{i,\tau}\right), \tag{10.67}$$

where $\sigma_i > 0$ is the volatility of the asset indexed by i, and $(W_{i,\tau})_\tau$ are possibly correlated Brownian motions. We call ρ the correlation factor.
The payoff at maturity is $\chi(x_{1,T}, x_{2,T})$. In what follows, we take

$$\chi(x) = \max(0, \min(K - x_1, K - x_2)),$$

so χ is piecewise linear.
We call $t = T - \tau$ the time to maturity.
In the Black-Scholes model, the price of the option is given by $P_t = u(t, x_{1,T-t}, x_{2,T-t})$,

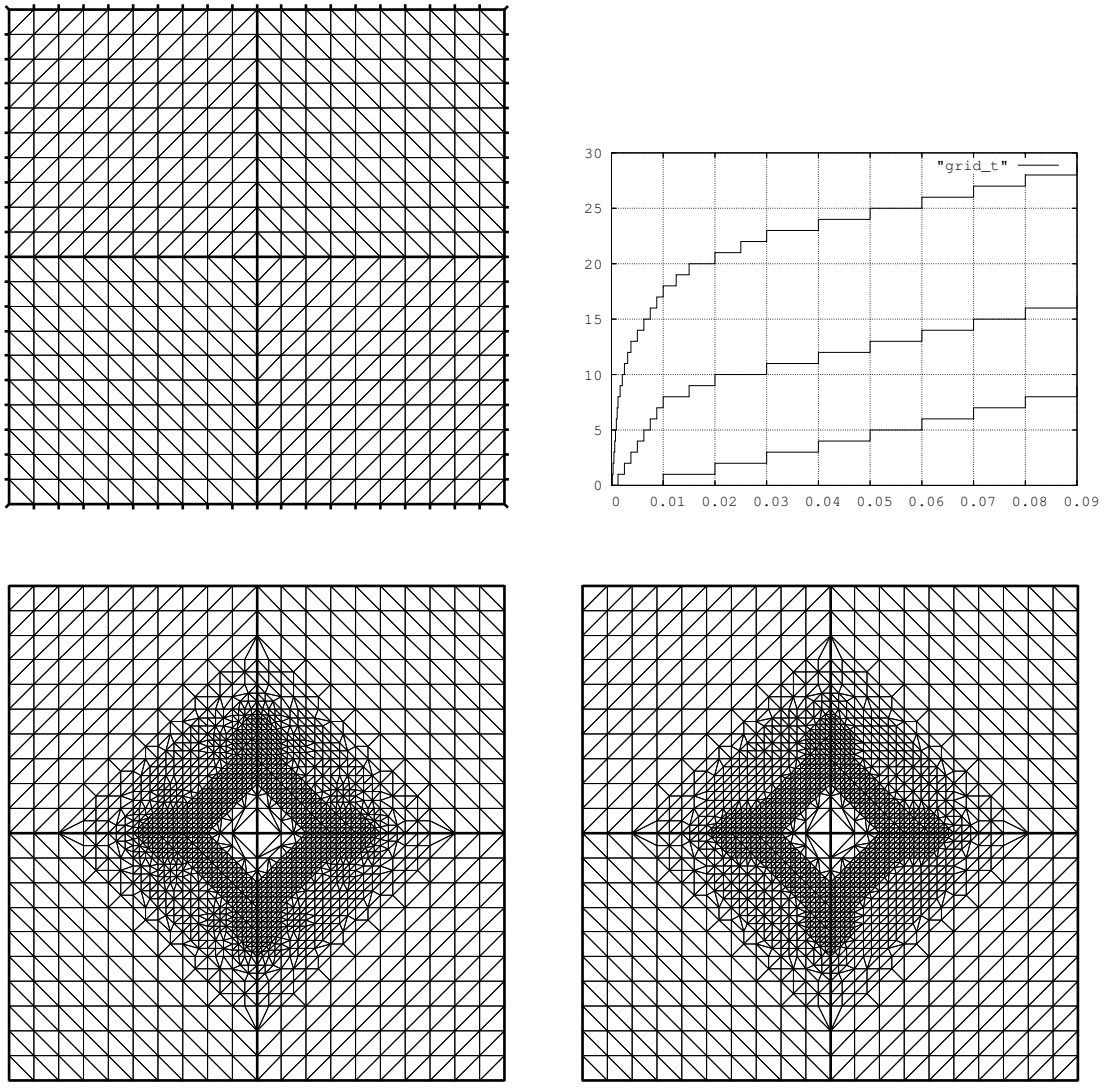

FIG. 10.2 – Top : the initial finite element mesh $\mathcal{T}^0$ (left) and three successively adapted time grids (right). Bottom : first adaption, the adapted meshes used near $t = 0$(left) and at $t = 0.1$(right)

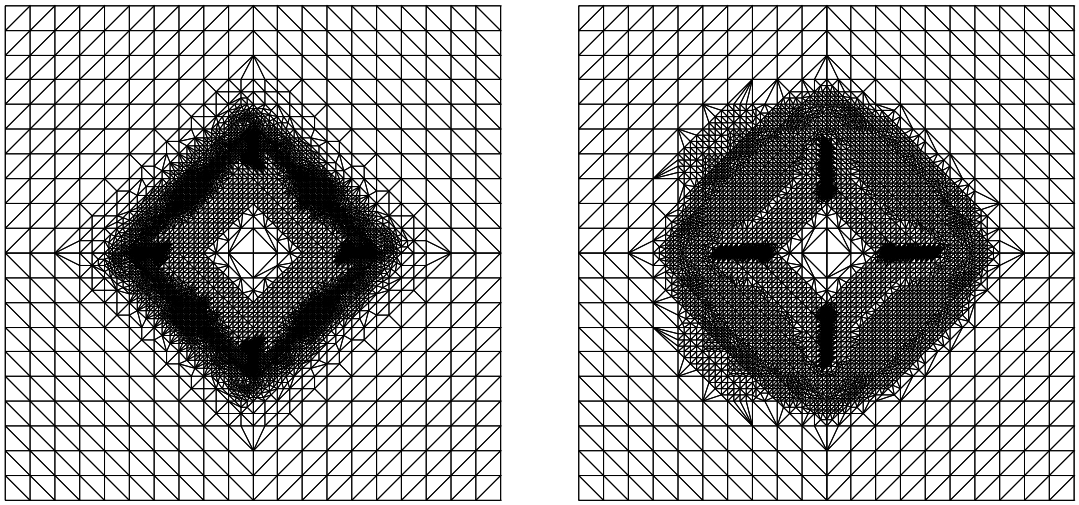

FIG. 10.3 – Second adaption, the adapted mesh used near $t = 0$(left) and at $t = 0.1$(right)

where

$$\begin{array}{rcll} \frac{\partial u}{\partial t} - Lu + ru & \geq & 0, & \text{in } \mathbb{R}^2_+ \times (0,T], \\ u & \geq & \chi, & \text{in } \mathbb{R}^2_+ \times (0,T], \\ \left(\frac{\partial u}{\partial t} - Lu + ru\right)(u-\chi) & = & 0, & \text{in } \mathbb{R}^2_+ \times (0,T], \\ u(t=0) & = & \chi, & \text{in } \mathbb{R}^2_+, \end{array} \tag{10.68}$$

and

$$Lu = \frac{1}{2}\sum_{i=1}^{2}\sum_{j=1}^{2}\Xi_{i,j}x_i x_j \frac{\partial^2 u}{\partial x_i \partial x_j} + \sum_{j=1}^{2} r x_j \frac{\partial u}{\partial x_j}. \tag{10.69}$$

Here $r > 0$ is the interest rate of the risk-free asset and $\Xi_{1,1} = \sigma_1^2$, $\Xi_{2,2} = \sigma_2^2$, $\Xi_{1,2} = \Xi_{2,1} = \rho\sigma_1\sigma_2$.
The pricing function is computed in the rectangular domain $\Omega = (0, \overline{x})^2$ with $\overline{x}$ large enough. We impose homogeneous Dirichlet boundary conditions at the artificial boundary

$$\Sigma_0 = \{x \in \partial\Omega;\ \max(x_1, x_2) = \overline{x}\}. \tag{10.70}$$

We introduce the Hilbert space

$$V = \left\{ v :\ v \in L^2(\Omega), x_i \frac{\partial v}{\partial x_i} \in L^2(\Omega), i = 1, \ldots, 2 \right\}, \tag{10.71}$$

with the norm $\|v\|_V = \left(\|v\|^2_{L^2(\Omega)} + \sum_{i=1}^{2} \|x_i \frac{\partial v}{\partial x_i}\|^2_{L^2(\Omega)} \right)^{\frac{1}{2}}$, and V^0 :

$$V^0 = \{v \in V;\ v|_{\Sigma_0} = 0\} \tag{10.72}$$

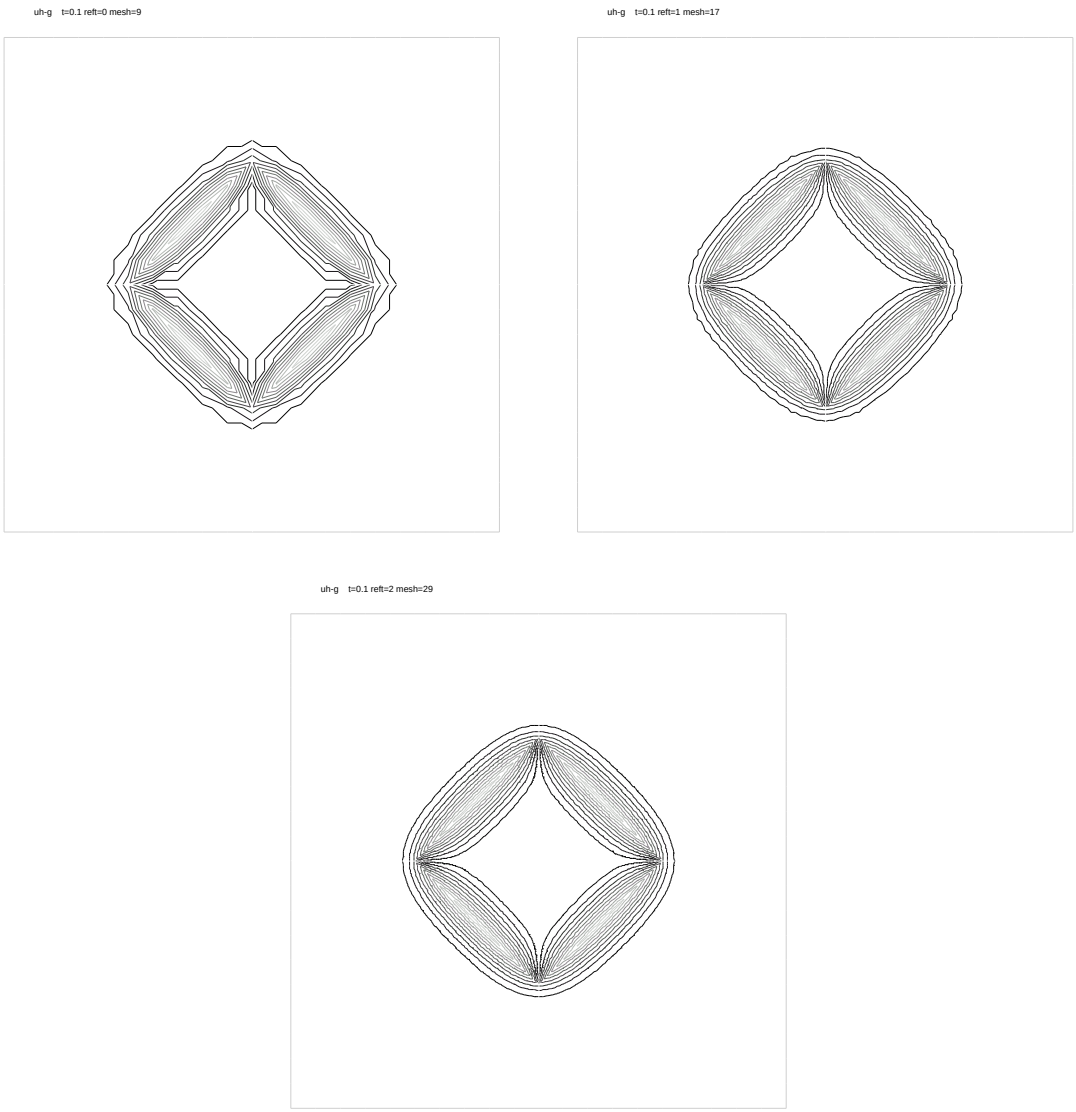

FIG. 10.4 – The contours of the function $u_{\Delta t,h} - \chi$ at time 0.1 computed with successively adapted time-space meshes.

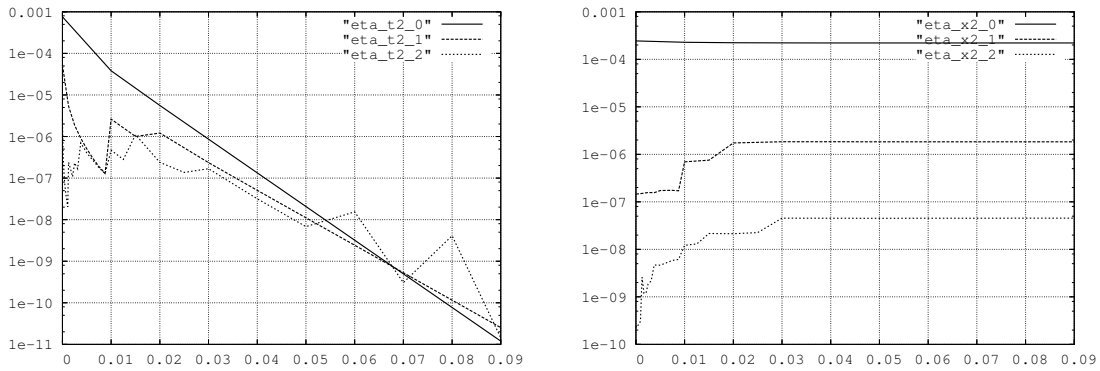

FIG. 10.5 – The indicators $\Delta t_{n-1}\eta_n^2$ (left) and $\Delta t_{n-1}\sum_{z\in\mathcal{N}_{n,h}}\eta_{n,z}^2$ (right) in log scale for three successive time/steps refinements, as a function of t_n

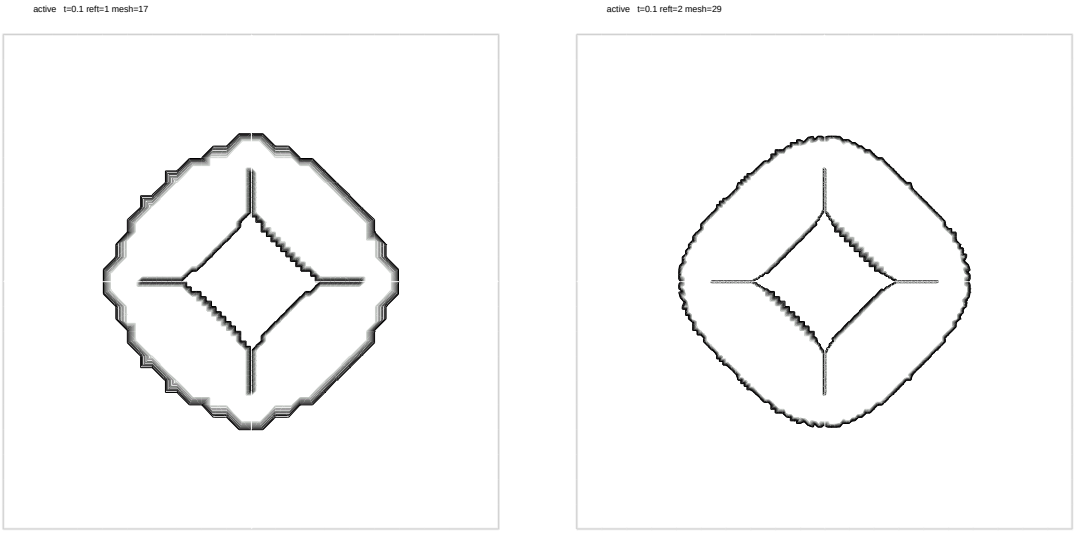

FIG. 10.6 – The boundary of the contact region at time $t = 0.1$ computed with successively adapted time-space meshes.

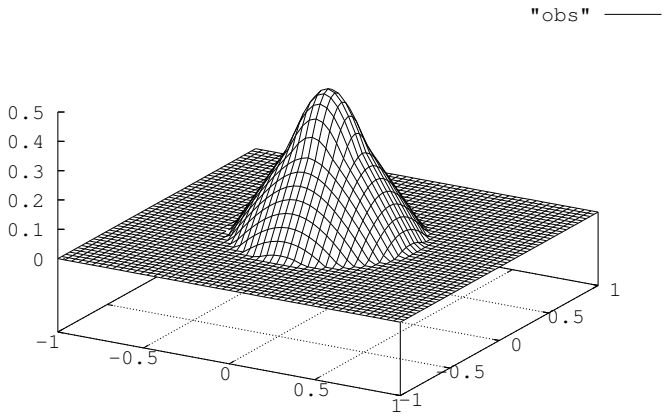

FIG. 10.7 – The function χ

Note that $|v|_V = \left(\sum_{i=1}^2 \|x_i \frac{\partial v}{\partial x_i}\|^2_{L^2(\Omega)}\right)^{\frac{1}{2}}$ defines a norm on V equivalent to the one above, see [AP05]. We introduce the bilinear form a on $V \times V$:

$$\begin{aligned} a(u,v) &= \frac{1}{2}\sum_{i=1}^{2}\sum_{j=1}^{2}\int_\Omega \Xi_{i,j}x_i x_j \frac{\partial u}{\partial x_j}\frac{\partial v}{\partial x_i} \\ &\quad -\sum_{j=1}^{2}\int_\Omega \left(rx_j - \frac{1}{2}\sum_{i=1}^{2}\frac{\partial}{\partial x_i}\left(\Xi_{i,j}x_i x_j\right)\right)\frac{\partial u}{\partial x_j}v + r\int_\Omega uv. \end{aligned} \tag{10.73}$$

With the new definitions of $\mathbb{K}$

$$\mathbb{K} = \{v \in V^0,\ v \geq \chi \text{ a.e. in } \Omega\}, \tag{10.74}$$

and of $\mathcal{K}$

$$\mathcal{K} = \{v \in L^2(0,T;V^0) \text{ s.t. } v(t) \in \mathbb{K} \text{ for a.a. } t \in (0,T)\}, \tag{10.75}$$

the variational formulation of (10.68) is to find $u \in \mathcal{K} \cap C^0([0,T];L^2(\Omega))$, such that $\frac{\partial u}{\partial t} \in L^2(0,T;(V^0)')$ and $u(t=0) = \chi$ and satisfying for a.a. $t \in (0,T)$,

$$\left\langle \frac{\partial u}{\partial t}(t), v - u(t) \right\rangle + a(u(t), v - u(t)) \geq 0, \qquad \forall v \in \mathbb{K}. \tag{10.76}$$

The sequence of meshes $(\mathcal{T}_{n,h})_h$ of Ω is such that Assumption 10.3 holds. The discrete spaces are

$$V_{n,h} = \left\{v_h \in V,\ \forall \omega \in \mathcal{T}_{n,h}, v_{h|\omega} \in \mathcal{P}_1\right\}, \qquad V^0_{n,h} = V_{n,h} \cap V^0, \tag{10.77}$$

and the discrete scheme is (10.10) with $f = 0$ and $\mathbb{K}_{n,h} = \{v_h \in V^0_{n,h},\ v_h \geq \chi\}$.
In order to define the full contact zone, the Lagrange multiplier and the error indicators,

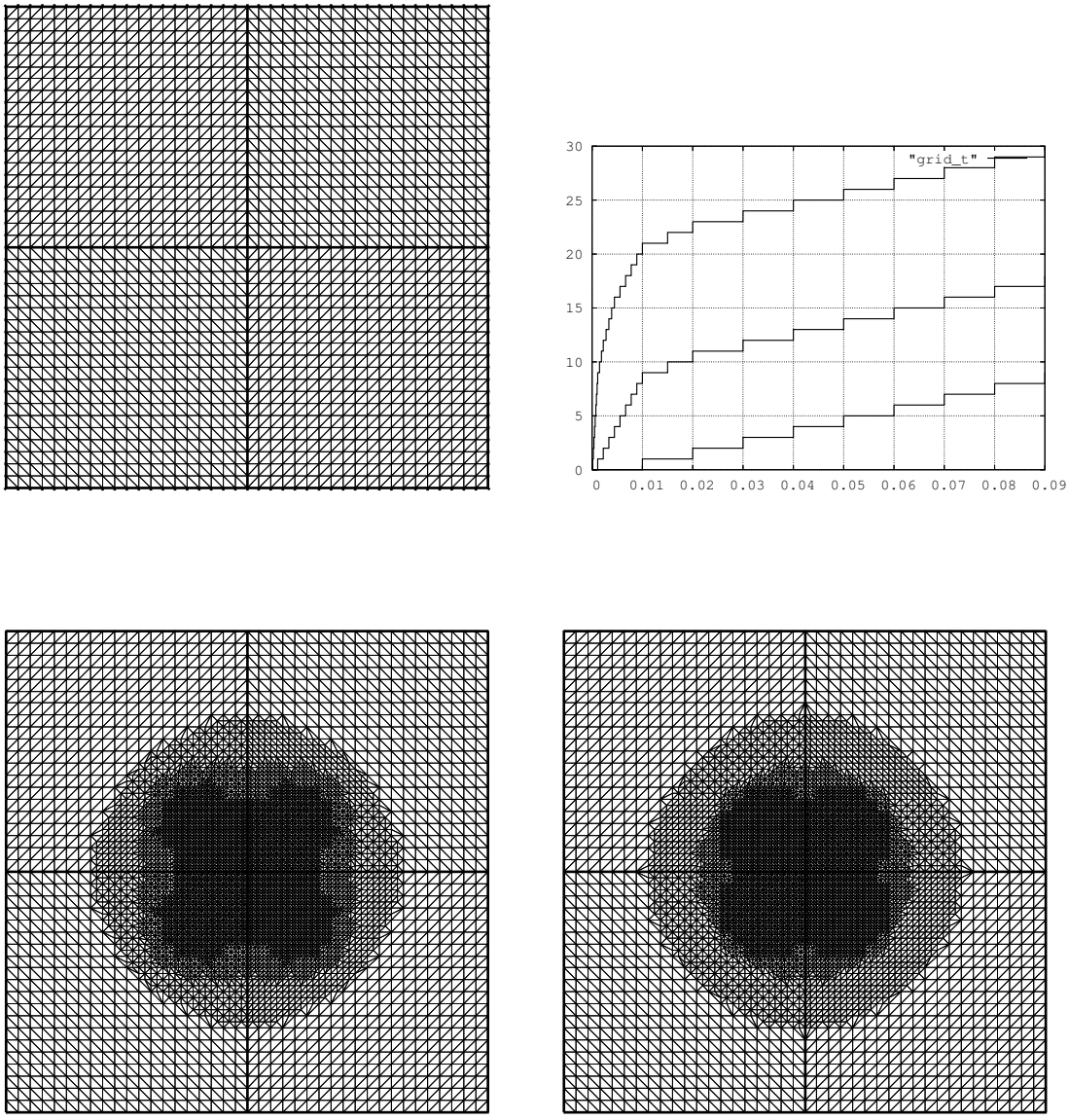

FIG. 10.8 – Top : the initial finite element mesh $\mathcal{T}^0$ (left) and three successively adapted time grids (right). Bottom : first adaption, the adapted meshes used near $t = 0$(left) and at $t = 0.1$(right)

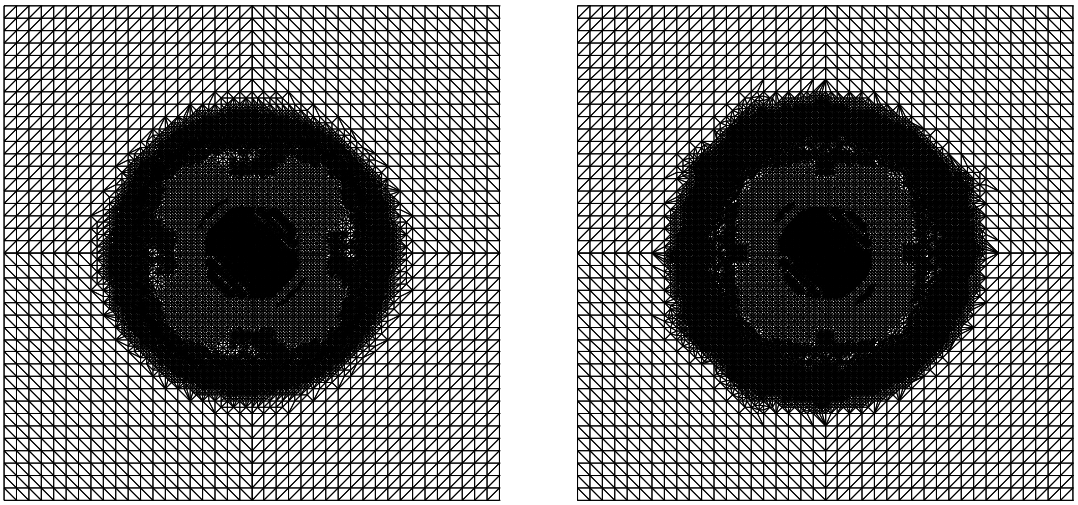

FIG. 10.9 – Second adaption, the adapted mesh used near $t = 0$(left) and at $t = 0.1$(right)

we introduce the jumps J_h^n : if $S \subset \Gamma_{n,h}$ is the common side of the two triangles κ^- and κ^+ in $\mathcal{T}_{n,h}$, then

$$J_h^n|_S = \frac{1}{2}\sum_{i=1}^{2} \boldsymbol{n}_i \left[\sum_{j=1}^{2} \Xi_{i,j} x_i x_j \frac{\partial u_h^n}{\partial x_j},\right], \tag{10.78}$$

where $\boldsymbol{n}$ is the unit normal vector to S pointing from κ_- to κ_+. We do not need to consider the edges S on $\partial\Omega\backslash\Sigma_0$ because $\sum_{i=1}^{2} \boldsymbol{n}_i \sum_{j=1}^{2} \Xi_{i,j} x_i x_j \frac{\partial u_h^n}{\partial x_j} = 0$ on S.
For $n = 1, \ldots, N$, we introduce the set $\mathcal{C}_{n,h}$ of the full contact nodes at t_n by :

$$\mathcal{C}_{n,h} = \left\{ z \in \mathcal{N}_{n,h} \text{ s.t. } \left| \begin{array}{l} u_h^n = \chi \text{ and } u_h^n - u_h^{n-1} \geq r\Delta t_{n-1} \left(\sum_{i=1}^{2} x_i \frac{\partial u_h^n}{\partial x_i} - u_h^n\right) \text{ in } \omega_z, \\ J_h^n \leq 0 \text{ on } \gamma_z \end{array} \right. \right\}. \tag{10.79}$$

The sets $\Omega_{n,h}^0$, $\Omega_{n,h}^+$, $\Gamma_{n,h}^0$, $\Gamma_{n,h}^+$ are defined as in § 10.2.2.2. The indicators η_n and $\eta_{n,z}$ are given by

$$\eta_n = \sigma |u_h^n - u_h^{n-1}|_V, \tag{10.80}$$

where $\sigma = \max(\sigma_1, \sigma_2)$ and

$$\eta_{n,z}^2 = \frac{1}{\alpha_z^2} \left(\begin{array}{l} h_z^2 \left\| \dfrac{u_h^n - u_h^{n-1}}{\Delta t_{n-1}} - r\left(\sum_{i=1}^{2} x_i \dfrac{\partial u_h^n}{\partial x_i} - u_h^n\right) \right\|^2_{L^2(\omega_z \cap \Omega^+_{n,h})} \\ + h_z \|J_h^n\|^2_{L^2(\gamma_z \cap \Gamma^+_{n,h})} + h_z 1 \left\{ \begin{array}{l} z \in \mathcal{N}^0_{n,h} \backslash \mathcal{C}_{n,h} \\ u_h^n(z) = \chi(z) \end{array} \right\} \|\widetilde{J_h^n}\|^2_{L^2(\gamma_z \cap \Gamma^+_{n,h})} \end{array} \right), \tag{10.81}$$

where $\alpha_z = \min_{z \in \omega_z}(z_1, z_2) + h_z$ and $\widetilde{J_h^n}$ is obtained by replacing u_h^n by $u_h^n - \chi$ in (10.78). It is possible to prove that Theorem 10.8 holds, with $H^1(\Omega)$ replaced with V and $H^{-1}(\Omega)$ replaced with V_0'.

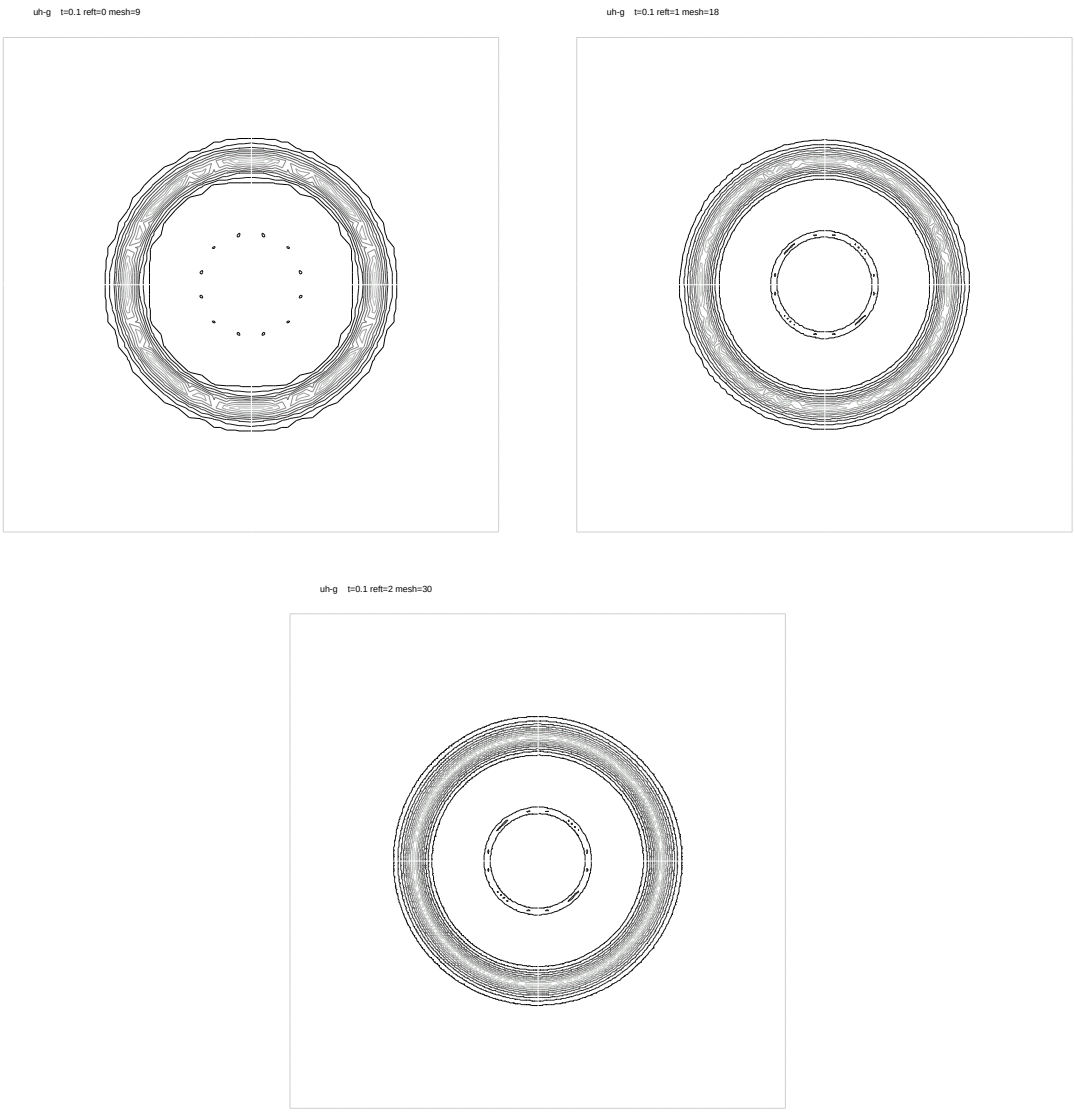

FIG. 10.10 – The contours of the function $u_{\Delta t,h} - \chi$ at time 0.1 computed with successively adapted time-space meshes.

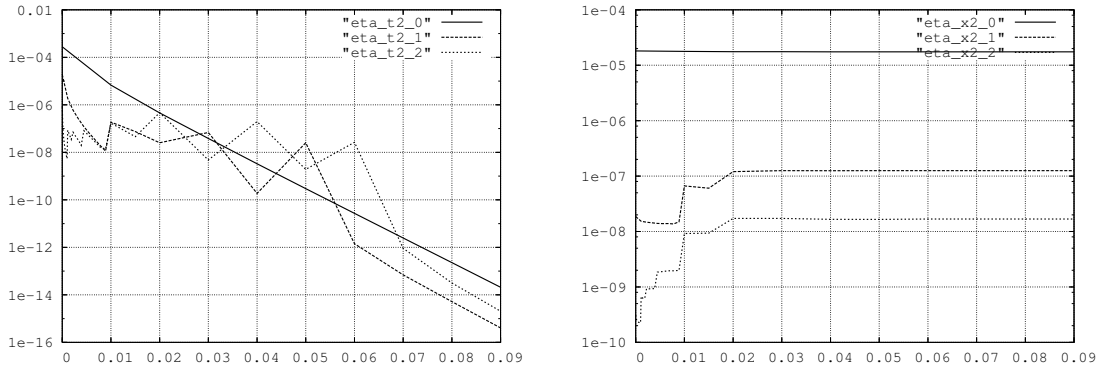

FIG. 10.11 – The indicators $\Delta t_{n-1}\eta_n^2$ (left) and $\Delta t_{n-1}\sum_{z\in\mathcal{N}_{n,h}}\eta_{n,z}^2$ (right) in log scale for three successive time/space refinements, as a function of t_n

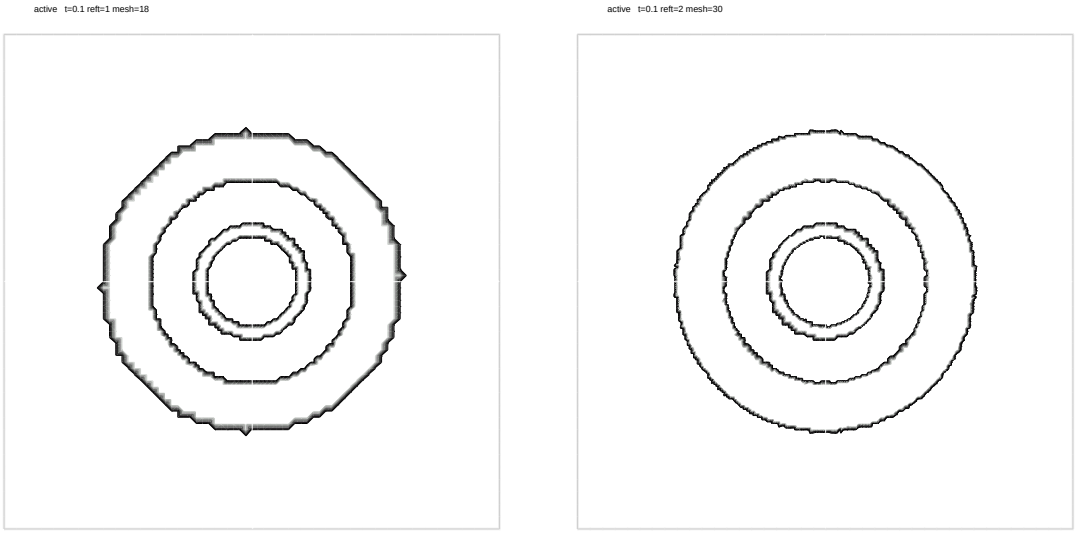

FIG. 10.12 – The boundary of the contact region at time $t = 0.1$ computed with successively adapted time-space meshes.

10.6.2 Numerical Results

We choose the following values for the relevant parameters :

$$\sigma_1 = 0.2, \quad \sigma_2 = 0.1, \quad \rho = -0.8, \quad r = 0.05, \quad T = 1,$$

and the payoff function is $\chi(x) = (K - \max(x_1, x_2))_+$ with $K = 25$. We choose $\overline{x} = 100$. Three successively adapted grids in the time variable are represented in the top/right part of Figure 10.13. The time-grids are very much like the one obtained in § 10.5 above. The initial spatial mesh is in the top/left side of Figure 10.13 : it is uniform with 40×40 nodes, so χ belongs to the finite element space. The spatial meshes corresponding to the first level of refinement and the dates $t = 0$ and $t = 1$ are respectively plotted in the bottom/left and bottom/right parts of Figure 10.13. The spatial meshes corresponding to the third level of refinement and the dates $t = 0$ and $t = 1$ are respectively plotted in the left and right parts of Figure 10.14. We see that the mesh is refined near the free boundary (exercise boundary in the language of financial options), but not near the diagonal $x_1 = x_2$ although the payoff function is singular there. In Figure 10.15 we show the contours of $u_{\Delta t,h}(t = T) - \chi$ obtained with the initial time-space mesh and after the first and third adaption steps. The graph of the error indicator $\Delta t_{n-1}\eta_n^2$ as a function of time for the three levels of adaption is plotted in the left part of Figure 10.16. The right part of this figure shows the graphs of the Hilbertian sum of the error indicators $\Delta t_{n-1}\eta_{n_z}^2$ as a function of time for the three levels of adaption. Finally, the exercise boundary at time $t = T$ obtained after the first and third refinment stages are shown in Figure 10.17. The shape of the exercise region (contact zone) is nontrivial and singular.

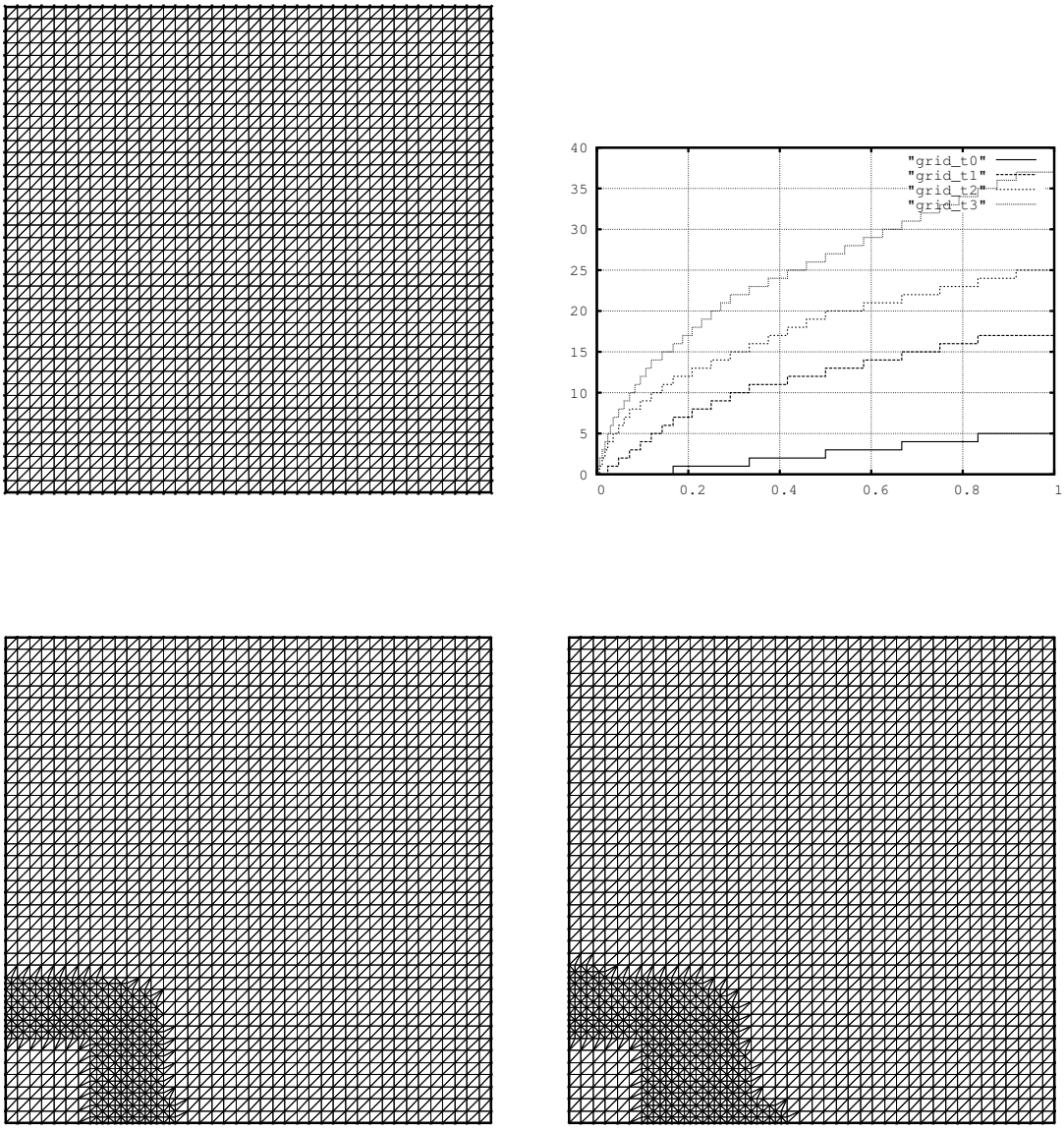

FIG. 10.13 – Top : the initial finite element mesh $\mathcal{T}^0$ (left) and three successively adapted time grids (right). Bottom : first adaption, the adapted meshes used near $t = 0$(left) and at $t = 1$(right)

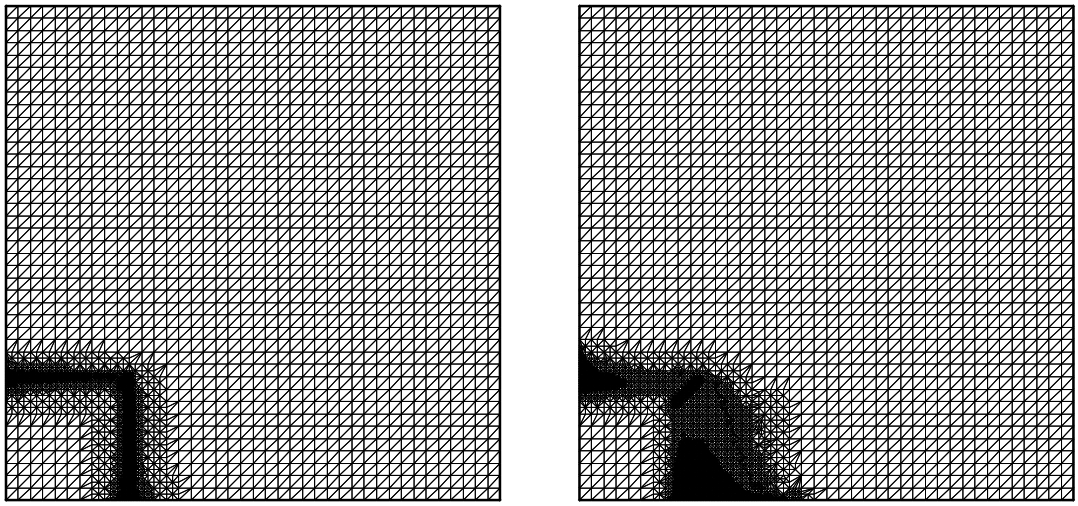

FIG. 10.14 – Third adaption, the adapted mesh used near $t = 0$(left) and at $t = 1$(right)

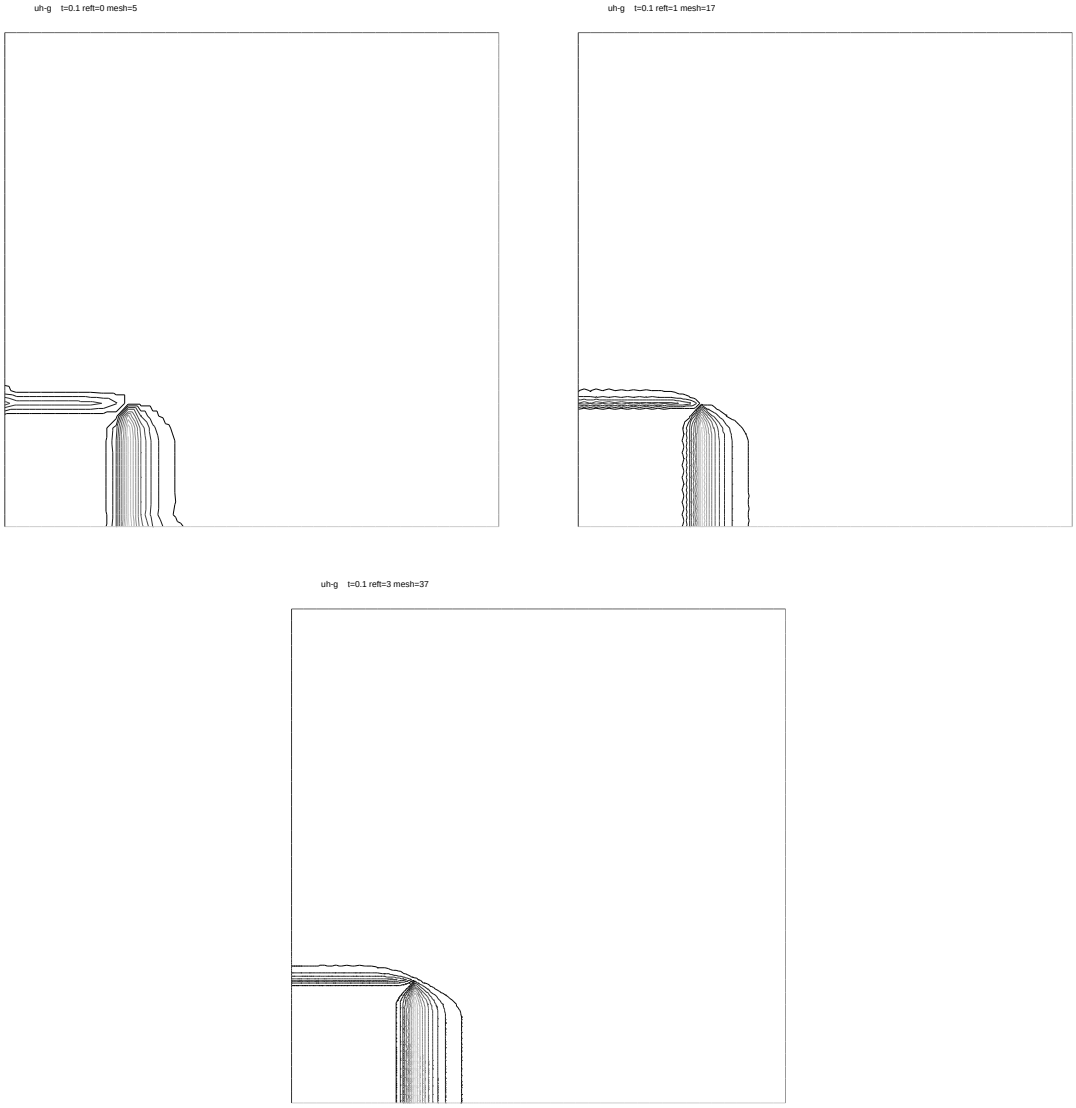

FIG. 10.15 – The contours of the function $u_{\Delta t,h} - \chi$ at time 1 computed with successively adapted time-space meshes.

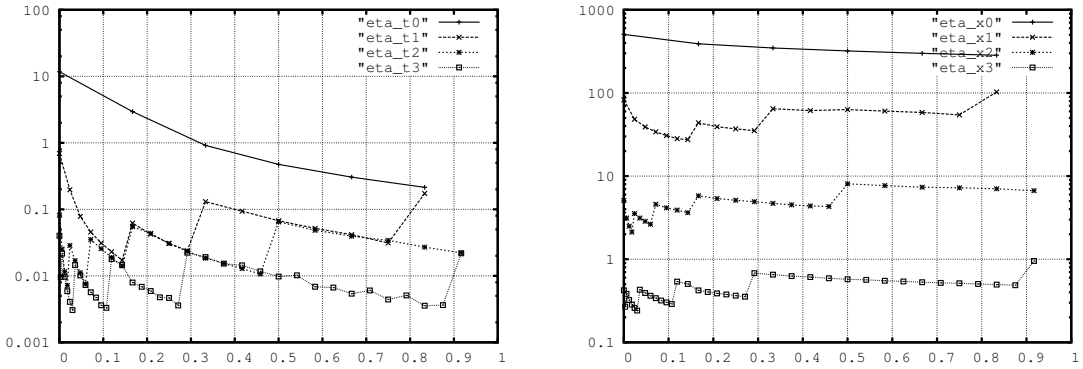

FIG. 10.16 – The indicators $\Delta t_{n-1}\eta_n^2$ (left) and $\Delta t_{n-1}\sum_{z\in\mathcal{N}_{n,h}}\eta_{n,z}^2$ (right) in log scale for three successive time/space refinements, as a function of t_n

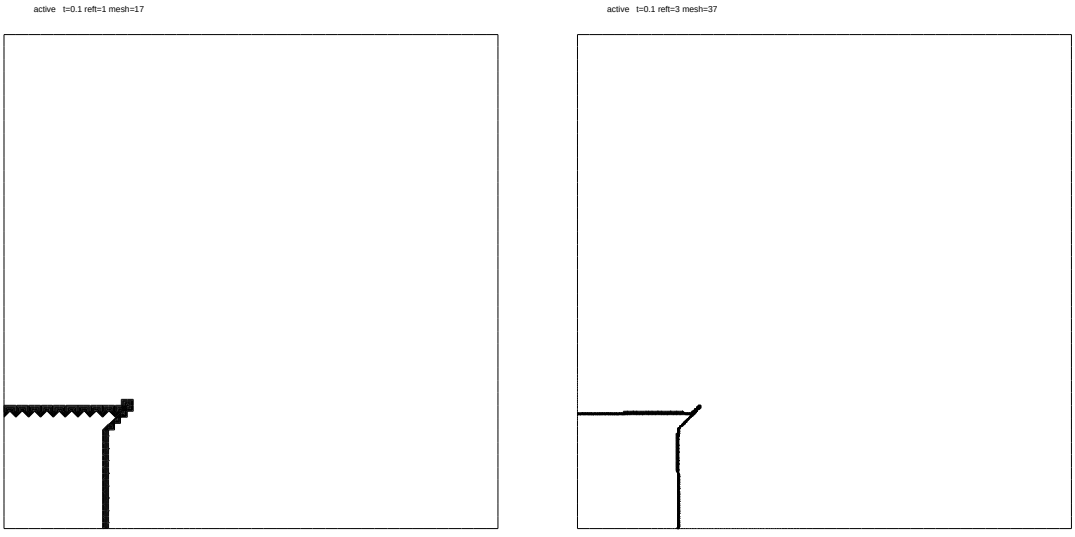

FIG. 10.17 – The boundary of the contact region at time $t = 1$ computed with successively adapted time-space meshes.

Annexes

Annexe A

Prime de risque sur le modèle à volatilité stochastique

Reprenons le modèle à volatilité stochastique présenté au chapitre 5 dans lequel le sous-jacent suit la dynamique donnée par l'équation différentielle stochastique suivante :

$$dS_t = \mu S_t dt + \sigma_t S_t dW_t, \tag{A.1}$$

où $\mu S_t dt$ représente le terme de « drift », (W_t) est un brownien et (σ_t) représente la volatilité. Nous supposons que (σ_t) est un processus stochastique à valeurs dans $\mathbb{R}^+$, qui satisfait lui aussi une équation différentielle stochastique. Nous introduisons le brownien $\widetilde{W}_t$, corrélé à W_t. Nous supposons que $\sigma_t = v(t)e^{X_t}$, où X_t est un processus gaussien de la forme

$$X_t = \int_0^t K(t,s)\, d\widetilde{W}_s$$

où le noyau K appartient à $L^2\left([0,T]^2\right)$ avec $K(t,t) \neq 0$. Le processus $(X_t)_{t\geq 0}$ vérifie l'EDS suivante :

$$dX_t = \left(\int_0^t \partial_1 K(t,s) d\widetilde{W}_s\right) dt + K(t,t) d\widetilde{W}_t = \gamma_t dt + K(t,t) d\widetilde{W}_t. \tag{A.2}$$

Le processus de volatilité vérifie l'équation :

$$d\sigma_t = \sigma_t \left(\phi(t)dt + \gamma_t dt + K(t,t) d\widetilde{W}_t\right), \tag{A.3}$$

où $\phi(t) = \dfrac{v'(t)}{v(t)}$, et $K(t,s) = \displaystyle\sum_{k=1}^{n} \beta_k e^{-\lambda_k (t-s)}$.

A.1 Équation de valorisation sur le processus gaussien

Considérons un contrat européen sur le sous-jacent qui suit l'EDS (A.1). Notons T la date d'expiration du contrat et h la fonction payoff. La question de la dimension du problème se pose naturellement. Nous avons cherché à réduire le nombre de variables intervenant dans l'équation de valorisation. La variable X_t n'est pas markovienne. Néanmoins,

il existe des résultats sur les processus gaussiens et, plus particulièrement, sur les « browniens factionnaires » [NT06, Nec04, BØSW04] qui permettent d'établir une équation aux dérivées partielles sur ces processus non-markoviens. Cependant, cette méthode est fortement liée au choix de la probabilité risque neutre. Notre modèle fixe cette probabilité risque neutre. Nous affirmons que l'équation de valorisation d'une option européenne dans le cas d'un modèle à m facteurs est de dimension $m+1$ en espace ($m+2$ si nous ajoutons la variable temps).

Pour cela, nous montrons qu'il n'est pas possible de construire un portefeuille de réplication qui reproduit le prix de l'option européenne dans notre modèle de diffusion, une fois la probabilité risque neutre fixée.

Supposons que le prix P du contrat dépend des trois processus (S_t, σ_t, t), en notant r_t le taux d'intérêt. La formule d'Itô multidimensionnelle permet d'exprimer dP comme combinaison linéaire de dt, dW_t et $\gamma_t dt$ et $d\widetilde{W}_t$,

$$\begin{aligned} dP(S_t,\sigma_t,t) =& \partial_t P dt + \partial_s P dS_t + \partial_\sigma P d\sigma_t \\ &+ \left(\frac{1}{2}\sigma_t^2 S_t^2 \partial_s^2 P + \sigma_t^2 S_t K(t,t)\rho \partial_{s,\sigma}^2 P + \frac{1}{2}\sigma_t^2 K(t,t)^2 \partial_\sigma^2 P\right) dt. \end{aligned} \tag{A.4}$$

En introduisant les équations des dynamiques (A.1) et (A.3),

$$\begin{aligned} dP(S_t,\sigma_t,t) = & \partial_t P dt + S_t \partial_s P dt + \partial_s P \sigma_t S_t dW_t + \partial_\sigma P \sigma_t \left(\phi(t)dt + \gamma_t dt + K(t,t)d\widetilde{W}_t\right) \\ &+ \left(\frac{1}{2}\sigma_t^2 S_t^2 \partial_s^2 P + \sigma_t^2 S_t K(t,t)\rho \partial_{s,\sigma}^2 P + \frac{1}{2}\sigma_t^2 K(t,t)^2 \partial_\sigma^2 P\right) dt. \end{aligned} \tag{A.5}$$

A.1.1 Couverture avec les contrats sur la variance future

L'idée du modèle à volatilité stochastique consiste à définir le processus de diffusion sur la variance future. Le prix de ce contrat est donné par

$$V_t(T) = \mathbb{E}\left[\sigma_T^2 \middle| \mathcal{F}_t\right],$$

pour une filtration $(\mathcal{F}_t)_{t\geq 0}$. Nous supposerons que le contrat sur la variance future actualisé de maturité T_i, noté U_t^i, défini par

$$U_t^i \stackrel{\text{def}}{=} e^{-r(T_i - t)} V_t(T_i), \quad i \in \{1 \dots m\}, \tag{A.6}$$

est traité sur le marché.

Remarque A.1 *Le contrat sur la variance future n'est pas un contrat traité sur le marché. Toutefois, il peut être approché par la différence entre deux contrats sur swap de variance définis par*

$$VS_t(\theta) = \left(\mathbb{E}\left[\frac{1}{\theta}\int_t^{t+\theta} \sigma_s^2 \middle| \mathcal{F}_t\right]\right)_{t\geq 0} = \left(\frac{1}{\theta}\int_t^{t+\theta} V_t(s) ds\right)_{t\geq 0}.$$

Ce paragraphe a pour objectif de démontrer que si le prix est supposé être une fonction du spot S_t, de la variance instantanée σ_t^2 et du temps, alors il n'est pas possible de se couvrir parfaitement, *i.e.* de construire un portefeuille de réplication qui annule les sources

de risque. Suivant la méthode de couverture en Δ proposée par Black & Scholes [BS73b], nous souhaitons annuler la source de risque sur la volatilité $d\widetilde{W}_t$ par une couverture de l'option à l'aide de contrats actualisés sur la variance future.

Cet actif suit le taux sans risque sous la probabilité risque neutre. Les contrats sur la variance future $V_t(T_i)$ vérifient l'EDS suivante

$$\frac{dV_t(T)}{V_t(T)} = K(T,t)d\widetilde{W}_t, \tag{A.7}$$

avec la condition initiale $V_0(T)$. La solution de cette EDS est donnée, pour tout $T > 0$, par

$$V_t(T) = V_0(T)\exp\left(\int_0^t K(T,s)d\widetilde{W}_s - \frac{1}{2}g(T,t)\right), \tag{A.8}$$

où

$$g(T,t) = var\left(\frac{dV_t(T)}{V_t(T)}\right) = \int_0^t K\left(T,s\right)^2 ds.$$

Notons que

$$\begin{aligned}\sigma_t = \sqrt{V_t\left(t\right)} &= \sqrt{V_0(t)}\exp\left(\frac{1}{2}\int_0^t K\left(t,s\right)d\widetilde{W}_s - \frac{1}{4}g(t,t)\right)\\ &= \sigma_0\left(t\right)\exp\left(\frac{1}{2}\int_0^t K\left(t,s\right)d\widetilde{W}_s\right)\end{aligned} \tag{A.9}$$

Nous essayons de construire une stratégie de réplication de l'option européenne à l'aide du portefeuille défini par :

- une quantité a_t de sous-jacent S_t
- m contrats actualisés sur la variance future $\left(U_t^i\right)_{1\leq i\leq m}$ de maturité T_i en quantité b_i,
- une option de date d'expiration T au prix $P = P\left(S_t,\sigma_t,t\right)$, où $P = P\left(t,S_t,U_t^1,\ldots,U_t^M\right)$.

Notons c_t la valeur de ce portefeuille. Le principe de non arbitrage implique que, pour tout $t < T$,

$$\begin{aligned}dc_t = a_t dS_t + dP + \sum_{i=1}^m b_t^i dU_t(T_i) &= r_t c_t dt = r_t\left(a_t S_t + P + \sum_{i=1}^m b_t^i U_t(T_i)\right)dt\\ a_t dS_t + dP + \sum_{i=1}^m b_t^i\left(r_t U_t(T_i)dt + e^{-r(T_i-t)}dV_t(T_i)\right) &= r_t\left(a_t S_t + P + \sum_{i=1}^m b_t^i U_t(T_i)\right)dt\\ a_t dS_t + dP + \sum_{i=1}^m b_t^i e^{-r(T_i-t)}dV_t(T_i) &= r_t(a_t S_t + P)dt,\\ a_t\left(r_t S_t dt - dS_t\right) + r_t P dt - \sum_{i=1}^m b_t^i e^{-r(T_i-t)}dV_t(T_i) &= dP.\end{aligned} \tag{A.10}$$

En introduisant l'Eq.(A.5), avec $e^{-r(T_i-t)}dV_t(T_i) = K\left(T_i,t\right)U_t\left(T_i\right)d\widetilde{W}_t$

$$\begin{aligned}\partial_t P dt + \partial_s P dS_t + \partial_\sigma P d\sigma_t + \left(\frac{1}{2}\sigma_t^2 S_t^2 \partial_s^2 P + \sigma_t^2 S_t K(t,t)\rho\partial_{s,\sigma}^2 P + \frac{1}{2}\sigma_t^2 K(t,t)^2\partial_\sigma^2 P\right)dt \\ = a_t\left(r_t S_t dt - dS_t\right) + r_t dt P - \sum_{i=1}^{m} b_t^i K\left(T_i,t\right)U_t\left(T_i\right)d\widetilde{W}_t.\end{aligned} \tag{A.11}$$

Les processus $(a_t)_{t\geq 0}$ et $\left(b_t^i\right)_{t\geq 0}$, $1 \leq i \leq m$ qui constituent notre portefeuille c_t sont choisis de manière à annuler les termes de risque. En posant a_t tel que

$$a_t + \frac{\partial P}{\partial s} = 0, \tag{A.12}$$

alors le risque lié à W_t est nul. De même, si

$$K(t,t)\sigma\partial_\sigma P + \sum_{i=1}^{m} b_t^i K(T_i,t)U_t(T_i) = 0, \tag{A.13}$$

alors le risque lié à $\widetilde{W}_t$ est annulé. L'Eq(A.11) devient, en divisant par dt

$$\begin{array}{rcl}\partial_t P & - & r_t S_t \partial_s P + \sigma_t\left(\phi(t) + \gamma_t\right)\partial_\sigma P \\ & + & \frac{1}{2}\sigma_t^2 S_t^2\partial_s^2 P + \sigma_t^2 S_t K(t,t)\rho\partial_{s,\sigma}^2 P + \frac{1}{2}\sigma_t^2 K(t,t)^2\partial_\sigma^2 P - r_t P = 0.\end{array} \tag{A.14}$$

Nous ne disposons plus de degré de liberté dans le choix des processus $\left(b_t^i\right)_{t\geq 0}$, $1 \leq i \leq m$ qui permettrait d'annuler le terme de drift stochastique dû au processus $(\gamma_t)_{t\geq 0}$.

A.1.2 Couverture à l'aide d'autres instruments financiers

Supposons que les contrats sur la variance future soient remplacés par des instruments dont la dynamique est donnée par

$$\frac{dV_t^\epsilon(T)}{V_t^\epsilon(T)} = \epsilon_T\left(T,t\right)dt + K(T,t)d\widetilde{W}_t. \tag{A.15}$$

Cette hypothèse revient à changer la probabilité risque neutre par rapport aux produits sur variance.

A partir d'Eq(A.7) et Eq(A.8), nous déduisons la relation liant $V_t^\epsilon(T)$ et $V_t(T)$:

$$\frac{V_t^\epsilon(T)}{V_0^\epsilon(T)} = \frac{V_t(T)}{V_0(T)} exp\left(\int_0^t \epsilon_T\left(T,u\right)du\right). \tag{A.16}$$

Nous supposons que $V_0^\epsilon(T) = V_0(T)$ et donc

$$V_t^\epsilon(T) = V_t(T)\exp\left(\int_0^t \epsilon_T\left(T,u\right)du\right). \tag{A.17}$$

Notons $U_t^\epsilon(T)$ le nouvel instrument actualisé. L'Eq(A.10) devient

$$dc_t = a_t dS_t + dP + \sum_{i=1}^m b_t^i dU_t^\epsilon(T_i) = r_t c_t dt = r_t \left(a_t S_t + P + \sum_{i=1}^m b_t^i U_t^\epsilon(T_i) \right) dt$$

$$a_t dS_t + dP + \sum_{i=1}^m b_t^i \left(r U_t^\epsilon(T_i) dt + e^{-r(T_i - t)} dV_t^\epsilon(T_i) \right) = r_t \left(a_t S_t + P + \sum_{i=1}^m b_t^i U_t^\epsilon(T_i) \right) dt$$

$$a_t dS_t + dP + \sum_{i=1}^m b_t^i e^{-r(T_i - t)} dV_t^\epsilon(T_i) = r_t (a_t S_t + P) dt,$$

$$a_t dt (r_t S_t - dS_t) + r_t dt P + \sum_{i=1}^m b_t^i U_t^\epsilon(T_i) \left(\epsilon_i (T_i, t)\, dt + K(T_i, t) d\widetilde{W}_t \right) = dP. \tag{A.18}$$

A nouveau, la stratégie, qui consiste à choisir a_t défini par Eq(A.12) et

$$K(t,t)\sigma \partial_\sigma P + \sum_{i=1}^m b_t^i K(T_i, t) U_t^\epsilon(T_i) = 0, \tag{A.19}$$

permet d'annuler le terme de risque sur W_t. L'Eq(A.14) devient

$$\begin{aligned} &\partial_t P - r_t S_t \partial_s P + \sigma_t \left(\phi(t) + \gamma_t \right) \partial_\sigma P + \frac{1}{2} \sigma_t^2 S_t^2 \partial_s^2 P \\ &+ \sigma_t^2 S_t K(t,t) \rho \partial_{s,\sigma}^2 P + \frac{1}{2} \sigma_t^2 K(t,t)^2 \partial_\sigma^2 P - r_t P = \sum_{i=1}^m b_t^i \epsilon_i \left(T_i, t \right) U_t^\epsilon \left(T_i \right). \end{aligned} \tag{A.20}$$

Supposons qu'il existe $\left(b_t^i \right)_{t \geq 0}$, $1 \leq i \leq m$ tel que

$$\gamma_t \sigma \partial_\sigma P = \sum_{i=1}^m b_t^i \epsilon_i \left(T_i, t \right) U_t^\epsilon(T_i), \tag{A.21}$$

alors P est solution de l'équation aux dérivées partielles

$$\begin{aligned} &\partial_t P - r_t S_t \partial_s P + \sigma_t \phi(t) \partial_\sigma P + \frac{1}{2} \sigma_t^2 S_t^2 \partial_s^2 P \\ &+ \sigma_t^2 S_t K(t,t) \rho \partial_{s,\sigma}^2 P + \frac{1}{2} \sigma_t^2 K(t,t)^2 \partial_\sigma^2 P - r_t P = 0. \end{aligned} \tag{A.22}$$

La formulation suivante incite à passer « à la limite » pour définir la probabilité risque neutre c'est à dire la dynamique des variance future (A.15). Supposons que (A.21) soit vérifiée et notons P_ϵ le prix sous la probabilité définie par (A.15) alors P_ϵ est solution de (A.22) pour tout $\epsilon > 0$.

Nous allons à présent démontrer que, dans le cas $K(t,s) = \sum_{k=1}^m \beta_k e^{-\lambda_k (t-s)}$, il est possible de trouver $\left(b_t^i \right)_{t \geq 0}$ solution de (A.21). Afin de rendre lisible le système d'équations vérifié par la stratégie $\left(b_t^1, \ldots, b_t^m \right)$, nous introduisons les processus $\left(Y_t^k \right)_{t \geq 0}$ définis par

$$Y_t^k = \int_0^t \beta_k e^{-\lambda_k (t-s)} d\widetilde{W}_s. \tag{A.23}$$

Alors

$$\gamma_t = -\sum_{k=1}^{m} \lambda_k Y_t^k, \quad \int_0^t K(T_i,s)d\widetilde{W}_s = \sum_{k=1}^{m} e^{-\lambda_k(T_i-t)}Y_t^k, \quad \sigma_t = \sigma_0(t)\exp\left(\frac{1}{2}\sum_{k=1}^{m} Y_t^k\right).$$

La stratégie $\left(b_t^i\right)_{t>0,i=1,\ldots,m}$ doit vérifier (A.19) et (A.21) pour que le prix de l'option européenne ne dépende que des valeurs de $(S_t)_{t>0}$, $(\sigma_t)_{t>0}$ à l'instant t et ce même si le processus de volatilité n'est pas markovien.

$$\sum_{i=1}^{m} b_t^i q\left(T_i,t\right)\exp\left(\sum_{k=1}^{m} e^{-\lambda_k(T_i-t)}Y_t^k\right) = -K(t,t)\sigma_0(t)\exp\left(\frac{1}{2}\sum_{k=1}^{m} Y_t^k\right)\partial_\sigma P, \tag{A.24}$$

$$\sum_{i=1}^{m} b_t^i q^\epsilon\left(T_i,t\right)\exp\left(\sum_{k=1}^{m} e^{-\lambda_k(T_i-t)}Y_t^k\right) = -\left(\sum_{k=1}^{m}\lambda_k Y_t^k\right)\sigma_0(t)\exp\left(\frac{1}{2}\sum_{k=1}^{m} Y_t^k\right)\partial_\sigma P,$$

avec

$$q\left(T_i,t\right) = K(T_i,t)e^{-r(T_i-t)}V_0(T_i) \text{ et } q^\epsilon\left(T_i,t\right) = \epsilon_i\left(T_i,t\right)\exp\left(\int_0^t \epsilon_i\left(T_i,u\right)du\right)e^{-r(T_i-t)}V_0(T_i).$$

Le système admet au moins une solution si

$$\exists i,j\ 1\le i,j\le m, \quad q^\epsilon\left(T_i,t\right)q\left(T_j,t\right) - q^\epsilon\left(T_j,t\right)q\left(T_i,t\right) \neq 0, \forall t\ge 0.$$

Dans le cas $\epsilon\left(T_i,t\right) = \epsilon$, cette condition est automatiquement vérifiée si $T_i \neq T_j$, (rappelons que $\beta_k > 0$).

A.2 Équation de valorisation sur les processus markoviens

A.2.1 Équation de valorisation sur les processus Y_t^i

Supposons que le prix P du contrat dépende des $m+2$ processus $(S_t, Y_t^1, \ldots, Y_t^m, t)$, en notant r_t le taux d'intérêt. La formule d'Itô multi-dimensionnelle permet d'exprimer dP comme combinaison linéaire de dt, dW_t et $d\widetilde{W}_t$. Cette formule est donnée par l'Eq(5.12). Nous cherchons alors à répliquer l'option à l'aide de m contrats actualisés sur la variance future. La dynamique du portefeuille de réplication c_t est donnée par l'Eq(A.10). Suivant la méthode décrite dans la section précédente, nous cherchons le système vérifié par les processus $(a_t)_{t\ge0}$ et $(b_t)_{t\ge0}$ pour que le portefeuille réplique l'option P.

Le processus $(a_t)_{t\ge0}$ est caractérisé par l'Eq(A.12). La condition de l'Eq(A.13) devient

$$\sum_{i}^{m}\beta_i\partial_{y_i}P + \sum_{i=1}^{m} b_t^i K(T_i,t)U_t(T_i) = 0, \tag{A.25}$$

Cette relation exprimée sur les processus $\left(Y_t^i\right)_{t>0}$, $1\le i\le m$ donne

$$\sum_{i}^{m}\beta_i\partial_{y_i}P + \sum_{i=1}^{m} b_t^i K(T_i,t)e^{-r_t(T_i-t)}V_0\left(T_i\right)\exp\left(\sum_{k=1}^{m} e^{-\lambda_k(Ti-t)}Y_t^k\right) = 0, \tag{A.26}$$

La stratégie définie par l'Eq(A.12) et

$$b_t^i\left(Y_t^1,\ldots,Y_t^m\right) = \partial_{y_i}P\frac{e^{r_t(T_i-t)}\beta_i}{K(T_i,t)V_0\left(T_i\right)}\exp\left(-\sum_{k=1}^{m} e^{-\lambda_k(Ti-t)}Y_t^k\right), \tag{A.27}$$

permet de répliquer l'option européenne P.

A.2.2 Équation de valorisation sur les processus U_t^i

Le prix de l'option européenne obtenue en considérant comme variables les processus $(S_t)_{t\geq 0}$ et $\left(U_t^i\right)_{t\geq 0}$, $\quad 1 \leq i \leq m$, vérifie l'équation

$$\frac{\partial P}{\partial t} + \frac{1}{2}\sigma^2 s^2 \frac{\partial^2 P}{\partial s^2} + \sum_i \rho\sigma^2 s u_i K(T_i,t)\frac{\partial^2 P}{\partial s \partial u_i} + \frac{1}{2}\sum_{i,j} u_i u_j K(T_i,t)K(T_j,t)\frac{\partial^2 P}{\partial u_i \partial u_j}$$
$$+r\left(s\frac{\partial P}{\partial s} - P\right) + r\sum_{i=1}^{n} u_i \frac{\partial P}{\partial u_i} = 0. \tag{A.28}$$

Preuve En utilisant la formule d'Itô et la stratégie donnée par $b_i^t\left(U_t^i\right) = -\partial_{u_i} P$ ■

Remarque A.2 *Cette équation permet de calculer directement les instruments de couverture.*

Remarque A.3 *Nous disposons d'une équation sous la forme multi-sous jacent avec une volatilité locale ce qui constitue une piste pour la recherche d'une formule semi-fermée.*

Expression de la volatilité en fonction des contrats sur la variance future Nous démontrons qu'il existe une bijection entre les processus $\left(Y_t^i\right)_{t>0}, 1 \leq i \leq m$ et $\left(V_t^k\right)_{t>0}, 1 \leq k \leq m$ pour un M fixé. Nous aurons besoin des processus

$$Z_t^k = \ln\left(\frac{V_t(T_k)}{V_0(t_K)}\right) = \int_0^t K(T_k,s)d\widetilde{W}_s = \sum_{i=1}^{m} e^{-\lambda_i(T_k-t)}Y_t^i. \tag{A.29}$$

En inversant cette relation matricielle, nous obtenons

$$\sum_{k=1}^{m} \alpha_{i,k}(t) Z_t^k = Y_t^i. \tag{A.30}$$

L'expression de la volatilité s'en déduit

$$\exp\left(\frac{1}{2}\sum_{i=1}^{m} Y_t^i\right) = \prod_{k=1}^{m}\left(\frac{V_t(T_i)}{V_0(T_i)}\right)^{\frac{1}{2}\sum_{i=1}^{m}\alpha_{i,k}(t)}. \tag{A.31}$$

Annexe B

Équation intégro-différentielle

Nous revenons sur les équations intégro-différentielles associées à une classe de processus markoviens, les processus de *Feller* dont font partie les diffusions de Lévy.

Par définition, un processus de Lévy (X_t) est un processus stationnaire à accroissements indépendants. Cette propriété implique que la distribution $\mu_t = L(X_{s+t} - X_s) = L(X_t - X_0)$ des incréments forme un semi-groupe pour l'opérateur de convolution dans l'espace des mesures de probabilité.

$$\mu_0 = \mathcal{I}, \qquad \mu_{t+s} = \mu_t \star \mu_s.$$

Cette propriété de semi-groupe pour l'opérateur de convolution implique l'existence d'une fonction $q : \mathbb{R}^d \to \mathbb{C}$ continue telle que

$$\widehat{\mu_t} = \exp(-tq(\xi)).$$

Pour toute distribution $(\mu_t)_{t\geq 0}$, la fonction q, nommée *l'exposant caractéristique*, caractérise complétement le processus de Lévy $(X_t)_{t\geq 0}$.

De plus, le semi-groupe pour l'opérateur de convolution $(\mu_t)_{t\geq 0}$ induit un semi-groupe noté T_t dans l'espace des fonctions $\mathcal{C}^\infty\left(\mathbb{R}^d\right)$, tel que $T_t u = u \star \mu_t$ qui est un semi-groupe de *Feller*. Dans cette situation, nous pouvons définir un *générateur infinitésimal* pour toute fonction $u \in \mathcal{C}^\infty\left(\mathbb{R}^d\right)$:

$$\begin{aligned} \mathcal{L}u(x) = \lim_{t\to 0} \frac{1}{t}\left(T_t u(x) - u(x)\right) = \lim_{t\to 0} \int_{\mathbb{R}^d} e^{\imath\langle x,\xi\rangle} \frac{e^{-tq(\xi)} - 1}{t} \widehat{u}(\xi) d\xi \\ = -\int_{\mathbb{R}^d} e^{\imath\langle x,\xi\rangle} q(\xi)\, \widehat{u}(\xi) d\xi. \end{aligned} \tag{B.1}$$

Dans le cas de processus qui ne sont pas invariants par translation, il est raisonable de supposer que le *générateur infinitésimal*, lorsqu'il existe, est de la forme

$$\mathcal{L}u(x) = -\int_{\mathbb{R}^d} e^{\imath\langle x,\xi\rangle} q(x,\xi)\, \widehat{u}(\xi) d\xi.$$

Nous supposerons que *le symbole de Fourier* $q : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{C}$ est une fonction continue définie négative pour tout $x \in \mathbb{R}^d$.

Définition B.1 (voir [Hoh94]) *Une fonction* $q : \mathbb{R}^d \to \mathbb{C}$ *est une fonction* définie négative *si pour tout* $m \in \mathbb{N}$ *et pour tout* $m-$*uplet* $\xi^j \in \mathbb{R}^n$, $1 \leq j \leq m$, *la matrice*

$$\left(q\left(\xi^i\right) + \overline{q\left(\xi^j\right)} - q\left(\xi^i - \xi^j\right)\right)_{i,j=1,\ldots,m}$$

est une matrice Hermitienne, i.e. *pour tout* $c_1, \ldots, c_m \in \mathbb{C}$

$$\sum_{i,j=1}^{m} \left[q\left(\xi^i\right) + \overline{q\left(\xi^j\right)} - q\left(\xi^i - \xi^j\right)\right] c_i \overline{c}_j \geq 0.$$

L'objet de cette partie est de présenter succintement la notion de symbole de Fourier et d'opérateur pseudo-différentiel. Ces notions nous permettront de formaliser trois problèmes :

1. Le problème d'existence et d'unicité de la *formulation faible* d'une équation parabolique associée à l'opérateur $\mathcal{L}$.
2. Les propriétés de décroissance du noyau de convolution associé à l'opérateur $\mathcal{L}$. Nous cherchons des estimateurs de la forme (2.186).
3. Une méthode de calcul des coefficients de la matrice de rigidité associée à la formulation faible pour mettre en oeuvre des méthodes numériques.

Sur ces trois questions, nous ne disposons que de résultats partiels, le plus souvent obtenus pour des symboles bornés par rapport à x.

B.1 Opérateur pseudo-différentiel

Soit la transformée de Fourier définie dans $L^2\left(\mathbb{R}^d\right)$,

$$\widehat{u}(\xi) = \mathcal{F}(u)(\xi) = \int_{\mathbb{R}^d} u(x) \exp\left(-\imath \langle \xi, x \rangle\right) dx, \qquad \forall\ u \in L^2\left(\mathbb{R}^d\right), \tag{B.2}$$

et l'inverse de cette transformation

$$\mathcal{F}^{-1}(\hat{u})(x) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \hat{u}(\xi) \exp\left(\imath \langle \xi, x \rangle\right) d\xi. \tag{B.3}$$

Définition B.2 (Opérateur pseudo-différentiel) *La fonction* q *est définie comme le* symbole de Fourier *de l'opérateur pseudo-différentiel* $\mathcal{L}$ *si*

$$\begin{aligned}(\mathcal{L}u)(x) &= -\int_{\mathbb{R}^d \times \mathbb{R}^d} \exp\left(\imath \langle x - y, \xi \rangle\right) q(x,\xi) u(y) \frac{d\xi\ dy}{(2\pi)^d} \\ &= -\int_{\mathbb{R}^d} q(x,\xi) \mathcal{F}(u)(\xi) \exp\left(\imath \langle x, \xi \rangle\right) \frac{d\xi}{(2\pi)^d}.\end{aligned} \tag{B.4}$$

L'intégrale est définie si le *symbole de Fourier* $q(x,\xi)$ vérifie les propriétés suivantes (voir (B.5) ci-dessous) :

- la fonction $q(x,\xi)$ et ses dérivées doivent avoir une croissance tempérée lorsque $|\xi| \to \infty$.
- la fonction $q(x,\xi)$ doit avoir une variation lente par rapport aux variables d'espace x.

Une classe particulière d'opérateur borné en x est donnée ci-dessous. Pour ces opérateurs, il est possible d'obtenir des résultats sur la formulation faible du problème parabolique associé. Ces résultats sont une conséquence des deux propriétés rappellées ci-dessous.

Définition B.3 (Classe des symboles d'ordre m) *Soient q une fonction lisse de $\mathcal{C}^\infty\left(\mathbb{R}^d \times \mathbb{R}^d\right)$ à valeur dans $\mathbb{C}$ et m un nombre réel quelconque. La classe $S^m\left(\mathbb{R}^d \times \mathbb{R}^d\right)$ des symboles d'ordre m est définie comme*

$$\left\{q \,\middle|\, \forall\ \alpha, \beta\ \exists C \in \mathbb{R}^+ \text{ telles que } \left|\partial_\xi^\alpha \partial_x^\beta q(x,\xi)\right| \leq C\left(1\ + |\xi|\right)^{m-|\alpha|}\ \forall\ x \in \mathbb{R}^d,\ \forall \xi \in \mathbb{R}^d\right\}. \tag{B.5}$$

Propriété B.1 ([Hör07] théorème 18.1.13) *Si q appartient à $S^m\left(\mathbb{R}^d \times \mathbb{R}^d\right)$ alors l'opérateur pseudo-différentiel $\mathcal{L}$ associé à q par la relation (B.4) est continu de $H^s\left(\mathbb{R}^d\right)$ dans $H^{s-m}\left(\mathbb{R}^d\right)$.*

Propriété B.2 ([Hör07] théorème 18.1.14) *Si q appartient à $S^{2m+1}\left(\mathbb{R}^d \times \mathbb{R}^d\right)$ et $\Re(q) \geq 0$, alors la forme bilinéaire associée à l'opérateur pseudo-différentiel vérifie*

$$\Re\left(\mathcal{L}u, u\right) \geq -C\|u\|^2_{H^m\left(\mathbb{R}^d\right)}, \quad u \in D\left(\mathbb{R}^d\right).$$

B.2 Quelques résultats sur les processus de Feller

B.2.1 Symbole de Fourier & exposant caractéristique

Le calcul proposé ici est purement formel, le lecteur trouvera dans [Jac98] la justification du résultat suivant (sous certaines hypothèses de régularité sur le symbole q). Notons $\mathbb{E}\left[X_t^{0,x}\right] = \mathbb{E}^x\left[X_t\right]$, alors

$$-\,q(x,\xi) = \frac{d}{dt}\mathbb{E}^x\left[e^{\imath\langle X_t - x, \xi\rangle}\right]_{|t=0} = \frac{d}{dt}\Phi_{X_t^{0,x}}(\xi)_{|t=0}. \tag{B.6}$$

En effet, par définition du symbole de Fourier q (B.4),

$$\mathcal{L}\left(u\right)(x) = \mathcal{F}^{-1}\left[-q\left(x,\cdot\right)\mathcal{F}(u)\right](x), \tag{B.7}$$

or

$$\begin{aligned}
\mathcal{L}\left(u\right)(x) &= \lim_{t\to 0}\frac{1}{t}\left(\mathbb{E}\left[u\left(X_t^{0,x}\right)\right] - u(x)\right) \\
&= \mathcal{F}^{-1}\left[\lim_{t\to 0}\frac{1}{t}\left(\mathbb{E}\left[\mathcal{F}\left(u\left(X_t^{0,x}\right)\right)\right] - \mathcal{F}(u)\right)\right](x) \\
&= \mathcal{F}^{-1}\left[\lim_{t\to 0}\frac{1}{t}\left(\Phi_{X_t^{0,x}}\mathcal{F}\left(u\right) - \mathcal{F}(u)\right)\right](x) \\
&= \mathcal{F}^{-1}\left[\lim_{t\to 0}\frac{1}{t}\left(\Phi_{X_t^{0,x}} - 1\right)\mathcal{F}(u)\right](x).
\end{aligned} \tag{B.8}$$

En identifiant (B.7) et (B.8), nous obtenons la formulation suivante

$$-q(x,\xi)=\lim_{t\to 0}\frac{1}{t}\left(\Phi_{X_t^{0,x}}(\xi)-1\right)=\frac{d}{dt}\Phi_{X_t^{0,x}}(\xi)_{|t=0}\,. \tag{B.9}$$

Afin de généraliser la représentation de Lévy-Khintchine au cas des diffusions de Lévy, *l'exposant caractéristique* est défini par (B.9). De cette manière, *l'exposant caractéristique* est toujours égal au *symbole de Fourier* du *générateur infinitésimal* associé au processus de Feller.

B.2.2 Symbole de Fourier d'une diffusion de Lévy

Proposition B.3 *Le symbole de Fourier d'une diffusion de Lévy est donné, d'après la formule de Itô, par*

$$\begin{aligned} q(t,x,\xi)=&\imath\langle\alpha(x),\xi\rangle-\tfrac{1}{2}\sum_{i,j=1}^{n}\xi_i\Xi_{ij}(x)\xi_j \\ &+\int_{\mathbb{R}^d}\left[\sum_{k=1}^{\ell}e^{\imath\gamma^{(k)}(t,x,z)\xi_k}-1-\imath\gamma^{(k)}(t,x,z)\xi_k 1_{|z|\le R}\right]\nu(dz)\ . \end{aligned} \tag{B.10}$$

B.2.3 Condition pour qu'une diffusion de Lévy soit martingale

Lemme B.4 *Soient $X_t^{0,x}=X_t$ une diffusion de Lévy à valeurs dans $\mathbb{R}^d$ et le triplet caractéristique (Ξ,ν,α). Supposons que $\int_{|z|\ge R}e^{\gamma_k(x,z)}\nu(dz)\le\infty,\ k=1,\ldots,d$. Alors e^{X^k} est une martingale par rapport à la filtration canonique $\mathcal{F}_t$ de X si*

$$\frac{\Xi_{kk}}{2}+\alpha_k+\int_{\mathbb{R}^d}\left(e^{\gamma_k(x,z)}-1-\gamma_k(x,z)1_{|z|\le R}\right)\nu(dz)=0. \tag{B.11}$$

Preuve A partir de (B.6), nous obtenons pour $\xi_k=-i,\ \xi_j=0, j\neq k$,

$$-q(x,(0,\ldots,0,-\imath,0,\ldots,0))=\frac{d}{dt}\mathbb{E}^x\left[e^{X_t^k-x^k}\right]_{|t=0}=e^{-x^k}\frac{d}{dt}\mathbb{E}^x\left[e^{X_t}\right]_{|t=0}. \tag{B.12}$$

Nous en déduisons que si $\dfrac{d}{dt}\mathbb{E}^x\left[e^{X_t}\right]=0$, alors $q(x,(0,\ldots,0,-\imath,0,\ldots,0))=0$, ce qui implique (B.11). ■

B.3 Méthode de Galerkin & Opérateur pseudo-différentiel

Nous discutons dans cette partie de l'existence d'une formulation faible dans le cas de l'équation parabolique

$$\frac{\partial u}{\partial t}-\mathcal{L}u=0,\quad u(0)=u_0, \tag{B.13}$$

où l'opérateur $\mathcal{L}$ est associé au symbole de Fourier $q:\mathbb{R}^d\times\mathbb{R}^d$ par (B.4). Nous étudierons par la suite les symboles de Fourier associés à des *diffusions de Lévy*. Dans ce cas $\mathcal{L}u$ est à valeurs réelles si u est à valeurs réelles.

L'objectif de cette partie est double : nous souhaitons, d'une part, obtenir un résultat d'existence et d'unicité et, d'autre part, pouvoir calculer les coefficients de la matrice de rigidité associée à l'opérateur $\mathcal{L}$.

B.3.1 Formulation faible

Considérons la forme de Dirichlet $a(\cdot,\cdot)$ associée à l'opérateur $\mathcal{L}$.

Rappels sur les équations paraboliques pour les opérateurs pseudo-différentiels Soit $\mathcal{L} : \mathcal{V} \to \mathcal{V}'$ une forme linéaire continue avec $\mathcal{V}'$ l'espace dual de $\mathcal{V}$, et $a(\cdot,\cdot) : \mathcal{V}\times\mathcal{V} \to \mathbb{R}$ la forme bilinéaire continue $a(u,v) = \langle \mathcal{L}u, v\rangle$. S'il existe un espace $\mathcal{H}$ tel que

$$\mathcal{V} \hookrightarrow \mathcal{H} \approx \mathcal{H}' \hookrightarrow \mathcal{V}', \quad \text{avec injection dense,}$$

et si $a(\cdot,\cdot)$ vérifie une inégalité de Gårding : il existe C, c telles que

$$a(u,u) \geq C\|u\|^2_{\mathcal{V}} - c\|u\|^2_{\mathcal{H}},$$

alors le problème (B.13) avec $u_0 \in \mathcal{H}$ est bien posé dans $L^2(]0,T[;\mathcal{V}) \cap \mathcal{C}([0,T];\mathcal{H})$ et $\dfrac{\partial u}{\partial t} \in L^2(]0,T[;\mathcal{V}')$.

Nous allons montrer l'inégalité de Gårding dans le cas de l'opérateur associé au symbole de Fourier q. Dans ce qui suit, nous considérons $\mathcal{H} = L^2\left(\mathbb{R}^d\right)$.

Formellement,

$$a(u,v) = \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} \mathcal{F}(u)(\xi)\,\overline{\mathcal{F}(-\overline{q}(\cdot,\xi)\,v)}(\xi)\,d\xi. \tag{B.14}$$

En effet,

$$a(u,v) = (\mathcal{L}u,v)_{L^2(\mathbb{R}^d)} = \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} -q(x,\xi)\exp(\imath\langle x,\xi\rangle)\,v(x)\mathcal{F}(u)(\xi)\,d\xi dx. \tag{B.15}$$

En intégrant suivant x, nous obtenons(B.14) en remarquant que $v(x) \in \mathbb{R}$ et

$$q(x,\xi)\exp(\imath\langle x,\xi\rangle) = \overline{\overline{q}(x,\xi)\exp(-\imath\langle x,\xi\rangle)}.$$

Hypothèse B.1 *Nous supposerons dans tout ce qui suit que*

$$\infty > a(u,u) \geq 0,\ \forall u \in D\left(\mathbb{R}^d\right), \tag{B.16}$$

où $D\left(\mathbb{R}^d\right)$ est l'ensemble des fonctions continûment dérivables à support compact.

Sur l'espace $D\left(\mathbb{R}^d\right)$ nous introduisons

– la semi-norme :

$$|u|_{\mathcal{H}_q(\mathbb{R}^d)} = \sqrt{a(u,u)}. \tag{B.17}$$

– la norme :

$$\|u\|_{\mathcal{H}_q(\mathbb{R}^d)} = \sqrt{\|u\|^2_{L^2(\mathbb{R}^d)} + a(u,u)}. \tag{B.18}$$

– et le produit scalaire

$$\langle u,v\rangle_{\mathcal{H}_q\left(\mathbb{R}^d\right)}=\langle u,v\rangle_{L^2\left(\mathbb{R}^d\right)}+\frac{1}{2}\left(a\left(u,v\right)+a\left(v,u\right)\right). \tag{B.19}$$

Soit l'espace $\mathcal{H}_q\left(\mathbb{R}^d\right)$ défini comme l'adhérence de $D\left(\mathbb{R}^d\right)$ pour la norme $\|\cdot\|_{\mathcal{H}_q\left(\mathbb{R}^d\right)}$, alors $\mathcal{H}_q\left(\mathbb{R}^d\right)$ est un espace de Hilbert. La norme sur $\mathcal{H}_q\left(\mathbb{R}^d\right)$ définie par (B.18) implique $\mathcal{H}_q\left(\mathbb{R}^d\right)\hookrightarrow L^2\left(\mathbb{R}^d\right)$, avec injection dense puisque les fonctions régulières à support compact sont denses dans $L^2\left(\mathbb{R}^d\right)$.

Remarque B.1 *Pour certains opérateurs pseudo-différentiels, il est possible d'établir l'équivalent du lemme de Friedrich obtenu dans le cas différentiel à coefficients lipschitziens [Hör07]. Certains de ces résultats sont démontrés dans [Car71].*

Théorème B.5 *Sur l'espace $\mathcal{H}_q$ muni de la norme $\|\cdot\|_{\mathcal{H}_q\left(\mathbb{R}^d\right)}$, la forme bilinéaire définie sur $\mathcal{H}_q\left(\mathbb{R}^d\right)\times\mathcal{H}_q\left(\mathbb{R}^d\right)$ vérifie l'inégalité de Gårding :*

$$a\left(u,u\right)\geq\|u\|^2_{\mathcal{H}_q\left(\mathbb{R}^d\right)}-\|u\|^2_{L^2\left(\mathbb{R}^d\right)}. \tag{B.20}$$

Nous introduisons la partie symétrique et la partie non-symétrique de a :

$$\begin{aligned} a_s\left(u,v\right)&=\frac{1}{2}\left(a(u,v)+a(v,u)\right), \\ a_a\left(u,v\right)&=\frac{1}{2}\left(a(u,v)-a(v,u)\right). \end{aligned} \tag{B.21}$$

Suivant cette définition,

$$a(u,u)=a_s(u,u). \tag{B.22}$$

Nous distinguons deux cas :

1. *Si il existe une constante C telle que*

$$\mid a_a\left(u,v\right)\mid\leq C\mid a_s\left(u,v\right)\mid,\qquad\forall u,v\in\mathcal{H}_q\left(\mathbb{R}^d\right), \tag{B.23}$$

alors la forme bilinéaire a est continue sur $\mathcal{H}_q\left(\mathbb{R}^d\right)\times\mathcal{H}_q\left(\mathbb{R}^d\right)$.
Nous en déduisons que, pour un temps $T>0$ et $u_0\in L^2\left(\mathbb{R}^d\right)$, il existe une unique fonction $u\in L^2\left(]0,T[;\mathcal{H}_q\left(\mathbb{R}^d\right)\right)\cap\mathcal{C}\left([0,T];L^2\left(\mathbb{R}^d\right)\right)$ telle que u vérifie l'équation (B.13).

2. *Si il existe deux constantes $C1$ et $C2$ telles que*

$$|a_a\left(u,v\right)|\leq C_1\left|a_s(u,v)\right|+C_2\|u\|_{H^1\left(\mathbb{R}^d\right)}\|v\|_{H^1\left(\mathbb{R}^d\right)}. \tag{B.24}$$

alors la forme bilinéaire a est continue de $\left(\mathcal{H}_q\left(\mathbb{R}^d\right)\cap H^1\left(\mathbb{R}^d\right)\right)\times\left(\mathcal{H}_q\left(\mathbb{R}^d\right)\cap H^1\left(\mathbb{R}^d\right)\right)$. En suivant ce qui est proposé dans la partie 4.2.3, nous pouvons montrer que le problème régularisé

$$\frac{\partial u_\varepsilon}{\partial t}-\mathcal{L}u_\varepsilon-\epsilon\Delta u_\varepsilon=0,\quad u_\varepsilon\left(0\right)=u_0, \tag{B.25}$$

admet une unique solution. Nous pouvons en déduire que pour un temps $T > 0$ *et* $u_0 \in H^1\left(\mathbb{R}^d\right)$*, il existe une unique fonction* $u \in L^2\left(]0,T[;\mathcal{H}_q\left(\mathbb{R}^d\right)\right) \cap \mathcal{C}\left([0,T];L^2\left(\mathbb{R}^d\right)\right)$ *telle que u vérifie l'équation (B.13) et qui soit obtenue comme limite de la suite des solutions du problème régularisé (B.25).*

Preuve

Le résultat (B.20) se déduit de la définition de la norme (B.18) sur l'espace $\mathcal{H}_q\left(\mathbb{R}^d\right)$.

$$|a(u,v)|^2 = \left(a_s(u,v) + a_a(u,v)\right)^2 \leq C\left(a_s(u,v)^2 + a_a(u,u)^2\right). \tag{B.26}$$

Si (B.23) est vérifiée alors

$$|a(u,v)|^2 = \left(a_s(u,v) + a_a(u,v)\right)^2 \leq C\left(a_s(u,v)^2\right) \leq Ca(u,u)a(v,v). \tag{B.27}$$

■

Remarque B.2 *Une question délicate consiste à trouver les hypothèses sur le symbole q pour que (B.16) soit vérifiée. Dans le cas d'un symbole constant par rapport à x, cette hypothèse est équivalente à l'hypothèse que* $\Re q\left(\xi\right) \geq 0, \quad \forall\ \xi$*, la partie anti-symétrique étant liée à la partie imaginaire du symbole. En reprenant (B.4) :*

$$a(u,v) + a(v,u) = -\int_{\left\{\mathbb{R}^d\right\}^3} \exp\left(\imath\left\langle x-y,\xi\right\rangle\right) q(\xi)\left(u(y)v(x) + v(y)u(x)\right)\frac{d\xi\ dy\ dx}{(2\pi)^d} \tag{B.28}$$

$$\begin{aligned}
&= -\int_{\left\{\mathbb{R}^d\right\}^3} \exp\left(\imath\left\langle x-y,\xi\right\rangle\right) q(\xi)u(y)v(x)\frac{d\xi\ dy\ dx}{(2\pi)^d} \\
&\quad -\int_{\left\{\mathbb{R}^d\right\}^3} \exp\left(\imath\left\langle -x+y,\xi\right\rangle\right) \overline{q}(\xi)v(y)u(x)\frac{d\xi\ dy\ dx}{(2\pi)^d} \\
&= -\int_{\left\{\mathbb{R}^d\right\}^3} \exp\left(\imath\left\langle x-y,\xi\right\rangle\right)\left(q(\xi) + \overline{q}(\xi)\right)u(y)v(x)\frac{d\xi\ dy\ dx}{(2\pi)^d} \\
&= -\int_{\left\{\mathbb{R}^d\right\}^3} \exp\left(\imath\left\langle x-y,\xi\right\rangle\right) 2\Re q(\xi)u(y)v(x)\frac{d\xi\ dy\ dx}{(2\pi)^d}
\end{aligned} \tag{B.29}$$

Remarque B.3 *D'après la propriété B.1 si l'opérateur* $q \in S^m,\ m \geq 1$*, alors* $\mathcal{L}$ *est continu de* $H^1\left(\mathbb{R}^d\right)$ *dans* $L^2\left(\mathbb{R}^d\right)$*. Nous nous trouvons alors dans le cadre d'application du théorème sous sa version « limite de problèmes régularisés ».*

B.3.2 Calcul des coefficients de la matrice de rigidité

Dans le cas d'un symbole de Fourier q ne dépendant pas de x, le calcul des coefficients de la matrice de rigidité, associée la forme bilinéaire (B.14), est donnée par une fonction calculable pour la plupart des symboles *i.e.*$\left(\xi^\alpha, 0 \leq \alpha \leq 2;\quad \exp\left(-\frac{(\xi-\mu)^2}{2\eta^2}\right), \ldots\right)$.

Proposition B.6 *Soient $q : \mathbb{R}^d \to \mathbb{C}$ un symbole de Fourier associé à l'opérateur bilinéaire a et $\psi_{\ell,i}$ la base d'ondelettes. Les coefficients de la matrice de rigidité associée à la forme bilinéaire $a(\psi_{\ell,i}, \psi_{\ell,j})$ sont obtenus par la transformée de Fourier inverse de $\xi \to q\left(2^\ell \xi\right)\left|\widehat{\psi}(\xi)\right|^2$.*

La démonstration découle des deux lemmes suivants.

Lemme B.7 *La relation entre l'ondelette $\psi_{\ell,i}$ et l'ondelette mère ψ permet de caractériser la transformée de Fourier de $\psi_{\ell,i}$ à partir de celle de ψ :*

$$\widehat{\psi}_{\ell,i}(\xi) = 2^{-\ell/2} e^{-\imath\langle 2^{-\ell}\xi, i\rangle}\widehat{\psi}\left(\xi 2^{-\ell}\right). \tag{B.30}$$

Preuve

$$\widehat{\psi}_{\ell,i}(\xi) = \int_{\mathbb{R}} 2^{\ell/2}\psi\left(2^\ell x - i\right) e^{-\imath\langle\xi,x\rangle} dx = \int_{\mathbb{R}} 2^{-\ell/2}\psi(y)\, e^{-\imath\left\langle 2^{-\ell}\xi, y+i\right\rangle} dy.$$

■

Lemme B.8 *Soit $q : \mathbb{R}^d \to \mathbb{C}$ un symbole de Fourier associé à l'opérateur bilinéaire a, alors l'égalité de Parseval permet de montrer que*

$$a(u,v) = \frac{1}{(2\pi)^d}\left\langle -q\ \widehat{u}, \widehat{v}\right\rangle_{L^2(\mathbb{R}^d)}. \tag{B.31}$$

Preuve de la proposition B.6

$$a(\psi_{\ell,i}, \psi_{\ell,j}) = \frac{-1}{2\pi} 2^{-\ell}\int_{\mathbb{R}} q(\xi)\left|\widehat{\psi}\left(2^{-\ell}\xi\right)\right|^2 e^{-\imath\langle\xi, i-j\rangle} d\xi, = \frac{-1}{2\pi}\int_{\mathbb{R}} q\left(2^\ell\xi\right)\left|\widehat{\psi}(\xi)\right|^2 e^{-\imath\langle\xi,i-j\rangle} d\xi. \tag{B.32}$$

Soit

$$\widehat{g}_\ell(\xi) = -q\left(2^\ell\xi\right)\left|\widehat{\psi}(\xi)\right|^2, \quad \text{alors } a(\psi_{\ell,i}, \psi_{\ell,j}) = g_\ell(-(i-j)). \tag{B.33}$$

■

Remarque B.4 *La transformée de Fourier de la fonction chapeau φ est donnée par*

$$\widehat{\varphi}(\xi) = \left(\frac{\sin(\xi/2)}{\xi/2}\right)^2. \tag{B.34}$$

Annexe C

Méthode de Galerkin sur une base d'ondelettes de $[a,b]$

Nous détaillons dans cette partie le calcul sur les fonctions chapeaux de coefficients de matrices de masse et de rigidité. Le premier paragraphe concerne le passage d'un domaine $[a,b]$ à un domaine $[0,1]$ et la reconstruction des coefficients des matrices d'opérateurs sur $[a,b]$ à partir des coefficients obtenus sur $[0,1]$.Un second paragraphe donne les formules explicites pour les coefficients de différentes formes bilinéaires sur une base de fonctions chapeaux de raffinement ℓ définie sur $[0,1]$.

Pour rappel,

$$\varphi(x) = \begin{cases} 1-|x| & \text{si } x \in [-1,1], \\ 0 & \text{sinon.} \end{cases}$$

C.1 Passage de $[a,b]$ à $[0,1]$ et réciproque

Les fonctions de base sur $[a,b]$, notées ${}^{[a,b]}\varphi_{\ell,i}$, sont choisies de sorte que leur norme L^2 ne dépendent pas du domaine.

$$\begin{aligned} {}^{[a,b]}\varphi_{\ell,i}(x) &= (b-a)^{\frac{1}{2}}\,\varphi_{\ell,i}\left(\frac{x-a}{b-a}\right) = (b-a)^{\frac{1}{2}}\,2^{\ell/2}\varphi\left(2^{\ell}\frac{x-a}{b-a}-i\right), \\ {}^{[a,b]}\varphi'_{\ell,i}(x) &= (b-a)^{-\frac{1}{2}}\,2^{3\ell/2}\varphi'\left(2^{\ell}\frac{x-a}{b-a}-i\right). \end{aligned} \tag{C.1}$$

A partir de ces définitions, nous appliquons les formules de changement de base pour passer des coefficients de la matrice de rigidité sur $[0,1]$ à ceux sur $[a,b]$.

$$\begin{aligned} \int f(x)\,{}^{[a,b]}\varphi_{\ell,i}(x)\,{}^{[a,b]}\varphi_{\bar{\ell},\bar{i}}(x)\,dx &= \int f\left((1-\theta)a+\theta b\right)\varphi_{\ell,i}(\theta)\,\varphi_{\bar{\ell},\bar{i}}(\theta)\,d\theta \\ \int f(x)\,{}^{[a,b]}\varphi'_{\ell,i}(x)\,{}^{[a,b]}\varphi_{\bar{\ell},\bar{i}}(x)\,dx &= \frac{1}{b-a}\int f\left((1-\theta)a+\theta b\right)\varphi'_{\ell,i}(\theta)\,\varphi_{\bar{\ell},\bar{i}}(\theta)\,d\theta \\ \int f(x)\,{}^{[a,b]}\varphi'_{\ell,i}(x)\,{}^{[a,b]}\varphi'_{\bar{\ell},\bar{i}}(x)\,dx &= \frac{1}{(b-a)^2}\int f\left((1-\theta)a+\theta b\right)\varphi'_{\ell,i}(\theta)\,\varphi'_{\bar{\ell},\bar{i}}(\theta)\,d\theta. \end{aligned} \tag{C.2}$$

C.2 Quelques matrices associées à des opérateurs

Quelques matrices d'opérateurs classiques.

1. Matrice de Masse :

$$\int \varphi_{\ell,i}(x)\,\varphi_{\ell,\bar{i}}(x)\,dx = \begin{cases} \frac{2}{3} & \text{si} \quad i=\bar{i} \text{ et } \frac{1}{3} \text{ au bord} \\ \frac{1}{6} & \text{si} \quad i=\bar{i}+1 \text{ ou } i=\bar{i}-1 \\ 0 & \text{sinon} \end{cases} \tag{C.3}$$

2. Matrice de Masse avec poids exponentiel :

$$\int e^{\alpha x}\varphi_{\ell,i}(x)\,\varphi_{\ell,\bar{i}}(x)\,dx = \frac{e^{\gamma i}}{\gamma^2}\begin{cases} 2\left(\frac{e^{\gamma}-e^{-\gamma}}{\gamma}-2\right) & \text{si} \quad i=\bar{i} \\ e^{-\gamma}\left(1+\frac{2}{\gamma}\right)+1-\frac{2}{\gamma} & \text{si} \quad i=\bar{i}+1 \\ e^{\gamma}\left(1-\frac{2}{\gamma}\right)+1+\frac{2}{\gamma} & \text{si} \quad i=\bar{i}-1 \\ 0 & \text{sinon} \end{cases}, \tag{C.4}$$

avec $\gamma = \alpha 2^{-\ell}$.

3. Matrice associée à l'opérateur $\dfrac{\partial}{\partial x}$:

$$\int \varphi'_{\ell,i}(x)\,\varphi_{\ell,\bar{i}}(x)\,dx = 2^{\ell-1}\begin{cases} 0 & \text{si} \quad i=\bar{i}, \quad 1 \text{ si } i=2^{\ell} \text{ et } \; -1 \text{ si } i=0 \\ 1 & \text{si} \quad i=\bar{i}+1 \\ -1 & \text{si} \quad i=\bar{i}-1 \\ 0 & \text{sinon} \end{cases}. \tag{C.5}$$

4. Matrice associée à l'opérateur $x\partial_x$:

$$\int x\varphi'_{\ell,i}(x)\,\varphi_{\ell,\bar{i}}(x)\,dx = \begin{cases} \frac{-1}{3} & \text{si} \quad i=\bar{i}, \quad 2^{\ell-1}-\frac{1}{6} \text{ si } i=2^{\ell} \text{ et } \frac{-1}{6} \text{ si } i=0 \\ \frac{-1}{3}+\frac{i}{2} & \text{si} \quad i=\bar{i}+1 \\ \frac{-1}{3}-\frac{i}{2} & \text{si} \quad i=\bar{i}-1 \\ 0 & \text{sinon} \end{cases}. \tag{C.6}$$

5. Matrice associée à l'opérateur $e^{\alpha x}\partial_x$:

$$\int e^{\alpha x}\varphi'_{\ell,i}(x)\,\varphi_{\ell,\bar{i}}(x)\,dx = 2^{\ell}\frac{e^{\gamma i}}{\gamma}\begin{cases} 2+\frac{e^{-\gamma}-e^{\gamma}}{\gamma} & \text{si} \quad i=\bar{i} \\ -e^{-\gamma}\left(1+\frac{1}{\gamma}\right)+\frac{1}{\gamma} & \text{si} \quad i=\bar{i}+1 \\ -e^{\gamma}\left(1-\frac{1}{\gamma}\right)-\frac{1}{\gamma} & \text{si} \quad i=\bar{i}-1 \\ 0 & \text{sinon} \end{cases}, \tag{C.7}$$

avec $\gamma = \alpha 2^{-\ell}$.

6. Matrice associée au Laplacien :

$$\int \varphi'_{\ell,i}(x)\,\varphi'_{\ell,\bar{i}}(x)\,dx = 2^{2\ell}\begin{cases} 2 & \text{si} \quad i=\bar{i} \\ -1 & \text{si} \quad i=\bar{i}+1 \text{ ou } i=\bar{i}-1 \\ 0 & \text{sinon} \end{cases}. \tag{C.8}$$

7. Matrice associée à l'opérateur de second ordre $x^2\partial_x^2$:

$$\int x^2\varphi'_{\ell,i}(x)\,\varphi'_{\ell,\bar{i}}(x)\,dx = \begin{cases} 2\left(\dfrac{1}{3}+i^2\right) & \text{si} \quad i=\bar{i} \\ \dfrac{-1}{3}+i-i^2 & \text{si} \quad i=\bar{i}+1 \\ \dfrac{-1}{3}-i-i^2 & \text{si } i=\bar{i}-1 \\ 0 & \text{sinon} \end{cases}. \tag{C.9}$$

Remarque C.1 *Les calculs sont comparés avec les résultats obtenus à l'aide de Matlab. Dans le cas des poids exponentiels, un développement limités montre que les formules sont cohérentes.*

Preuve

– Équation (C.4),

$$\begin{aligned} I_{\ell,i,\bar{i}} &= \int e^{\alpha x}\varphi_{\ell,i}(x)\,\varphi_{\ell,\bar{i}}(x)\,dx \\ &= \int_{-1}^{1} e^{\alpha 2^{-\ell}(\theta+i)}\varphi(\theta)\,\varphi(\theta-(\bar{i}-i))\,d\theta \\ \text{nous notons} \quad & \gamma=\alpha 2^{-\ell} \\ &= e^{\gamma i}\int_{-1}^{1} e^{\gamma\theta}\varphi(\theta)\,\varphi(\theta-(\bar{i}-i))\,d\theta. \end{aligned} \tag{C.10}$$

alors,

$$\begin{aligned} I_{\ell,i,i} &= e^{\gamma i}\int_{-1}^{1} e^{\gamma\theta}\varphi^2(\theta)\,d\theta \\ &= e^{\gamma i}\left(\int_{-1}^{0} e^{\gamma\theta}(1+\theta)^2\,d\theta+\int_{0}^{1} e^{\gamma\theta}(1-\theta)^2\,d\theta\right) \\ &= e^{\gamma i}\left(\int_{0}^{1}\left(e^{\gamma\theta}+e^{-\gamma\theta}\right)(1-\theta)^2\,d\theta\right) \\ &= \frac{e^{\gamma i}}{\gamma^2}\left(\frac{e^{\gamma}-e^{-\gamma}}{\gamma}-2\right) \end{aligned} \tag{C.11}$$

Faisons le calcul pour $\bar{i}=i+1$,

$$\begin{aligned} I_{\ell,i,i+1} &= e^{\gamma i}\int_{-1}^{1} e^{\gamma\theta}\varphi(\theta)\,\varphi(\theta-1)\,d\theta \\ &= e^{\gamma i}\int_{0}^{1} e^{\gamma\theta}(1-\theta)\,\theta d\theta \\ &= \frac{e^{\gamma i}}{\gamma^2}\left(e^{\gamma}\left(1-\frac{2}{\gamma}\right)+1+\frac{2}{\gamma}\right). \end{aligned} \tag{C.12}$$

Puis, pour $\bar{i} = i - 1$,

$$\begin{aligned} I_{\ell,i,i-1} &= e^{\gamma i} \int_0^1 e^{-\gamma\theta} (1-\theta)\, \theta d\theta \\ &= \frac{e^{\gamma i}}{\gamma^2} \left(e^{-\gamma} \left(1 + \frac{2}{\gamma} \right) + 1 - \frac{2}{\gamma} \right). \end{aligned} \tag{C.13}$$

Au voisinage de $\gamma = 0$, un développement de la fonction exp permet de montrer que nous retrouvons le coefficient de la matrice de masse. Nous donnons le résultat pour $\bar{i} = i$ et $\bar{i} = i + 1$:

$$\begin{aligned} I_{\ell,i,i} &= \frac{e^{\gamma i}}{\gamma^2} \left(\frac{2\gamma^2}{3} + O\left(\gamma^3\right) \right) = e^{\gamma i} \frac{2}{3} + O(\gamma), \\ I_{\ell,i,i+1} &= e^{\gamma} \left(1 - \frac{2}{\gamma} \right) + 1 + \frac{2}{\gamma} \approx \frac{\gamma^2}{2} - \frac{2\gamma^2}{6} + O(\gamma^3). \end{aligned} \tag{C.14}$$

– Équation (C.5),

$$\begin{aligned} I &= \int \varphi'_{\ell,i}(x)\, \varphi_{\ell,\bar{i}}(x)\, dx \\ &= \int 2^{2\ell} \varphi' \left(2^\ell x - i \right) \varphi \left(2^\ell x - \bar{i} \right) dx \\ &= 2^\ell \int_{-1}^1 \varphi'(\theta)\, \varphi(\theta - (\bar{i} - i))\, d\theta \\ &= 2^{\ell-2} \left(-\delta(\bar{i} = i+1) + \delta(\bar{i} = i-1) \right). \end{aligned} \tag{C.15}$$

– Équation (C.6),

$$\begin{aligned} I &= \int x \varphi'_{\ell,i}(x)\, \varphi_{\ell,\bar{i}}(x)\, dx \\ &= \int x 2^{2\ell} \varphi' \left(2^\ell x - i \right) \varphi \left(2^\ell x - \bar{i} \right) dx \\ &= \int_{-1}^1 (\theta + i)\, \varphi'(\theta)\, \varphi(\theta - (\bar{i} - i))\, d\theta \\ &= \int_{-1}^1 \theta \varphi'(\theta)\, \varphi(\theta - (\bar{i} - i))\, d\theta \\ &\quad + i \int_{-1}^1 \varphi'(\theta)\, \varphi(\theta - (\bar{i} - i))\, d\theta \\ &= \left(\frac{2}{3} \delta(\bar{i} = i) - \frac{1}{3} \delta(\bar{i} = i+1) - \frac{1}{3} \delta(\bar{i} = i-1) \right) \\ &\quad + \frac{i}{2} \left(-\delta(\bar{i} = i+1) + \delta(\bar{i} = i-1) \right) \end{aligned} \tag{C.16}$$

– Équation (C.7),

$$\begin{aligned} I_{\ell,i,\bar{i}} &= \int e^{\alpha x} \varphi'_{\ell,i}(x)\, \varphi_{\ell,\bar{i}}(x)\, dx \\ &= 2^\ell \int_{-1}^1 e^{\alpha 2^{-\ell}(\theta+i)} \varphi'(\theta)\, \varphi(\theta - (\bar{i} - i))\, d\theta \\ &\quad \text{avec } \gamma = \alpha 2^{-\ell} \\ &= 2^\ell e^{\gamma i} \int_{-1}^1 e^{\gamma\theta} \varphi'(\theta)\, \varphi(\theta - (\bar{i} - i))\, d\theta \end{aligned} \tag{C.17}$$

$$\begin{aligned} I_{\ell,i,i} &= 2^{\ell} e^{\gamma i} \int_0^1 \left(e^{-\gamma\theta} - e^{\gamma\theta}\right)(1-\theta)\, d\theta \\ &= 2^{\ell} \frac{e^{\gamma i}}{\gamma}\left(2 + \frac{e^{-\gamma} - e^{\gamma}}{\gamma}\right) \end{aligned} \tag{C.18}$$

$$\begin{aligned} I_{\ell,i,i+1} &= 2^{\ell} e^{\gamma i} \int_0^1 -e^{\gamma\theta}\theta d\theta \\ &= -2^{\ell} \frac{e^{\gamma i}}{\gamma}\left(e^{\gamma}\left(1 - \frac{1}{\gamma}\right) + \frac{1}{\gamma}\right) \end{aligned} \tag{C.19}$$

$$\begin{aligned} I_{\ell,i,i-1} &= 2^{\ell} e^{\gamma i} \int_0^1 e^{-\gamma\theta}\theta d\theta \\ &= -2^{\ell} \frac{e^{\gamma i}}{\gamma}\left(e^{-\gamma}\left(1 + \frac{1}{\gamma}\right) - \frac{1}{\gamma}\right) \end{aligned} \tag{C.20}$$

Comme précédemment, un développement de l'exponentielle permet de retrouver le résultat.

$$e^{-\gamma}\left(1+\frac{1}{\gamma}\right) - \frac{1}{\gamma} = +\frac{1}{\gamma} - 1 + \frac{\gamma}{2} - \frac{\gamma^2}{6} - \frac{1}{\gamma} = -\frac{\gamma}{2} + \frac{\gamma^2}{3} \quad \text{et} \quad I_{\ell,i,i-1} \approx 2^{\ell-1} e^{\gamma i}\left(1 - \frac{\gamma}{3}\right). \tag{C.21}$$

– Équation (C.9),

$$\begin{aligned} I_{\ell,i,\bar{i}} &= \int x^2 \varphi'_{\ell,i}(x)\, \varphi'_{\ell,\bar{i}}(x)\, dx \\ &= \int_{-1}^{1} (\theta + i)^2 \varphi'(\theta)\, \varphi'(\theta - (\bar{i} - i))\, d\theta \\ &= \int_0^1 -\left((-\theta + i)^2 \varphi'(\theta + (\bar{i} - i)) - (\theta + i)^2 \varphi'(\theta - (\bar{i} - i))\right) d\theta \\ &= \delta(\bar{i} = i)\left(\frac{2}{3} + 2i^2\right) - \delta(\bar{i} = i - 1)\left(\frac{1}{3} - i + i^2\right) - \delta(\bar{i} = i + 1)\left(\frac{1}{3} + i + i^2\right) \end{aligned} \tag{C.22}$$

■

Bibliographie

[Ach08] Y. Achdou. An inverse problem for a parabolic variational inequality with an integro-differ ential operator. *SIAM journal of Control and optimization*, 2008.

[Ada75] R. A. Adams. *Sobolev Spaces*. Academic Press, 1975.

[AFT05] Yves Achdou, Bruno Franchi, and Nicoletta Tchou. A partial differential equation connected to option pricing with stochastic volatility : regularity results and discretization. *Math. Comp.*, 74(251) :1291–1322 (electronic), 2005.

[AP05] Y. Achdou and O. Pironneau. *Computational methods for option pricing*, volume 30 of *Frontiers in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2005.

[AT02] Y. Achdou and N. Tchou. Variational analysis for the Black and Scholes equation with stochastic volatility. *M2AN Math. Model. Numer. Anal.*, 36(3) :373–395, 2002.

[Bab60] K. I. Babenko. Approximation of periodic functions of many variables by trigonometric polynomials. *Soviet Math. Dokl.(Engl. Transl.)*, 1 :513–516, 1960. Translated from : Dokl. Akad. Nauk SSS, 132 :247–250, 1960.

[Bac95] Louis Bachelier. *Théorie de la spéculation*. Les Grands Classiques Gauthier-Villars. [Gauthier-Villars Great Classics]. Éditions Jacques Gabay, Sceaux, 1995. Théorie mathématique du jeu. [Mathematical theory of games], Reprint of the 1900 original.

[Bar94] Guy Barles. *Solutions de viscosité des équations de Hamilton-Jacobi*, volume 17 of *Mathématiques & Applications (Berlin) [Mathematics & Applications]*. Springer-Verlag, Paris, 1994.

[Bat00] David S. Bates. Post-'87 crash fears in the S&P 500 futures option market. *J. Econometrics*, 94(1-2) :181–238, 2000.

[BBM05] A. Bergam, C. Bernardi, and Z. Mghazli. A posteriori analysis of the finite element discretization of some parabolic equations. *Math. Comp.*, 74(251) :1117–1138 (electronic), 2005.

[BDD04] Peter Binev, Wolfgang Dahmen, and Ron DeVore. Adaptive finite element methods with convergence rates. *Numer. Math.*, 97(2) :219–268, 2004.

[BG04] Hans-Joachim Bungartz and Michael Griebel. Sparse grids. *Acta Numerica*, 13 :1–123, 2004.

[BL84] Alain Bensoussan and Jacques-Louis Lions. *Impulse control and quasivariational inequalities*. μ. Gauthier-Villars, Montrouge, 1984. Translated from the French by J. M. Cole.

[Bla04] J. Blackham. Sparse grid solution to the libor market. Master's thesis, Magdalen College University of Oxford, 2004.

[BM97] Christine Bernardi and Yvon Maday. Spectral methods. In *Handbook of numerical analysis, Vol. V*, Handb. Numer. Anal., V, pages 209–485. North-Holland, Amsterdam, 1997.

[BMR04] Christine Bernardi, Yvon Maday, and Francesca Rapetti. *Discrétisations variationnelles de problèmes aux limites elliptiques*, volume 45 of *Mathématiques & Applications (Berlin) [Mathematics & Applications]*. Springer-Verlag, Berlin, 2004.

[BN96] S. Bertoluzza and G. Naldi. A wavelet collocation method for the numerical solution of partial differential equations. *Appl. Comput. Harmon. Anal.*, 3(1) :1–9, 1996.

[BNS01] Ole E. Barndorff-Nielsen and Neil Shephard. Modelling by Lévy processes for financial econometrics. In *Lévy processes*, pages 283–318. Birkhäuser Boston, Boston, MA, 2001.

[BNS02] Ole E. Barndorff-Nielsen and Neil Shephard. Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 64(2) :253–280, 2002.

[BØSW04] Francesca Biagini, Bernt Øksendal, Agnès Sulem, and Naomi Wallner. An introduction to white-noise theory and Malliavin calculus for fractional Brownian motion. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.*, 460(2041) :347–372, 2004. Stochastic analysis with applications to mathematical finance.

[BS73a] F. Black and M. Scholes. The pricing of options and corporate liabilities. *J. Pol. Econ.*, 81 :637–659, 1973.

[BS73b] Fischer Black and M. Scholes. The pricing of options and corporate liabilities,. *Journal of Political Economy,*, 81 :637–654, 1973.

[Car71] Fernando Cardoso. The identity of weak and strong extensions of pseudo-differential operators. *Proc. Amer. Math. Soc.*, 29 :118–122, 1971.

[CDD96] A. Cohen, W. Dahmen, and R. Devore. Multiscale decompositions on bounded domains, September 09 1996.

[CDF92] A. Cohen, I. Daubechies, and J. Feauveau. Biorthogonal bases of compactly supported wavelets. *Comm. Pure Appl. Math*, 45 :485–560, 1992.

[CGH76] J. C. Cavendish, W. J. Gordon, and C. A. Hall. Ritz-Galerkin approximations in blending function spaces. 26(2) :155–178, 1976.

[CGMY03] Peter Carr, Hélyette Geman, Dilip B. Madan, and Marc Yor. Stochastic volatility for Lévy processes. *Math. Finance*, 13(3) :345–382, 2003.

[Cia78] P. G. Ciarlet. *The Finite Element Method for Elliptic Problems*. North-Holland, Amsterdam, 1978.

[Cia91] P. G. Ciarlet. Basic error estimates for elliptic problems. In *Handbook of Numerical Analysis, Vol. II*, pages 17–351. North-Holland, Amsterdam, 1991.

[CM98a] P. Carr and D. Madan. Option valuation using the fast Fourier transform. *The journal of computational finance*, 2 :61–73, 1998.

[CM98b] A. Cohen and R. Masson. Wavelet adaptive method for second order elliptic problems boundary conditions and domain decomposition, March 06 1998.

[CN00] Z. Chen and R.H. Nochetto. Residual type a posteriori error estimates for elliptic obstacle problems. *Numer. Math.*, 84(4) :527–548, 2000.

[Coh00] Albert Cohen. Wavelet methods in numerical analysis. In *Handbook of numerical analysis, Vol. VII*, Handb. Numer. Anal., VII, pages 417–711. North-Holland, Amsterdam, 2000.

[Coh03] A. Cohen. *Numerical analysis of wavelet methods*. Elsevier, Amsterdam, 2003.

[CT03] R. Cont and P. Tankov. *Financial modelling with jump processes*. Chapman and Hall, 2003.

[CV05] Rama Cont and Ekaterina Voltchkova. A finite difference scheme for option pricing in jump diffusion and exponential Lévy models. *SIAM J. Numer. Anal.*, 43(4) :1596–1626 (electronic), 2005.

[CW02] Peter Carr and Liuren Wu. Time-changed levy processes and option pricing. Finance 0207011, EconWPA, August 2002.

[CY89] C.Bernardi and Y.Maday. Approximation results for spectral methods with domain decomposition. *Appl. Numer. Math.*, 6(1/2), 1989. MSC2000.

[Dah97] W. Dahmen. Wavelet and multiscale methods for operator equations. *Acta Numerica*, 6 :55–228, 1997.

[Dau92] Ingrid Daubechies. *Ten Lectures on Wavelets*. SIAM, 1992.

[DD89] G. Deslauriers and S. Dubuc. Symmetric iterative interpolation processes. *Constructive Approximation*, 5 :49–68, 1989.

[DeV98] Ronald A. DeVore. Nonlinear approximation. In *Acta numerica, 1998*, volume 7 of *Acta Numer.*, pages 51–150. Cambridge Univ. Press, Cambridge, 1998.

[DM93] W. Dahmen and C. Micchelli. Using the refinement equation for evaluating integrals of wavelets. *SIAM J. NUMER. ANAL*, 30(2) :507–537, April 1993.

[Don92] David L. Donoho. Interpolating wavelet transforms. Technical report, 1992.

[DPS93] W. Dahmen, S. Prossdorf, and R. Schneider. Wavelet approximation methods for pseudodifferential equations II :Matrix compression and fast solution. *Adv. in Comp. Math.*, 1 :259–335, 1993.

[EGH00] Robert Eymard, Thierry Gallouët, and Raphaèle Herbin. Finite volume methods. In *Handbook of numerical analysis, Vol. VII*, Handb. Numer. Anal., VII, pages 713–1020. North-Holland, Amsterdam, 2000.

[EJ91] K. Eriksson and C. Johnson. Adaptive finite element methods for parabolic problems. I. A linear model problem. *SIAM J. Numer. Anal.*, 28(1) :43–77, 1991.

[EJ95] K. Eriksson and C. Johnson. Adaptive finite element methods for parabolic problems. II. Optimal error estimates in $L_\infty L_2$ and $L_\infty L_\infty$. *SIAM J. Numer. Anal.*, 32(3) :706–740, 1995.

[FPS00] Jean-Pierre Fouque, George Papanicolaou, and K. Ronnie Sircar. *Derivatives in financial markets with stochastic volatility*. Cambridge University Press, Cambridge, 2000.

[Fri44] K. O. Friedrichs. The identity of weak and strong extensions of differential operators. *Trans. Amer. Math. Soc.*, 55 :132–151, 1944.

[FS06] W. Farkas and C. Schwab. Anisotropic stable levy copula processes - analysis and numerical pricing methods. *SSRN eLibrary*, 2006.

[Gar05] J. Garcke. Sparse grid tutorial. http ://wwwmaths.anu.edu.au/ garcke/paper/sparseGridTutorial.pdf, 2005.

[Gau05] Walter Gautschi. Orthogonal polynomials (in Matlab). *J. Comput. Appl. Math.*, 178(1-2) :215–234, 2005.

[GG98] Thomas Gerstner and Michael Griebel. Numerical integration using sparse grids. *Numerical Algorithms*, 18(3-4) :209–232, 1998.

[GH08] T. Gerstner and M. Holtz. Valuation of performance-dependent options. *Applied Mathematical Finance*, 15(1) :1–20, 2008.

[GK00] M. Griebel and S. Knapek. Optimized tensor-product approximation spaces. *Constructive Approximation*, 16(4) :525–540, 2000.

[GO95] M. Griebel and P. Oswald. Tensor-product-type subspace splittings and multilevel iterative methods for anisotropic problems. *Advances of Computational Mathematics*, 4 :171–206, 1995.

[GOS99] M. Griebel, P. Oswald, and T. Schiekofer. Sparse grids for boundary integral equations. *Numer. Mathematik*, 83(2) :279–312, 1999. also as SFB 256 report 554, Universität Bonn.

[GR86] Vivette Girault and Pierre-Arnaud Raviart. *Finite element methods for Navier-Stokes equations*, volume 5 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 1986. Theory and algorithms.

[Gri98] M. Griebel. Adaptive sparse grid multilevel methods for elliptic PDEs based on finite differences. *Computing*, 61(2) :151–179, 1998.

[GSZ92] M. Griebel, M. Schneider, and C. Zenger. A combination technique for the solution of sparse grid problems. In P. de Groen and R. Beauwens, editors, *Iterative Methods in Linear Algebra*, pages 263–281. IMACS, Elsevier, North Holland, 1992. also as SFB Bericht, 342/19/90 A, Institut für Informatik, TU München, 1990.

[Hes93] S. Heston. A closed form solution for options with stochastic volatility with application to bond and currency options. *Review with Financial Studies*, pages 327–343, 1993.

[HIK02] M. Hintermüller, K. Ito, and K. Kunisch. The primal-dual active set strategy as a semismooth Newton method. *SIAM J. Optim.*, 13(3) :865–888 (electronic) (2003), 2002.

[HK94] R. H. W. Hoppe and R. Kornhuber. Adaptive multilevel methods for obstacle problems. *SIAM J. Numer. Anal.*, 31(2) :301–323, 1994.

[Hoh94] Walter Hoh. The martingale problem for a class of pseudo-differential operators. *Math. Ann.*, 300(1) :121–147, 1994.

[Hör67] L. Hörmander. Hypoelliptic second order differential equations. *Acta Mathematica*, 119 :147–171, 1967.

[Hör07] Lars Hörmander. *The analysis of linear partial differential operators. III.* Classics in Mathematics. Springer, Berlin, 2007. Pseudo-differential operators, Reprint of the 1994 edition.

[IK03] K. Ito and K. Kunisch. Semi-smooth Newton methods for variational inequalities of the first kind. *M2AN Math. Model. Numer. Anal.*, 37(1) :41–62, 2003.

[Jac98] Niels Jacob. Characteristic functions and symbols in the theory of Feller processes. *Potential Anal.*, 8(1) :61–68, 1998.

[Joh92] C. Johnson. Adaptive finite element methods for the obstacle problem. *Math. Models Methods Appl. Sci.*, 2(4) :483–487, 1992.

[Jou04] B. Jourdain. Loss of martingality in asset price models with lognormal stochastic volatility. Preprint 2004-267, CERMICS, 2004.

[Kal02] Olav Kallenberg. *Foundations of modern probability.* Probability and its Applications (New York). Springer-Verlag, New York, second edition, 2002.

[Kna00] S. Knapek. *Approximation und Kompression mit Tensorprodukt-Multiskalenräumen.* Doktorarbeit, Universität Bonn, April 2000.

[Kor57] N. M. Korobov. Approximate calculation of multiple integrals with the aid of methods in the theory of numbers. *Dokl. Akad. Nauk SSSR*, 115 :1062–1065, 1957.

[Kor96] R. Kornhuber. A posteriori error estimates for elliptic variational inequalities. *Comput. Math. Appl.*, 31(8) :49–60, 1996.

[Kos00] F. Koster. A proof of the consistency of the finite difference technique on sparse grids. *Computing*, 65 :247–261, 2000. also as Report SFB 256, No. 642, Univ. Bonn, 2000.

[Kro65] Aleksandr Semenovich Kronrod. *Nodes and weights of quadrature formulas. Sixteen-place tables.* Authorized translation from the Russian. Consultants Bureau, New York, 1965.

[KS02] Koster F. Knapek S. Integral operators on sparse grids. *SIAM J. Num. Anal.*, 2002.

[Lau01] Dirk P. Laurie. Computation of Gauss-type quadrature formulas. *J. Comput. Appl. Math.*, 127(1-2) :201–217, 2001. Numerical analysis 2000, Vol. V, Quadrature and orthogonal polynomials.

[LL97] D. Lamberton and B. Lapeyre. *Introduction au calcul stochastique appliqué à la finance.* Ellipses, 1997.

[LM61] J.L Lions and E. Magenes. Problemi ai limiti non omogenei(iii). *Ann. Scuola Norm. Pisa*, 15 :41–103, 1961.

[LM68] Jacques-Louis Lions and Enrico Magenes. *Problèmes aux limites non homogènes et applications*, volume I and II. Dunod, Paris, 1968.

[LP94] E. Lanconelli and S. Polidoro. On a class of hypoelliptic evolution operators. *Rend. Sem. Mat. Univ. Politec. Torino*, 52(1) :29–63, 1994. Partial differential equations, II (Turin, 1993).

[Mal98] Stéphane Mallat. *A Wavelet Tour of Signal Processing.* Academic Press, San Diego, 1998.

[Mas99] Roland Masson. *Méthodes d'Ondelettes en Simulation Numérique pour les Problèmes Elliptiques et de Point Selle*. PhD thesis, Université Paris 6, 1999.

[MDF05] A. Pascucci M. Di Francesco. On a class of degenerate parabolic equations of kolmogorov type. *AMRX Appl. Math. Res. Express*, 3 :77–116, 2005.

[Mer73] Robert C. Merton. Theory of rational option pricing. *Bell J. Econom. and Management Sci.*, 4 :141–183, 1973.

[Mer76] R.C. Merton. Option pricing when underlying stock returns are discontinuous. *J. Financ. Econ.*, 3 :125–144, 1976.

[Mey90a] Y. Meyer. *Ondelettes et Opérateurs I.* Hermann, 1990.

[Mey90b] Y. Meyer. *Ondelettes et Opérateurs II. Opérateurs de Caldéron-Zygmund*. Hermann, 1990.

[MPS04] A-M. Matache, T.Von Petersdorff, and C. Schwab. Fast deterministic pricing of options on lévy driven assets. *Mathematical Modelling and Numerical Analysis*, 38 :37–72, 2004.

[MSW06] A.-M. Matache, C. Schwab, and T. P. Wihler. Linear complexity solution of parabolic integro-differential equations. *Numer. Math.*, 104(1) :69–102, 2006.

[MvPS04] A.M. Matache, T. von Petersdoff, and C. Schwab. Fast deterministic pricing of Lévy driven assets. *Mathematical Modelling and Numerical Analysis*, 38(1) :37–72, 2004.

[Nec04] Ciprian Necula. Option pricing in a fractional Brownian motion environment. *Math. Rep. (Bucur.)*, 6(56)(3) :259–273, 2004.

[NSV03] R.H. Nochetto, K.G. Siebert, and A. Veeser. Pointwise a posteriori error control for elliptic obstacle problems. *Numer. Math.*, 95(1) :163–195, 2003.

[NSV05] R.H. Nochetto, K.G. Siebert, and A. Veeser. Fully localized a posteriori error estimators and barrier sets for contact problems. *SIAM J. Numer. Anal.*, 42(5) :2118–2135 (electronic), 2005.

[NT06] David Nualart and Murad S. Taqqu. Wick-Itô formula for Gaussian processes. *Stoch. Anal. Appl.*, 24(3) :599–614, 2006.

[ØS07] Bernt Øksendal and Agnès Sulem. *Applied stochastic control of jump diffusions*. Universitext. Springer, Berlin, second edition, 2007.

[Pan02] Jun Pan. The jump-risk premia implicit in options : evidence from an integrated time-series study. *Journal of Financial Economics*, 63(1) :3–50, January 2002.

[Pat68] T. N. L. Patterson. On some Gauss and Lobatto based integration formulae. *Math. Comp. 22 (1968), 877–881; addendum, ibid.*, 22(104, loose microfiche suppl.) :D1–D4, 1968.

[Pat89] T. N. L. Patterson. Algorithm 672 : generation of interpolatory quadrature rules of the highest degree of precision with preassigned nodes for general weight functions. *ACM Trans. Math. Software*, 15(2) :137–143, 1989.

[PH] O. Pironneau and F. Hecht. *FREEFEM*. www.ann.jussieu.fr.

[Pro04a] Philip E. Protter. *Stochastic integration and differential equations*, volume 21 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin, second edition, 2004. Stochastic Modelling and Applied Probability.

[Pro04b] Philip E. Protter. *Stochastic integration and differential equations*, volume 21 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin, second edition, 2004. Stochastic Modelling and Applied Probability.

[PS96] Tobias-Von Petersdorff and Christoph Schwab. Wavelet approximations for first kind boundary integral equations on polygons. *Numerische Mathematik*, 74(4) :479–516, 1996.

[PS04] Tobias-Von Petersdorff and Christoph Schwab. Numerical solution of parabolic equations in high dimensions. *M2AN Math. Model. Numer. Anal.*, 38(1) :93–127, 2004.

[PW01] L. Plaskota and G. W. Wasilkowski. The exact exponent of sparse grid quadratures in the weighted case. *Journal of Complesity*, 17 :840–849, 2001.

[PWW00] L. Plaskota, G. W. Wasilkowski, and H. Woźniakowski. A New Algorithm and Worst Case Complexity for Feynman–Kac Path Integration. *Journal of Computational Physics*, 164 :335–353, 2000.

[Qua91] A. Quarteroni. An introduction to spectral methods for partial differential equations. In *Advances in numerical analysis, Vol. I (Lancaster, 1990)*, Oxford Sci. Publ., pages 96–146. Oxford Univ. Press, New York, 1991.

[Rei04] C. Reisinger. *Numerische Methoden für hochdimensionale parabolische Gleichungen am Beispiel von Optionspreisaufgaben*. PhD thesis, Universitat Heidelberg, 2004.

[Rei08] Nils Reich. Wavelet compression of integral operators on sparse tensor spaces.construction, consistency and asymptotically optimal complexity. Research report 2008-24, ETH-SAM, Zurich, 2008.

[RM94] Robert D. Richtmyer and K. W. Morton. *Difference methods for initial-value problems*. Robert E. Krieger Publishing Co. Inc., Malabar, FL, second edition, 1994.

[RW07] Christoph Reisinger and Gabriel Wittum. Efficient hierarchical approximation of high-dimensional option pricing problems. *SIAM J. Sci. Comput.*, 29(1) :440–458 (electronic), 2007.

[Saa96] Y. Saad. *Iterative Methods for Sparse Linear Systems*. PWS Publishing Company, 1996.

[Sat99] Ken-iti Sato. *Lévy processes and infinitely divisible distributions*, volume 68 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 1999. Translated from the 1990 Japanese original, Revised by the author.

[Sch98a] T. Schiekofer. *Die Methode der Finiten Differenzen auf dünnen Gittern zur Lösung elliptischer und parabolischer partieller Differentialgleichungen*. PhD thesis, Universität Bonn, 1998.

[Sch98b] R. Schneider. *Multiskalen-und Wavelet-Matrixkompression : analysisbasierte Methoden zur effizienten Losung grosser vollbesetzter Gleichungssysteme*. Teubner Verlag, 1998.

[Smo63] S. A. Smolyak. Interpolation and Quadrature Formulas for the classes W_s^a and E_s^a. *Soviet mathematic*, 1 :384–387, 1963.

[SS01] Dominik Schötzau and Christoph Schwab. *hp*-discontinuous Galerkin time-stepping for parabolic problems. *C. R. Acad. Sci. Paris Sér. I Math.*, 333(12) :1121–1126, 2001.

[SS08] Christoph Schwab and Rob Stevenson. Adaptive wavelet algorithms for elliptic PDE's on product domains. *Math. Comp.*, 77(261) :71–92 (electronic), 2008.

[SST07] C. Schwab, E. Süli, and R.A. Todor. Sparse finite element approximation of high-dimensional transport-dominated diffusion problems. Technical report, ETH Zurich - D-MATH - SAM, 2007.

[SZ07] M. O. Souza and J. P. Zubelli. On the asymptotics of fast mean-reversion stochastic volatility models. *International Journal of Theoretical and Applied Finance*, 2007.

[Tod03] R.-A. Todor. A new approach to energy-based sparse fe spaces, 2003.

[Tri92] Hans Triebel. *Theory of function spaces. II*, volume 84 of *Monographs in Mathematics*. Birkhäuser Verlag, Basel, 1992.

[Tri06] Hans Triebel. *Theory of function spaces. III*, volume 100 of *Monographs in Mathematics*. Birkhäuser Verlag, Basel, 2006.

[Vee01] A. Veeser. Efficient and reliable a posteriori error estimators for elliptic obstacle problems. *SIAM J. Numer. Anal.*, 39(1) :146–167 (electronic), 2001.

[Ver96] R. Verfurth. *A review of a posteriori error estimation and adaptive mesh-refinement techniques*. Wiley Chichester, 1996.

[Vol05] Ekaterina Voltchkova. *Integro-differential evolution equations : numerical methods and applications in finance*. PhD thesis, CMAP, EP - CMAP Centre de Mathématiques Appliquées, EP/X, 2005.

[WGSS01] T. Werder, K. Gerdes, D. Schötzau, and C. Schwab. *hp*-discontinuous Galerkin time stepping for parabolic problems. *Comput. Methods Appl. Mech. Engrg.*, 190(49-50) :6685–6708, 2001.

[WW95] Grzegorz W. Wasilkowski and Henryk Woźniakowsk. Explicit Cost Bounds of Algorithms for Multivariate Tensor Product Problems. *Journal of Complexity*, 11(1) :1–56, 1995.

[Zen91] C. Zenger. Sparse grids. In W. Hackbusch, editor, *Parallel Algorithms for Partial Differential Equations*, Notes on Numerical Fluid Mechanics ; 31, pages 241–251. Vieweg, Braunschweig, 1991.

[ZT00] O. C. Zienkiewicz and R. L. Taylor. *The finite element method. Vol. 1.* Butterworth-Heinemann, Oxford, fifth edition, 2000. The basis.

[Zum00] G. W. Zumbusch. A sparse grid PDE solver. In H. P. Langtangen, A. M. Bruaset, and E. Quak, editors, *Advances in Software Tools for Scientific Computing*, volume 10 of *Lecture Notes in Computational Science and Engineering*, chapter 4, pages 133–177. Springer, Berlin, Germany, 2000. (Proceedings SciTools '98).

[Zum03] G. Zumbusch. *Parallel Multilevel Methods. Adaptive Mesh Refinement and Loadbalancing*. Teubner, 2003.

Résumé

Cette thèse regroupe plusieurs travaux relatifs à la résolution numérique d'équations aux dérivées partielles et d'équations intégro-différentielles issues de la modélisation stochastique de produits financiers.

La première partie des travaux est consacrée aux méthodes de Sparse Grid appliquées à la résolution numérique d'équations en dimension supérieure à trois. Deux types de problèmes sont abordés. Le premier concerne l'évaluation d'options vanilles dans un modèle à sauts avec une volatilité stochastique multi-facteurs. La résolution numérique de l'équation de valorisation, posée en dimension 4, est obtenue à l'aide d'une méthode de différences finies sparse et d'une méthode de collocation pour la discrétisation de l'opérateur intégral. Le second problème traite de l'évaluation de produits sur un panier de plusieurs sous-jacents. Il nécessite le recours à une méthode de Galerkin sur une base d'ondelettes obtenue à l'aide d'un produit tensoriel sparse.

La seconde partie des travaux concerne des estimations d'erreur a posteriori pour des options américaines sur un panier de plusieurs actifs.

mots clés : options sur panier, options sur des processus a sauts avec volatilité stochastique, équations aux dérivées partielles en grande dimension, méthodes numériques sparse, estimations d'erreur a posteriori, options américaines.

Abstract

In this work, we present some numerical methods to approximate Partial Differential Equations (PDEs) or Partial Integro-Differential Equations (PIDEs) commonly arising in finance.
This thesis is split into three part. The first one deals with the study of Sparse Grid techniques. In an introductory chapter, we present the construction of Sparse Grid spaces and give some approximation properties. The second chapter is devoted to the presentation of a numerical algorithm to solve PDEs on these spaces. This chapter gives us the opportunity to clarify the finite difference method on Sparse Grid by looking at it as a collocation method. We make a few remarks on the practical implementation.
The second part of the thesis is devoted to the application of Sparse Grid techniques to mathematical finance. We will consider two practical problems. In the first one, we consider a European vanilla contract with a multivariate generalisation of the one dimensional Ornstein-Ulenbeck-based stochastic volatility model. A relevant generalisation is to assume that the underlying asset is driven by a jump process, which leads to a PIDE. Due to the curse of dimensionality, standard deterministic methods are not competitive with Monte Carlo methods. We discuss sparse grid finite difference methods for solving the PIDE arising in this model up to dimension 4. In the second problem, we consider a Basket option on several assets (five in our example) in the Black & Scholes model. We discuss Galerkin methods in a sparse tensor product space constructed with wavelets.
The last part of the thesis is concerned with a posteriori error estimates in the energy norm for the numerical solutions of parabolic obstacle problems allowing space/time mesh adaptive refinement. These estimates are based on a posteriori error indicators which can be computed from the solution of the discrete problem. We present the indicators for the variational inequality obtained in the context of the pricing of an American option on a two dimensional basket using the Black & Scholes model. All these techniques are illustrated by numerical examples.

key words : basket options, Stochastic volatility model with jump, parabolic equations in high dimensions, Sparse Grid techniques, posteriori error estimates, American options.