Méthodes numériques et modélisation pour certains problèmes multi-échelles

Mémoire

Présenté en vue d'obtenir

L'HABILITATION A DIRIGER DES RECHERCHES

par

Alexei LOZINSKI

Université Paul Sabatier, Toulouse 3

Spécialité :

Mathématiques Appliquées

Soutenance prévue le 8 décembre 2010
devant le jury composé de

M. Yves Achdou	Professeur	Université Paris 7
M. Claude Le Bris	Professeur	ENPC
M. Martin Gander	Professeur	Université de Genève
M. Patrick Laborde	Professeur	Université Paul Sabatier
M. Olivier Pironneau	Professeur	Université Paris 6
M. Jean-Paul Vila	Professeur	INSA de Toulouse

Table des matières

Table des matières i

Avant-propos iii

I Sur la résolution numérique de problèmes elliptiques multi-échelles 1

1 Présentation de l'approche dite "Zoom Numérique" 3

2 La méthode des patchs d'éléments finis 9
2.1 Les estimations d'erreur *a priori* pour la méthode des patchs 9
2.2 La vitesse de convergence de la méthode itérative 13
2.3 Une accélération : patchs harmoniques 17

3 Sur la méthode Chimère 20
3.1 Convergence de l'algorithme de Schwarz sur des maillages arbitraires non-conformes 20
3.2 Une extension au cas des problèmes non stationnaires 22

4 Une comparaison numérique des algorithmes précédents 25

5 "Zoom Numérique" pour "multi-modèles" 30

6 MsFEM 37

7 Perspectives 46

II Sur la modélisation des fluides viscoélastiques 47

8 Construction du modèle de chaînes bille-ressort 49

8.1 Le modèle prenant en compte l'inertie des billes 50
8.2 La limite sans inertie des billes 53
8.3 Le cas des écoulements localement homogènes 56

9 Sur le choix de la force de ressort 58
9.1 Motivation physiques 58
9.2 Des résultats théoriques sur le modèle avec réflexion 60
9.3 Simulations numériques 65

10 Perspectives 69

III Autres travaux 71

11 Méthodes numériques pour les écoulements biphasiques 73
11.1 Sur la discrétisation en temps en simulation des écoulements avec des particules . . . 73
11.2 Une méthode du type domaines fictifs pour des écoulements avec des bulles de gaz . 74

12 Problèmes fortement anisotropes 75
12.1 Un schéma du type "Asymptotic-Preserving" pour une équation de diffusion anisotrope 75
12.2 Méthodes adaptatives sur des maillages anisotropes pour les problèmes non stationnaires 76

13 Méthodes numériques non probabilistes pour le modèle des chaînes 77

14 Travaux pendant la thèse et avant 79

Bibliographie 81

Table des matières i

IV Articles annexés 87

Numerical Zoom for Advection Diffusion Problems with Localized Multiscales 91

On discretization in time in simulations of particulate 105

Motion of gas bubbles, considered as massless bodies, affording deformations within a prescribed family of shapes, in an incompressible fluid under the action of gravitation and surface tension 129

Duality-based Asymptotic-Preserving method for highly anisotropic diffusion equations 165

An anisotropic error estimator for the Crank-Nicolson method : application to a parabolic problem 195

Sparse tensor-product Fokker-Planck-based methods for nonlinear bead-spring chain models of dilute polymer solutions 225

Avant-propos

On présent dans ce travail la problématique "multi-échelles" de deux points de vue :

– Les problèmes où l'aspect "multi-échelles" apparaît au niveau de fortes variations dans les données ou les coefficients des équations. On proposera des méthodes numériques adaptées à ces situations : patchs d'éléments finis ou le Zoom Numérique, Chimère, les méthodes de type "éléments finis multi-échelles - MsFEM".

– Les problèmes où l'aspect "multi-échelles" est pris en compte déjà au niveau des équations du modèle, comme par exemple dans la théorie cinétique des polymères. On présentera des travaux sur la dérivation des modèle des fluides polymériques qui permettent le couplage entre les échelles microscopique (les molécules du polymère) et les échelle macroscopiques (écoulement du fluide).

Les problème du premier type évoqués ci-dessus sont très répandus dans les calculs industriels actuels. On peut citer l'exemple des structures aéronautiques, où le calcul de structure de grande taille (par exemple un avion) doit prendre en compte le comportement des nombreux rivets ou des petites fissures. Un autre exemple de motivation se trouve dans des problèmes du contrôle de la qualité d'air, dans lesquels les sources d'émission des polluants sont de taille beaucoup plus petite que la taille de la région affectée par la pollution. Pour faire une simulation précise sur une grande échelle, on doit calculer une solution numérique dans un sous-domaine autour de la source de pollution en utilisant un maillage beaucoup plus fin.

Afin de traiter ce type de difficultés, je présenterai une méthode de résolution numérique pour des problèmes elliptiques avec des données multi-échelles qui utilise plusieurs niveaux des maillages, pas nécessairement emboîtés (patchs d'éléments finis). J'ai participé au développement de ces méthodes qui a été mené conjointement par les équipes à l'EPFL (Jacques Rappaz) et à l'Université de Houston (Roland Glowinski et Jiwen He). On résout le problème dans un domaine Ω à l'intérieur duquel on considère des patchs $\Lambda_1,\ \Lambda_2,\ \ldots$ où une précision plus élevée est désirable. On calcule alors successivement des corrections à la solution globale dans les patchs. Nous avons démontré des estimations d'erreur a priori et étudié la vitesse de convergence de la méthode itérative. Ces résultats théoriques sont confirmés par des expériences numériques.

Plus récemment, en collaboration avec Olivier Pironneau (Paris 6) et Patrick Laborde (IMT, Toulouse 3), on a développé des nouvelles variantes de ces méthodes qui sont plus faciles à mettre en

œuvre. Elles peuvent être appliquées dans les situations où les multi-échelles sont incrustées dans les équations aux dérivées partielles du problème (et pas seulement dans les données). Cela ouvre la route au couplage de plusieurs modèles dans une seule simulation. Par exemple, en mécanique des structures, on peut superposer un calcul fin local autour d'une petite fissure sur le calcul global relativement grossier de toute la structure. Nous avons aussi effectué des comparaisons numériques avec la méthode de Schwarz traditionnelle et une analyse théorique de cette dernière.

J'ai aussi participé dans le groupe de travail mené par Pierre Degond (IMT, Toulouse 3) sur le développement des méthode du type MsFEM (Multi-scale Finite Element Method). A la différence des méthodes du type "Zoom numérique" évoquées plus haut, la MsFEM est destinée aux problèmes où le comportement multi-échelles n'est pas localisé dans une petite région, mais il est présent partout dans le domaine du calcul. Le but est de pouvoir faire des calcul assez précis sur un maillage de pas plus grand que la longueur caractéristique de la micro-échelle en se permettant d'effectuer des calculs fins et coûteux sur chaque cellule du maillage grossier une fois pour toutes.

Tous ces travaux sont présentés dans la première partie de cette thèse d'habilitation.

La deuxième partie du manuscrit est consacrée aux problèmes où l'aspect "multi-échelles" est pris en compte déjà au niveau des équations du modèle. On s'intéresse notamment à quelques questions autour de la modélisation des polymères au niveau microscopique et autour du couplage avec le niveau macroscopique de l'écoulement d'une solution polymérique. On utilise une approche de la théorie cinétique où les molécules du polymère sont représentées par des "dumbbells" – deux billes reliés par un ressort. On va présenter une modélisation complète qui prend en compte les effets d'inertie des billes et en déduire des modèles classiques sans inertie comme cas limite, à l'aide de techniques de développements asymptotiques. La deuxième question traitée dans cette partie portera sur un nouveau choix de la loi de force des ressorts dans le modèle des dumbbells, qui prend en compte l'extensibilité maximale d'un ressort et qui mène aux équations différentielles stochastiques réfléchies. On établira rigoureusement la formule pour le tenseur des contraintes dans ce modèle.

La troisième partie est un résumé de mes autres travaux de recherche. Certains articles décrits dans cette partie sont annexés à la fin du manuscrit.

Première partie

Sur la résolution numérique de problèmes elliptiques multi-échelles

1 Présentation de l'approche dite "Zoom Numérique"

Dans les simulations éléments finis qui exigent une solution très fine dans une zone localisée, les ingénieurs font assez souvent un calcul grossier global d'abord, suivi par un calcul beaucoup plus fin sur un sous-ensemble (zoom) Λ du domaine entier Ω. Evidemment, un calcul sur le zoom ne suffit pas en général, sauf si le problème est non-linéaire ou non-stationnaire et les itérations dans le zoom sont une partie de la boucle non-linéaire ou temporelle. On alterne donc les calculs grossiers et fins dans une boucle. Les questions suivantes se posent alors : comment formuler le problème dans Λ et comment utiliser la solution correspondante dans le solveur grossier pour corriger la solution globale. On présente dans ce chapitre plusieurs implémentations possibles. Dans les chapitres suivants, on étudiera la convergence et les erreurs relatives à ces approches.

1. La méthode Chimère

Dans la méthode Chimère, proposée originellement par Steger [48] pour des problèmes non stationnaires, on creuse un trou D dans le domaine grossier strictement à l'intérieur de la région du zoom Λ. Par exemple, le calcul de la pression hydrostatique u d'un écoulement dans un milieu poreux est gouvernée par la loi de Darcy avec la porosité K,

$$\begin{aligned} -\nabla \cdot (K \nabla u) &= f \quad \text{dans} \quad \Omega \\ u &= g \quad \text{sur} \quad \Gamma = \partial\Omega, \end{aligned} \tag{1.1}$$

où g est la valeur donnée sur la frontière. On choisit D strictement à l'intérieur de Λ et on construit deux maillages : une triangulation grossière $\mathcal{T}_H$ sur $\Omega_H = \Omega \setminus \overline{D}$ et une triangulation fine $\mathcal{T}_h$ sur Λ, voir Fig. 1.1. On appelle S_h la frontière de Λ et S_H la partie de la frontière de Ω_H autre que Γ. On introduit aussi des espaces d'éléments finis V_H et V_h sur $\mathcal{T}_H$ et $\mathcal{T}_h$ respectivement. On note $V_{0H} = V_H \cap H_0^1(\Omega_H)$, $V_{gH} = \{v_H \in V_H$ telles que $v_H|_\Gamma$ est l'interpolé nodal de g sur $\Gamma\}$, et finalement $V_{0h} = V_h \cap H_0^1(\Lambda)$. L'algorithme Chimère consiste à alterner en boucle les calculs grossier sur Ω_H et fin sur Λ en prenant pour conditions aux limites sur S_H et S_h les résultats de derniers calculs disponibles dans l'autre sous-domaine. Pour passer l'information entre $\mathcal{T}_H$ et $\mathcal{T}_h$ on doit utiliser l'interpolation

Entrées : Une fonction $u_h^0 \in V_{h0}$.
pour $n = 1...N$ **faire**
Trouver $u_H^n \in V_{gH}$ tel que

$$\begin{aligned} \int_{\Omega_H} K\nabla u_H^n \cdot \nabla v_H &= \int_{\Omega_H} f v_H, \ \forall v_H \in V_{0H}, \\ u_H^n|_{S_H} &= r_H u_h^{n-1}|_{S_H}. \end{aligned} \tag{1.2}$$

Trouver $w_h^n \in V_{0h}$ tel que

$$\begin{aligned} \int_{\Lambda} K\nabla u_h^n \cdot \nabla v_h &= \int_{\Lambda} f v_h, \ \forall v_h \in V_{0h}, \\ u_h^n|_{S_h} &= r_h u_H^n|_{S_h}. \end{aligned} \tag{1.3}$$

fin

Algorithme 1.1: Chimère

d'un maillage à l'autre et on introduit donc r_H (respectivement r_h) les opérateurs d'interpolation nodale sur V_H (respectivement V_h).

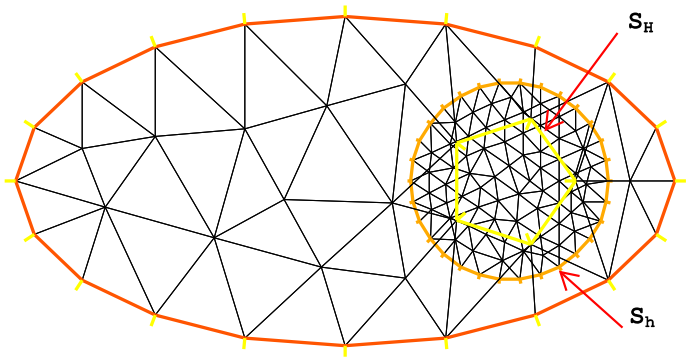


FIGURE 1.1: Triangulations $\mathcal{T}_H$ sur Ω_H (à l'extérieur de S_H) et $\mathcal{T}_h$ sur Λ (à l'intérieur de S_h).

Chimère est ainsi identique à l'algorithme alternatif de Schwarz, mais cette terminologie est utilisée dans la communauté de "Computational Fluid Dynamics". Les deux triangulations $\mathcal{T}_H$ et $\mathcal{T}_h$ ne sont pas nécessairement choisies de manière conforme. Par exemple, la frontière du trou ∂D peut couper arbitrairement les triangles de $\mathcal{T}_h$. Quoique cela fonctionne bien en général, ni la convergence de cet algorithme, ni la convergence de la solution approchée vers la solution exacte sous le raffinement des maillages ne sont toujours pas démontrées en général. La difficulté réside dans l'analyse de l'erreur d'interpolation de $\mathcal{T}_H$ sur $\mathcal{T}_h$ lorsque Λ n'est pas composé des éléments de la triangulation de Ω. Cependant, on présentera dans Chapitre 3 des éléments d'une analyse théorique de cette méthode

Entrées : une fonction $u_h^0 \in V_{0h}$.
pour $n = 1...N$ **faire**
Trouver $u_H^n \in V_{gH}$ tel que

$$\int_\Omega K\nabla u_H^n \cdot \nabla v_H = \int_\Omega f v_H - \int_\Omega K\nabla u_h^{n-1} \cdot \nabla v_H, \ \ \forall v_H \in V_{0H}. \tag{1.5}$$

Trouver $u_h^n \in V_{0h}$ tel que

$$\int_\Lambda K\nabla u_h^n \cdot \nabla v_h = \int_\Lambda f v_h - \int_\Lambda K\nabla u_H^n \cdot \nabla v_h, \ \ \forall v_h \in V_{0h}. \tag{1.6}$$

fin

Algorithme 1.2: Méthode des patchs d'éléments finis

en suivant [32]. Cette analyse est effectuée dans la norme de maximum et repose sur un principe de maximum fort au niveau discret.

2. La méthode des patchs d'éléments finis

Mis à part le faible taux de convergence de la méthode Chimère, souvent observé en pratique, l'inconvénient de cette méthode est la nécessité de changer la géométrie du problème grossier par rapport au problème original non raffiné. En effet, Ω_H dans (1.2) n'est pas le domaine Ω tout entier. Une idée alternative, introduite dans [53] et étudiée dans [39], [35], [16], consiste – au niveau continu – à garder le problème global tel quel (on notera ici sa solution par U) et à chercher une correction u à la solution grossier U dans l'espace des fonctions qui s'annulent en dehors de Λ, c'est-à-dire à trouver $U \in H^1(\Omega)$, $u \in H_0^1(\Lambda)$ avec $U|_\Gamma = g$ qui satisfont

$$\int_\Omega (K\nabla(U+u)\cdot\nabla(V+v) - f(V+v)) = 0, \ \ \forall V \in H_0^1(\Omega), v \in H_0^1(\Lambda) \tag{1.4}$$

Cette équation est facile à discrétiser, en prenant $u_H \approx U, \ u_h \approx u$, et en introduisant un schéma itératif de l'Algorithme 1.2. On utilise ici les mêmes notations qu'avant : V_H et V_h sont des espaces d'éléments finis sur des triangulations régulières $\mathcal{T}_H$ et $\mathcal{T}_h$ de Ω et Λ respectivement ; $V_{0H} = V_H \cap H_0^1(\Omega)$; V_{gH} est le sous-ensemble de V_H contenant les fonctions qui interpolent g sur Γ, et finalement $V_{0h} = V_h \cap H_0^1(\Lambda)$. Les approximations successives de la solution exacte u sont obtenues dans l'algorithme 1.2 par $u_{Hh}^n = u_H^n + u_h^n$. On montrera au chapitre 2, en suivant [16], l'estimation d'erreur pour la solution $u_{Hh} = \lim_{n\to\infty}(u_H^n + u_h^n)$ obtenue par l'algorithme :

$$\|u - u_{Hh}\|_{H^1(\Omega)} \leq C\left(H^r ||u||_{H^q(\Omega\setminus\overline{\Lambda})} + h^s ||u||_{H^q(\Lambda)}\right), \tag{1.7}$$

où r et s sont degrés maximaux des polynômes utilisés dans la construction de V_{0H} et V_{0h} respectivement et $q = \max(r,s)+1$.

Dans la suite on distinguera souvent les deux situations : des maillages $\mathcal{T}_h$ et $\mathcal{T}_H$ emboîtés, c'est-à-dire chaque triangle de $\mathcal{T}_h$ est contenu dans un seul triangle de $\mathcal{T}_H$, et des maillages $\mathcal{T}_h$ et $\mathcal{T}_H$ non emboîtés. Noter que la majoration (1.7) a lieu même si les deux maillages ne sont pas emboîtés l'un dans l'autre.

Entrées : une fonction $u_h^0 \in V_{0h}$.
pour $n = 1...N$ **faire**

Trouver $\lambda_H^n \in V_H^0$ tel que

$$\int_\Omega K\nabla\lambda_H^n \cdot \nabla\mu_H = \int_\Omega f\mu_H - \int_\Omega K\nabla u_h^{n-1} \cdot \nabla\mu_H, \ \ \forall\mu_H \in V_H^0. \tag{1.8}$$

Trouver $u_H^n \in V_{gH}$ tel que

$$\int_\Omega K\nabla u_H^n \cdot \nabla v_H = \int_\Omega f v_H - \int_\Omega K\nabla u_h^{n-1} \cdot \nabla v_H - \int_\Omega K\nabla\lambda_H^n \cdot \nabla v_H, \quad \forall v_H \in V_{0H}. \tag{1.9}$$

Trouver $u_h^n \in V_{0h}$ tel que

$$\int_\Lambda K\nabla u_h^n \cdot \nabla v_h = \int_\Lambda f v_h - \int_\Lambda K\nabla u_H^n \cdot \nabla v_h, \ \ \forall v_h \in V_{0h}. \tag{1.10}$$

fin

Algorithme 1.3: La méthode des patch harmoniques

3. La méthode des patchs harmoniques

L'inconvénient de la méthode des patchs consiste dans le fait que son taux de convergence peut être très bas pour certains paires de triangulations $\mathcal{T}_H$, $\mathcal{T}_h$ non emboîtées. Heureusement, la méthode peut être accélérée au prix de la résolution d'un problème auxiliaire typiquement peu coûteux à chaque itération. Cet algorithme accéléré a été introduit dans [18]. Il est résumé ci-après sous le nom de la méthode des patchs harmoniques, l'algorithme 1.3. On y utilise la notation suivante

$$V_H^0 = \{v_H \in V_{0H} : \text{supp } v_H \subset \Lambda\}.$$

V_H^0 est donc le sous-espace de V_{0H} qui contient les fonctions dont le support est à l'intérieur de Λ. Comme la taille de Λ n'est pas beaucoup plus grande que le pas du maillage grossier, la dimension de V_H^0 n'est pas trop élevée. Le problème pour λ_H^n est donc pas coûteux et la matrice correspondante est juste une sous-matrice de celle du problème grossier pour u_H^n. La nouvelle variable λ_H^n est purement auxiliaire. La solution approchée du problème global est toujours obtenue par $u_{Hh} = \lim_{n\to\infty}(u_H^n + u_h^n)$ comme dans le cas de l'algorithme 1.2. Le problème auxiliaire pour λ_H^n assure que les parties grossières u_H^n sont toujours a-orthogonales à V_H^0, c'est-à-dire elles sont approximativement a-harmoniques dans le cas $K = 1$, d'où le nom de l'algorithme.

L'idée des patchs harmoniques n'apporte rien de nouveau par rapport à l'algorithme 1.2 lorsque le maillage $\mathcal{T}_h$ est emboîté dans $\mathcal{T}_H$. En effet, l'algorithme 1.2 converge normalement très vite dans ce cas. De plus, ces deux algorithmes 1.2 et 1.3 sont identiques (toujours dans le cas des maillages emboîtés) en ce sens que les sommes $u_{Hh}^n = u_H^n + u_h^n$ sont exactement les mêmes dans les algorithmes 1.2 et 1.3 sur toutes les itérations $n \geq 1$, bien que u_H^n et u_h^n puissent être différents d'un algorithme à l'autre. En effet, l'intersection $V^0 = V_{0H} \cap V_{0h}$ coïncide dans ce cas avec V_H^0 et on peut décomposer $u_H^n = u_H^{n,0} + u_H^{n,\perp}$ et $u_h^n = u_h^{n,0} + u_h^{n,\perp}$ avec $u_H^{n,0},\ u_h^{n,0} \in V^0$ et $u_H^{n,\perp}$, $u_h^{n,\perp}$ qui appartiennent aux compléments orthogonaux de V^0 dans V_{0H} et V_{0h} respectivement. L'orthogonalité est comprise ici

Entrées : fonctions $u_H^0 \in V_{gH}$ et $w_h^0 \in V_h$.
pour $n = 1...N$ **faire**
Trouver $u_H^n \in V_{gH}$ tel que

$$\int_\Omega K\nabla u_H^n \cdot \nabla v_H = \int_\Omega f w_H + \int_\Lambda K\nabla(u_H^{n-1} - w_h^{n-1}) \cdot \nabla v_H, \ \ \forall v_H \in V_{0H}. \qquad (1.11)$$

Trouver $w_h^n \in V_h$ tel que

$$\int_\Lambda K\nabla w_h^n \cdot \nabla v_h = \int_\Lambda f v_h, \ \ \forall v_h \in V_{0h}, \qquad (1.12)$$
$$w_h^n|_{\partial\Lambda} = r_h u_H^n|_{\partial\Lambda}.$$

fin

Algorithme 1.4: Semi-Schwarz

au sens du produit scalaire $a(u,v) = \int_\Omega K\nabla u \cdot \nabla v$. En notant par $u^0 \in V^0$ l'unique solution de

$$a(u^0, v^0) = \int_\Omega f v^0, \quad \forall v^0 \in V^0$$

et en choisissant les fonctions test v_H et v_h dans V^0, on voit tout de suite que $u_H^{n,0} = u^0 - u_H^{0,0}$, $u_h^{n,0} = u_h^{0,0}$ à toutes les itérations de l'algorithme 1.2 et que $u_H^{n,0} = 0$, $u_h^{n,0} = u^0$ à toutes les itérations de l'algorithme 1.3. Donc les deux algorithmes donnent $u_H^{n,0} + u_h^{n,0} = u^0$. En choisissant maintenant les fonctions test orthogonales à V^0, on voit que les composantes $u_H^{n,\perp}$ et $u_h^{n,\perp}$ satisfont exactement les mêmes équations pour les deux algorithmes et donc $u_H^n + u_h^n$ sont les mêmes.

L'algorithme des patchs harmoniques prend toute sa valeur dans le cas général de triangulations non emboîtées. On verra que la convergence de l'algorithme 1.2 peut être extrêmement lente dans ce cas. Par contre, l'algorithme 1.3 conserve les bonnes propriétés de convergence du cas emboîté, même sur des triangulations quelconques. La solution finale u_{Hh} obtenue par l'algorithme 1.3 est en général légèrement différente de celle obtenue par l'Algorithme 1.2. On a toutefois la même majoration d'erreur (1.7).

4. La méthode "Semi-Schwarz"

La différence majeure entre les algorithmes des patchs 1.2, 1.3, d'un côté, et l'algorithme Chimère, de l'autre, réside dans la façon de récupérer la solution globale à partir des solutions grossière et fine. En effet, dans les algorithmes des patchs on pose $u_{Hh}^n = u_H^n + u_h^n$ tandis que dans Chimère u_{Hh}^n peut être défini comme $\{u_h^n$ à l'intérieur de Λ, u_H^n à l'extérieur de $\Lambda\}$. Si la triangulation $\mathcal{T}_h$ est emboîtée dans $\mathcal{T}_H$, alors il est facile de passer d'une représentation à l'autre en posant $w_h^n = u_H^n|_\Lambda + u_h^n$. En éliminant u_h^n de l'algorithme des patchs 1.2, on obtient l'algorithme 1.4 (proposé dans [22]). Les inconnues des itérations ici sont w_h^n et u_H^n et on passe l'information de u_H^n à w_h^n de la même manière que dans Chimère, autrement dit l'algorithme alternatif de Schwarz. Par contre, on ne peut plus faire la même chose dans l'autre direction, de w_h^n à u_H^n, d'où le nom de cet algorithme "Semi-Schwarz".

Les algorithmes 1.2 et 1.4 sont équivalents par construction dans le cas des maillages emboîtés. Dans une situation générale, les relations entre eux sont assez compliquées à cause de l'interpolation

Entrées : fonctions $u_H^0 \in V_{gH}$ et $\lambda_h^0 \in M_h$.
pour $n = 1...N$ **faire**
Trouver $u_H^n \in V_{gH}$ tel que

$$\int_\Omega K\nabla u_H^n \cdot \nabla v_H = \int_{\Omega\setminus\Lambda} f v_H - \int_{\partial\Lambda} \lambda_h^{n-1} v_H + \int_\Lambda K\nabla u_H^{n-1} \cdot \nabla v_H, \ \ \forall v_H \in V_{0H}. \quad (1.13)$$

Trouver $w_h^n \in V_h$ $\lambda_h^n \in M_h$ tels que

$$\begin{aligned} &\int_\Lambda K\nabla w_h^n \cdot \nabla v_h - \int_{\partial\Lambda} \lambda_h^n v_h = \int_\Lambda f v_h, \ \ \forall v_h \in V_{0h}, \quad (1.14)\\ &\int_{\partial\Lambda} w_h^n \mu_h = \int_{\partial\Lambda} u_H^n \mu_h, \quad \forall \mu_h \in M_h. \end{aligned}$$

fin

Algorithme 1.5: Semi-Schwarz-Lagrange

présente maintenant dans la relation $w_h^n \approx u_H^n|_\Lambda + u_h^n$. On voit, en pratique, que l'algorithme 1.4 est plus proche de la version harmonique 1.3, sa performance étant située entre celles de l'algorithme 1.2 et l'algorithme 1.3, voir [22]. On constate ainsi un certain avantage du "Semi-Schwarz" devant l'algorithme des patchs original et cela sans l'introduction d'un problème auxiliaire comme dans la version harmonique. Noter aussi que le sous-problème fin est plus facile à mettre en œuvre dans "Semi-Schwarz" que dans l'algorithme original parce que le terme mixte $\int_\Lambda K\nabla u_H^n \cdot \nabla v_h$, délicat du point de vue de l'implémentation, n'y figure plus.

5. La méthode "Semi-Schwarz-Lagrange"

On finit ce panorama des méthodes du type "Zoom Numérique" encore par une réécriture de l'algorithme 1.4. L'idée est d'interpréter le terme $\int_\Lambda K\nabla w_h^{n-1} \cdot \nabla v_H$ comme une approximation de $\int_{\partial\Lambda} K(\partial_n w_h^{n-1}) v_H$ et d'introduire une nouvelle variable auxiliaire λ_h^n (qui vit sur $\partial\Lambda$) comme une approximation de la dérivée normale $\partial_n w_h^n$). On introduit donc l'espace M_h des traces sur $\partial\Lambda$ de fonctions éléments finis de V_h. La valeur de $\lambda_h^n \in M_h$ peut être facilement récupérée en tant que multiplicateur de Lagrange dans le sous-problème fin. Cela donne l'algorithme 1.5 qu'on nomme "Semi-Schwarz-Lagrange". Pour symétriser le problème fin (1.14), on a choisi d'imposer les conditions aux limites pour w_h^n de manière faible.

Encore une fois, ce nouvel algorithme est équivalent à l'algorithme 1.4 et donc aux algorithmes 1.2, 1.3 dans le cas des maillages emboîtés. Mais en général, c'est une méthode différente de celles introduites plus haut. Elle a au moins cet avantage sur "Semi-Schwarz" que $\int_{\partial\Lambda} \lambda_h^{n-1} v_H$ est moins coûteux à évoluer en pratique que $\int_\Lambda K\nabla w_h^{n-1} \cdot \nabla v_H$. De plus, cette méthode peut être réutilisée dans les problèmes du type "multi-échelles" et "multi-modèles" comme décrit dans Chapitre 5.

2 La méthode des patchs d'éléments finis

On s'intéresse ici aux algorithmes des patchs 1.2 et 1.3 présentés dans le chapitre précédent. On commence par une étude des erreurs dans l'algorithme 1.2 et de sa convergence en tant qu'une méthode itérative. A la fin, on étudiera les mêmes questions pour la version accélérée : l'algorithme des patches harmoniques 1.3.

2.1 Les estimations d'erreur *a priori* pour la méthode des patchs

D'après les articles [16, 19].

On rappelle d'abord les notations : soit $\Omega \subset \mathbb{R}^2$ un domaine borné polygonal et $\Lambda \subset\subset \Omega$ un autre domaine polygonal "beaucoup plus petit" que Ω. Etant donné $f \in H^{-1}(\Omega)$ et une fonction positive $K \in L^\infty(\Omega)$, on considère le problème de Poisson-Dirichlet : trouver u tel que

$$-\nabla \cdot (K\nabla u) = f \text{ dans } \Omega \text{ et } u = 0 \text{ sur } \partial\Omega.$$

La généralisation au cas des conditions aux limites non-homogènes $u = g$ sur $\partial\Omega$ est évidente.

La formulation faible du problème s'écrit : trouver $u \in V = H_0^1(\Omega)$ tel que

$$a(u, v) = (f, v), \quad \forall v \in V, \tag{2.1}$$

où $a : V \times V \to \mathbb{R}$ est une forme bilinéaire symétrique, continue et coercive

$$a(u, v) = \int_\Omega K\nabla u \cdot \nabla v \tag{2.2}$$

et $(f, v) = \int_\Omega fv$. On s'intéresse par une situation "multi-échelles" où la donnée f de (2.1) a une forte variation dans Λ. Ainsi la discrétisation de (2.1) par une méthode des éléments finis ne peut pas

être faite en utilisant des grilles de la même résolution sur $\bar{\Omega} \setminus \Lambda$ et sur $\bar{\Lambda}$. C'est pourquoi on introduit deux espaces d'éléments finis :

$$V_{0H} = \{g \in C_0(\overline{\Omega}) \text{ telles que } \ g|_K \in \mathbb{P}_r(K), \forall K \in \mathcal{T}_H \text{ et } g = 0 \text{ sur } \partial\Omega\}$$

où $\mathcal{T}_H$ est une triangulation régulière de $\bar{\Omega}$ et $\mathbb{P}_r(K)$ est l'espace des polynômes de degré $\leq r$ sur un triangle $K \in \mathcal{T}_H$, et

$$V_{0h} = \{g \in C_0(\overline{\Omega}) \text{ tel que } \ g|_K \in \mathbb{P}_s(K), \forall K \in \mathcal{T}_h \text{ et } g = 0 \text{ sur } \overline{\Omega} \setminus \Lambda\}$$

où $\mathcal{T}_h$ est une triangulation régulière de $\bar{\Lambda}$. En posant $V_{Hh} = V_{0H} + V_{0h}$ et en supposant que l'algorithme 1.2 converge, on voit immédiatement que la solution limite $u_{Hh} = \lim_{n\to\infty} u^n_{Hh}$ est une approximation de la solution exacte u donnée par la méthode de Galerkin

$$a(u_{Hh}, v) = \langle f|v\rangle, \quad \forall v \in V_{Hh}. \tag{2.3}$$

On rappelle que $V_{0H} \cap V_{0h}$ ne se réduit pas nécessairement à l'élément nul et la matrice du système (2.3) peut donc être non inversible. C'est une des motivations pour introduire l'algorithme itératif 1.2 pour calculer u_{Hh}. De manière plus générale, l'intérêt de l'algorithme 1.2 est de réduire le problème éléments finis non standards (2.3) à des problèmes éléments finis classiques à chaque itération qui peuvent être implémentés à l'aide des logiciels existants.

Notre premier résultat donne une estimation d'erreur *a priori* optimale pour l'approximation (2.3) et donc pour l'algorithme 1.2 en supposant toujours qu'il converge.

Théorème 2.1 *On pose $q = \max(r, s) + 1$ et on suppose que la solution u de (2.1) appartient à $H^q(\Omega)$. Alors, on a l'estimation d'erreur* a priori *suivante pour la solution u_{Hh} de (2.3) :*

$$||u - u_{Hh}||_{H^1(\Omega)} \leq C \left(H^r ||u||_{H^q(\Omega\setminus\Lambda)} + h^s ||u||_{H^q(\Lambda)}\right), \tag{2.4}$$

avec une constante C qui ne dépend que de Ω et de régularité des maillages.

Démonstration. D'après les théorèmes bien connus d'extension des fonctions dans les espaces de Sobolev, il existe un opérateur borné d'extension $E : H^q(\Omega \setminus \overline{\Lambda}) \to H^q(\Omega)$ pour lequel $Ev|_{\Omega\setminus\overline{\Lambda}} = v|_{\Omega\setminus\overline{\Lambda}}$, $\forall v \in H^q(\Omega \setminus \overline{\Lambda})$. On pose $\tilde{u} = Eu$ si $||Eu||_{H^q(\Lambda)} \leq ||u||_{H^q(\Lambda)}$ et $\tilde{u} = u$ sinon. On a de toute façon $\tilde{u} = u$ dans $\Omega \setminus \Lambda$ et donc $u - \tilde{u} \in H^q_0(\Lambda)$. De plus

$$||\tilde{u}||_{H^q(\Omega)} \leq C||u||_{H^q(\Omega\setminus\overline{\Lambda})}, \tag{2.5}$$

avec une constante générique C, et

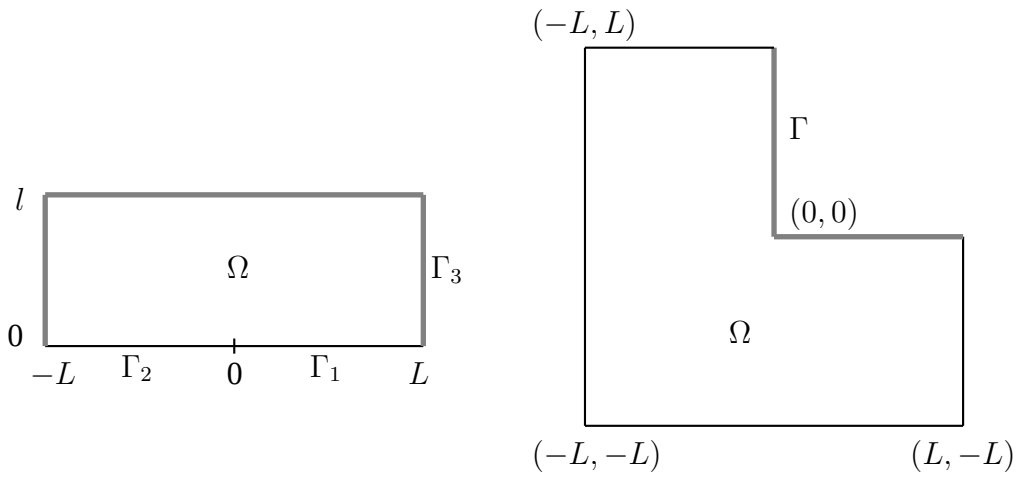
$$||\tilde{u}||_{H^q(\Lambda)} \leq ||u||_{H^q(\Lambda)}. \tag{2.6}$$

Notons par r_H et r_h les opérateurs standards d'interpolation aux espaces V_{0H} et V_{0h} respectivement. On introduit $\tilde{u}_H = r_H\tilde{u}$, $\tilde{u}_h = r_h(u - \tilde{u})$ et $\tilde{u}_{Hh} = \tilde{u}_H + \tilde{u}_h$. D'après le lemme de Céa

$$||u - u_{Hh}|| \leq ||u - \tilde{u}_{Hh}|| \tag{2.7}$$

où $||\cdot||$ est la norme d'énergie $||\cdot||^2 = a(\cdot,\cdot)$. En écrivant $u - \tilde{u}_{Hh} = (\tilde{u} - \tilde{u}_H) + [(u - \tilde{u}) - \tilde{u}_h]$, et en utilisant des résultats standards d'interpolation par les éléments finis, on voit que

$$||u - \tilde{u}_{Hh}|| \leq ||\tilde{u} - \tilde{u}_H|| + ||(u - \tilde{u}) - \tilde{u}_h|| \leq C \left(H^r ||\tilde{u}||_{H^q(\Omega)} + h^s ||u - \tilde{u}||_{H^q(\Lambda)}\right).$$



(a) (b)

FIGURE 2.1: Le domaine Ω : (a) cas du problème avec des conditions aux limites mixtes Dirichlet-Neuman ; (b) cas du domaine avec coin rentrant.

(2.6) implique $||u - \tilde{u}||_{H^q(\Lambda)} \leq 2||u||_{H^q(\Lambda)}$. Ainsi, en invoquant (2.5), on obtient

$$||u - \tilde{u}_{Hh}|| \leq C \left(H^r ||u||_{H^q(\Omega \setminus \Lambda)} + h^s ||u||_{H^q(\Lambda)} \right). \tag{2.8}$$

D'où le résultat du lemme. ■

L'algorithme des patchs peut s'avérer utile aussi dans les situations où la solution exacte n'a pas la régularité requise par le théorème 2.1. En visant par exemple les éléments finis P_1, on est dans cette situation quand u n'est pas dans $H^2(\Omega)$ même si le second membre f est dans $L^2(\Omega)$. C'est le cas, par exemple, pour le problème de Poisson-Dirichlet dans un domaine polygonale avec un coin rentrant (Fig. 2.1b) ou le problème de Poisson avec les conditions aux limites mixtes de Dirichlet-Neumann (Fig. 2.1a, avec condition de Dirichlet sur Γ_1 et celle de Neumann sur Γ_2). L'erreur de la méthode d'éléments finis standard P_1 n'est pas donc d'ordre optimal CH par rapport au pas de maillage H. On peut remarquer néanmoins que le caractère singulier est bien localisé dans ces deux exemples, autour du coin rentrant ou du point de changement du type des conditions aux limites. On peut donc améliorer le taux de convergence par rapport à H en appliquant un petit patch autour d'un point de singularité et en y introduisant un maillage fin de pas $h << H$.

Illustrons cette idée plus en détail sur l'exemple du problème de Poisson avec les conditions aux limites mixtes de Dirichlet-Neumann : soit le domaine Ω comme en Fig. 2.1a. Pour une fonction donnée $f \in L^2(\Omega)$, on cherche $u \in H^1(\Omega)$ telle que

$$\begin{cases} -\Delta u = f & \text{dans } \Omega, \\ u = 0 & \text{sur } \Gamma_1, \\ \frac{\partial u}{\partial n} = 0 & \text{sur } \Gamma_2, \\ u = 0 & \text{sur } \Gamma_3. \end{cases}$$

La formulation faible de ce problème s'écrit toujours comme (2.1) avec

$$V = \{u \in H^1(\Omega) \ / \ u|_{\Gamma_1 \cup \Gamma_3} = 0\}$$

et $K = 1$ dans (2.2). Il est bien connu que la solution u contient en général une partie singulière

$$u_s = c_0 \sqrt{r} \sin \frac{\theta}{2},$$

c'est-à-dire

$$u = u_r + u_s$$

où $u_r \in H^2(\Omega)$ est la partie régulière. Ici (r, θ) sont les coordonnées polaires et on sous-entend que le point de changement du type des conditions aux limites coïncide avec l'origine du système des coordonnées. Evidemment $u_s \notin H^2(\Omega)$, mais seulement $u_s \in H^\sigma(\Omega)$ pour tout $\sigma < 1/2$. Pour améliorer le taux de convergence de l'approximation par les éléments finis, on pose un patch Λ autour de l'origine, voir Fig. 2.2. Comme précédemment, on construit une triangulation régulière $\mathcal{T}_H$ de pas H sur Ω et une triangulation plus fine $\mathcal{T}_h$ sur Λ de pas $h << H$. La proposition suivante donne une estimation d'erreur en fonction des pas des maillages H et h et de la taille ε du patch. On va énoncer ce résultat pour une classe de solutions singulières qui inclut le cas ci-dessus, mais aussi le cas du problème avec un coin rentrant. Par exemple, la solution du problème de Poisson-Dirichlet sur le domaine dessiné en Fig. 2.1b contient la partie singulière de la forme $c_0 r^{2/3} \sin \frac{2}{3}\theta$.

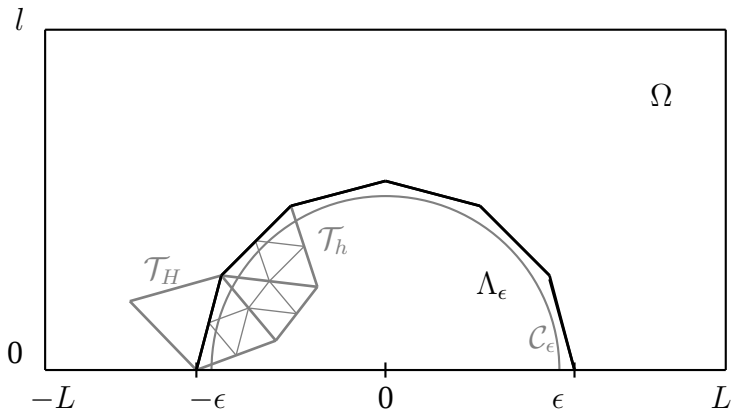


FIGURE 2.2: Position du patch.

Proposition 2.2 *Soit la solution exacte du problème (2.1) de la forme*

$$u = u_r + r^\alpha g(\theta)$$

avec $u_r \in H^2(\Omega)$, $0 < \alpha < 1$, $g(\theta) \in H^2([\theta_1, \theta_2])$ *où* θ_1, θ_2 *sont des angles tels que le domaine* Ω *est contenu dans le secteur* $\theta_1 < \theta < \theta_2$. *On suppose que* $C_\epsilon \subset \Lambda$ *où* C_ϵ *est la partie du cercle de rayon* ε *centré à* 0 *à l'intérieur de* Ω. *Alors, on a l'estimation d'erreur* a priori *suivante pour la solution* u_{Hh} *du problème (2.3) avec des élément finis* P_1 *(*$r = s = 1$*) :*

$$|u - u_{Hh}|_{H^1(\Omega)} \leq C \left(\frac{H}{\varepsilon^{1-\alpha}} + h^\alpha \right) ||g||_{H^2} + CH||u_r||_{H^2(\Omega)}$$

où $C > 0$ *est une constante qui ne dépend que de* Ω *et de régularité des maillages.*

On renvoie à [19] pour la preuve complète de cette proposition qui est assez similaire à celle du théorème 2.1. La seule différence est dans la construction de l'extension $\tilde{u}_s \in H^2(\Omega)$ de $u_s|_{\Omega \setminus \overline{\Lambda}}$ avec $u_s = r^\alpha g(\theta)$, la partie régulière u_r étant déjà dans l'espace convenable. On ne dispose pas ici d'un théorème d'extension connu mais on peut facilement construire cette extension en posant

$$\tilde{u}_s = \begin{cases} u_s, \text{ dans } \Omega \setminus C_\varepsilon, \\ \varepsilon^\alpha A(\frac{r}{\varepsilon}) g(\theta), \text{ dans } C_\varepsilon \end{cases}$$

avec une fonction appropriée $A \in C^2([0,1])$. La norme de $\tilde{u}_s$ dans $H^2(\Omega)$ donne le terme d'ordre $H/\varepsilon^{1-\alpha}$ dans l'estimation d'erreur. Remarquer que c'est vraiment le meilleur ordre possible parce que la norme de u_s dans $H^2(\Omega \setminus \overline{\Lambda})$ est aussi proportionnelle à $1/\varepsilon^{1-\alpha}$.

Le résultat de la Proposition 2.2 est important d'un point de vue pratique parce qu'il permet de choisir la taille du patch ε et le pas h de maillage associé de manière équilibrée. En revenant sur l'exemple du problème de Poisson avec les conditions aux limites mixtes de Dirichlet-Neumann (Fig. 2.1a), on a l'estimation d'erreur pour la méthode des patchs (en ne gardant que les termes essentiels)

$$|u - u_{Hh}|_{H^1(\Omega)} \leq C \left(\frac{H}{\sqrt{\varepsilon}} + \sqrt{h} \right).$$

Afin d'équilibrer les deux termes ici, on devra poser $\varepsilon \sim H^2/h$. Pour avoir une idée du bon choix de h en fonction de H, on note que le nombre d'inconnues dans le problème grossier sur $\mathcal{T}_H$ est proportionnel à $1/H^2$, tandis que celui du problème fin sur $\mathcal{T}_h$ est proportionnel à $(\varepsilon/h)^2 \sim (H/h)^4$. Pour avoir un coût de calcul sur le patch de même ordre que le coût pour le problème grossier, on choisira donc h tel que $1/H^2 \sim (H/h)^4$, c'est-à-dire

$$h \sim H^{3/2}, \quad \varepsilon \sim H^{1/2}.$$

Cela améliore l'erreur à $O(H^{3/4})$ par rapport à l'erreur de la discrétisation standard sur $\mathcal{T}_H$ qui est d'ordre $O(H^{1/2})$.

2.2 La vitesse de convergence de la méthode itérative

D'après l'article [16].

On procède maintenant à l'étude de convergence de la méthode itérative dans l'algorithme 1.2. Notons d'abord que cet algorithme s'inscrit dans le cadre général des méthodes de correction par sous-espaces. L'article [66] contient une preuve de convergence de telles méthodes. On établi ici des estimations plus précises de la vitesse de convergence, qui nous seront utiles dans la suite.

Le paramètre qui intervient habituellement dans les estimations de vitesse de convergence de méthodes de correction par sous-espaces est l'angle (ou plutôt son cosinus) entre les deux sous-espaces V_{0H} et V_{0h} deV_{Hh}. Il s'agit du nombre $\gamma \in [0;1]$ défini comme

$$\gamma = \sup_{\substack{v_h \in V_{0h}, v_h \neq 0 \\ v_H \in V_{0H}, v_H \neq 0}} \frac{a(v_h, v_H)}{||v_h|| ||v_H||}.$$

Cependant, on rappelle qu'en général $V^0 = V_{0H} \cap V_{0h}$ ne se réduit pas à l'élément zéro. Dans ce cas on a $\gamma = 1$ et le paramètre γ ne donne donc pas d'information pertinente sur la relation entre V_{0H} et V_{0h}. Une idée pour remédier à ce problème consiste à enlever l'intersection V_0 du calcul de l'angle. On introduit ainsi le paramètre suivant (cf. Fig. 2.3)

$$\tilde{\gamma} = \sup_{\substack{v_h \in \tilde{V}_h, v_h \neq 0 \\ v_H \in \tilde{V}_H, v_H \neq 0}} \frac{a(v_h, v_H)}{||v_h|| ||v_H||},$$

où $\tilde{V}_h = V_{0h} \cap V_0^{\perp}$, $\tilde{V}_H = V_{0H} \cap V_0^{\perp}$ et $V_0^{\perp}$ est le complément orthogonal (dans le produit scalaire a) de V^0 dans V_{Hh}.

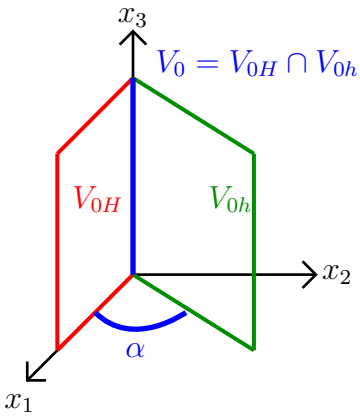


FIGURE 2.3: Une illustration géométrique de la définition du paramètre $\tilde{\gamma} = \cos\alpha$. On note que $\gamma = 1$ dans ce cas.

On introduit l'opérateur d'itération $B : V_{Hh} \to V_{Hh}$ pour exprimer l'évolution de l'erreur dans la méthode itérative 1.2

$$u_{Hh}^{n+1} - u_{Hh} = B(u_{Hh}^{n} - u_{Hh}).$$

Alors, on voit immédiatement que

$$B = (I - P_H)(I - P_h)$$

où $P_H : V_{Hh} \to V_{0H}$ et $P_h : V_{Hh} \to V_{0h}$ sont les projecteurs orthogonaux dans le produit scalaire $a(\cdot,\cdot)$ sur les sous-espaces V_{0H} et V_{0h}.

Proposition 2.3 *L'algorithme 1.2 converge, sa vitesse de convergence asymptotique (le rayon spectral de son opérateur d'itérations B) est égale à $\rho(B) = \tilde{\gamma}^2 < 1$. De plus, $||B|| = \tilde{\gamma} < 1$.*

Cette proposition découle du lemme sur la décomposition d'un espace vectoriel en une somme de plans invariants par rapport aux deux projecteurs orthogonaux donnés :

Lemme 2.4 *Soit V un espace vectoriel de dimension finie muni d'un produit scalaire $(\cdot,\cdot)$. Soient V_1 et V_2 deux sous-espaces tels que $V = V_1 + V_2$. Il existent $2p$ ($p \geq 0$) vecteurs $v_1^{(m)} \in V_1$ et $v_2^{(m)} \in V_2$, $m = 1,\ldots,p$, tels que*

$$||v_1^{(m)}|| = ||v_2^{(m)}|| = 1, \quad (v_1^{(m)}, v_2^{(m)}) = \gamma_m, \quad m = 1,\ldots,p,$$

avec

$$1 > \gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_p > 0, \tag{2.9}$$

et les plans $L_m = \mathrm{Vec}\{v_1^{(m)}, v_2^{(m)}\}$, $m = 1,\ldots,p$ sont invariants par rapport aux projecteurs orthogonaux $P_1 : V \to V_1$ et $P_2 : V \to V_1$.

L'espace V peut être décomposé en une somme directe

$$V = V_0 \oplus^{\perp} (V_1 \cap V_0^{\perp}) \oplus^{\perp} (V_2 \cap V_0^{\perp}) \oplus^{\perp} L_1 \oplus^{\perp} \cdots \oplus^{\perp} L_p, \tag{2.10}$$

où $V_0 = V_1 \cap V_2$, $m = 1,\ldots,p$. On observe que tous les termes dans (2.10) sont invariants par rapport à P_1, P_2 et

$$\gamma_1 = \sup_{\substack{v_1 \in (V_1 \cap V_0^{\perp}), v_1 \neq 0 \\ v_2 \in (V_2 \cap V_0^{\perp}), v_2 \neq 0}} \frac{a(v_1, v_2)}{||v_1||||v_2||}.$$

On renvoie à [16] pour les détails de la preuve de ce lemme. Elle consiste en une récurrence sur k, $0 \leq k \leq p$ dans laquelle on construit successivement les décompositions

$$V = V_0 \oplus^{\perp} W_k \oplus^{\perp} L_1 \oplus^{\perp} \cdots \oplus^{\perp} L_k \tag{2.11}$$

où W_k sont des sous-espaces de $V_0^{\perp}$ invariants par rapport aux opérateurs P_1 et P_2. Par exemple, pour $k = 0$, on pose $W_0 = V_0^{\perp}$ et, pour $k = 1$, on pose $V_1^{(1)} = V_1 \cap W_{k-1}$, $V_2^{(1)} = V_2 \cap W_{k-1}$,

$$\gamma_1 = \max_{v_1 \in V_1^{(1)}, v_2 \in V_2^{(1)} ||v_1||=||v_2||=1} (v_1, v_2). \tag{2.12}$$

On prend pour $v_1^{(1)}$ et $v_2^{(1)}$ les vecteurs sur lesquels le maximum dans (2.12) est atteint. On peut vérifier que les projecteurs P_1 et P_2 renvoient les vecteurs $v_1^{(1)}$ et $v_2^{(1)}$ l'un à l'autre et l'on procède de la même manière pour $k > 1$ jusqu'à l'épuisement des paires de vecteurs qui donnent un maximum positif dans les formules du type (2.12). On peut identifier alors ce qui reste de l'espace V avec $(V_1 \cap V_0^{\perp}) \oplus^{\perp} (V_2 \cap V_0^{\perp})$.

Avec le lemme 2.4, il est facile maintenant de démontrer la proposition 2.3. En effet, on pose

$$V = V_{Hh} = V_{0H} + V_{0h}, \quad V_1 = V_{0H}, \quad V_2 = V_{0h}.$$

La décomposition (2.10) montre alors que l'opérateur d'itération B agit sur chaque plan L_k séparément. Ainsi, son action sur L^k dans la base $(v_1^{(k)}, v_2)$ peut être représentée par la matrice

$$\mathbf{B}_k = \begin{pmatrix} 0 & -\gamma_k \\ 0 & \gamma_k^2 \end{pmatrix} \tag{2.13}$$

dont le rayon spectral est γ_k^2 et la norme γ_k. En parcourant tous les plans L_k ainsi que les autres sous-espaces dans (2.10) sur lesquels l'opérateur B se réduit à zéro, on obtient le résultat de la proposition 2.3.

Remarque 2.5 *L'algorithme 1.2 peut être considéré comme une méthode de Gauss-Seidel par blocs pour résoudre (2.3). Il était présenté dans [16] sous une forme légèrement différente qui permet d'y inclure une utilisation possible de sous- ou sur-relaxation ce qui le rend similaire à la méthode SOR par blocs. Cela donne l'algorithme 2.1. On retrouve l'algorithme des patchs original 1.2 en posant* $\omega = 1$.

L'analyse de cet algorithme est semblable au précedent. En effet, son opérateur d'itérations s'écrit aussi à l'aide des projecteurs

$$B = (I - \omega P_H)(I - \omega P_h).$$

On peut donc appliquer encore une fois le lemme 2.4 pour obtenir, après quelques calculs détaillés en [16], les formules pour le rayon spectral

$$\rho(B) = \begin{cases} \frac{\omega^2\tilde{\gamma}^2}{2} - \omega + 1 + \frac{\omega\tilde{\gamma}}{2}\sqrt{\omega^2\tilde{\gamma}^2 - 4\omega + 4}, & \textit{si } \omega \leq \omega_0(\tilde{\gamma}), \\ \omega - 1, & \textit{sinon}, \end{cases}$$

avec

$$\omega_0(\tilde{\gamma}) = \frac{2 - 2\sqrt{1 - \tilde{\gamma}^2}}{\tilde{\gamma}^2}$$

Entrées : une fonction $u_h^0 \in V_{0h}$ et un nombre $\omega \in (0,2)$.
pour $n = 1...N$ **faire**
Trouver $w_h \in V_{0h}$ tel que

$$a(w_h, v_H) = (f, v_H) - a(u_{Hh}^{n-1}, v) \ \ \forall v_h \in V_{0h}.$$

Poser

$$u_{Hh}^{n-\frac{1}{2}} = u_{Hh}^{n-1} + \omega w_h$$

. Trouver $w_H \in V_{0H}$ tel que

$$a(w_H, v) = \langle f|v\rangle - a(u^{n-\frac{1}{2}}, v), \ \ \forall v_H \in V_{0H}.$$

Poser

$$u_{Hh}^{n} = u_{Hh}^{n-\frac{1}{2}} + \omega w_H$$

fin

Algorithme 2.1: Méthode des patchs relaxée

et la norme de B

$$||B|| = \frac{1}{2}\omega\,(2-\omega)\,\tilde{\gamma} + \sqrt{\frac{1}{4}\omega^2\,(2-\omega)^2\,\tilde{\gamma}^2 + \,(\omega-1)^2}.$$

En particulier, $||B|| < 1$ *pour tous* $\tilde{\gamma} \in [0,1[$ *et* $\omega \in]0,2[$.

La vitesse de convergence asymptotique $\rho(B)$ *ne dépend donc que de* $\tilde{\gamma}$ *et du paramètre de relaxation* ω*. Cette dépendance est illustrée par Fig. 2.4.*

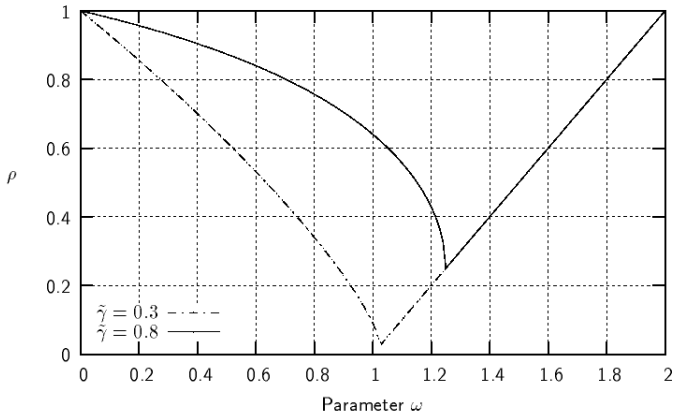


FIGURE 2.4: Le rayon spectral $\rho(B)$ en fonction de ω pour deux valeurs fixée de $\tilde{\gamma}$.

2.3 Une accélération : patchs harmoniques

D'après la note [18].

On a vu que l'algorithme 1.2 converge et son taux de convergence asymptotique est donné par $\tilde{\gamma}(V_{0H}, V_{0h})^2$. On peut constater ainsi que la convergence est lente si le paramètre $\tilde{\gamma}(V_{0H}, V_{0h})$ est proche de 1, c'est-à-dire lorsque il existe des paires des fonctions $v_H \in V_{0H}$ et $v_h \in V_{0h}$ telles que $v_h \approx v_H$. Noter que même une version plus générale de l'algorithme 1.2 utilisant une sur-relaxation peut permettre d'améliorer le taux de convergence de seulement $1 - \sqrt{8(1-\tilde{\gamma})} + O(1-\tilde{\gamma})$ quand $\tilde{\gamma} \to 1$. Cette très mauvaise convergence a été en effet observée dans la pratique, et c'est presque inévitable dans les situations des maillages non-emboîtées quelconques. En effet, n'importe quelle fonction $v_H \in V_{0H}$ avec supp$(v_H) \in \Lambda$ sera très proche d'une fonction $v_h \in V_{0h}$ en prenant pour v_h comme l'interpolé nodal de v_H sur $\mathcal{T}_h$. On aura ainsi $v_h \approx v_H$ bien que $v_H \notin V_{0h}$

Il y a toutefois au moins une situation dans laquelle l'algorithme 1.2 converge assez rapidement, à savoir lorsque les triangulations $\mathcal{T}_H$ et $\mathcal{T}_h$ sont emboîtées. Cette observation expérimentale est facile à interpréter au vu de la proposition 2.3 et du lemme 2.4. En effet, pour chaque fonction $v_H \in V_{0H}$ avec supp$(v_H) \in \Lambda$, on trouve une fonction $v_h \in V_{0h}$ telle que $v_h = v_H$. Comme l'égalité ici est exacte cette paire de fonctions n'intervient pas dans la formule pour $\tilde{\gamma}$ contrairement au cas non emboîté. Pour voir cela plus en détail, on définit

$$V_H^0 = \{v_H \in V_{0H} : v_H(x) = 0 \; \forall x \in \Omega \setminus \bar{\Lambda}\}, \tag{2.14}$$

de sorte que $V_H^0 = V_0 = V_{0H} \cap V_{0h}$. On introduit aussi $\bar{V}_H$ comme le complément orthogonal (au sens du produit scalaire a) de V_0 dans V_{0H} :

$$\bar{V}_H = \{v_H \in V_{0H} : a(v_H, \phi_H) = 0 \; \forall \phi_H \in V_H^0\}. \tag{2.15}$$

Cet ensemble se compose de fonctions éléments finis qui sont approximativement a-harmoniques à l'intérieur de Λ sur la triangulation $\mathcal{T}_H$. L'espace $\bar{V}_H$ est ainsi approximativement a-orthogonal à l'espace V_{0h}, parce que $a(\phi_H, \phi_h)$ doit être petit pour tout $\phi_H \in \bar{V}_H$ et pour tout $\phi_h \in V_{0h}$ (ϕ_h s'annule à la frontière $\partial\Lambda$ par définition de V_{0h}). C'est pourquoi on peut s'attendre à ce que le paramètre $\tilde{\gamma}(V_{0H}, V_{0h}) = \tilde{\gamma}(\bar{V}_H, V_{0h})$ soit petit dans le cas des triangulations emboîtés. Notons que V_H^0 est un sous-espace de V_{0h}. Ainsi l'approximation Galerkin u_{Hh} donnée par (2.3) ne changera pas si on y remplace $V_{Hh} = V_{0H} + V_{0h}$ par $\bar{V}_{Hh} = \bar{V}_H + V_{0h}$. De plus, on pourra remplacer l'espace V_{0H} par $\bar{V}_H$ sur toutes les itérations de l'algorithme 1.2 et cela ne changera pas les itérés u_{Hh}^n.

On s'inspire de la dernière remarque pour modifier l'algorithme 1.2 en remplaçant l'espace V_{0H} par son sous-espaces ne contenant que les fonctions approximativement harmoniques à l'intérieur de Λ (et c'est dans le cadre général des maillages non emboîtés) dans l'espoir de transférer les bonnes propriétés de convergence du cas emboîtés au cadre général de triangulations arbitraires. Plus précisément, on conserve les notations (2.14) et (2.15) pour définir aussi $\bar{V}_{Hh} = \bar{V}_H + V_{0h}$ dans un cas non emboîtés. Notons que V_H^0 se compose désormais de toutes les fonctions de V_{0H} avec des supports à l'intérieur de $\bar{\Lambda}_H$ où $\bar{\Lambda}_H$ est une union des triangles (fermés) de $\mathcal{T}_H$ qui se trouvent entièrement à l'intérieur de $\bar{\Lambda}$. Soit u la solution de (2.1). On cherche maintenant son approximation $\bar{u}_{Hh} \in \bar{V}_{Hh}$ telle que

$$a(\bar{u}_{Hh}, v) = (f, v), \quad \forall v \in \bar{V}_{Hh}. \tag{2.16}$$

Pour calculer $\bar{u}_{Hh}$ on va utiliser l'algorithme 1.2 en remplacant partout V_{0H} par $\bar{V}_H$. Ainsi, on rencontrera des problèmes du type : étant donné $g \in V'_{0H}$, trouver $u_H \in \bar{V}_H$ tel que

$$a(u_H, v) = \langle g|v\rangle, \quad \forall v \in \bar{V}_H. \tag{2.17}$$

Dans la pratique, on n'a pas de base explicite des éléments finis pour l'espace $\bar{V}_H$ de sorte qu'une implémentation directe de (2.17) est impossible. Pour contourner cette difficulté, on considère deux problèmes éléments finis standard :

(a) Trouver $\lambda_H \in V^0_{0H}$ tel que $a(\lambda_H, \mu) = \langle g|\mu\rangle, \quad \forall \mu \in V^0_H$.

(b) Trouver $u_H \in V_H$ tel que $a(u_H, v) = \langle g|v\rangle - a(\lambda_H, v), \quad \forall v \in V_H$.

On peut voir facilement que u_H dans (b) donne une solution de (2.17). En utilisant les problèmes du type (a), (b) à chaque itération, on arrive à l'algorithme 1.3 des patchs harmoniques.

En appliquant le lemme 2.4 avec $V_1 = \bar{V}_H$, $V_2 = V_{0h}$ on voit immédiatement que l'algorithme 1.3 converge vers la solution $\bar{u}_{Hh}$ de (2.16). Bien que l'espace $\bar{V}_{Hh}$ dans (2.16) soit plus pauvre que V_{Hh} dans (2.3), la solution $\bar{u}_{Hh}$ est une approximation de u à l'ordre optimal :

Proposition 2.6 *Soit $q = \max(r, s) + 1$, on suppose que la solution u de (2.1) est dans $H^q(\Omega)$. Si le patch Λ est convexe alors la solution $\bar{u}_{Hh}$ de (2.16) satisfait*

$$|u - \bar{u}_{Hh}|_{H^1(\Omega)} \leq C\left(H^r||u||_{H^q(\Omega\setminus\overline{\Lambda})} + h^s||u||_{H^q(\Lambda)}\right), \tag{2.18}$$

où H et h sont les diamètres maximaux des triangles de $\mathcal{T}_H$ et $\mathcal{T}_h$ respectivement et C est une constante indépendante de H et de h.

Démonstration. Par le lemme de Céa, il suffit de trouver un élément $\tilde{u}_{Hh} \in \bar{V}_{Hh}$ satisfaisant à l'inégalité (2.18) avec $\bar{u}_{Hh}$ remplacé par $\tilde{u}_{Hh}$. Pour cela, on choisit $\tilde{u} \in H^1_0(\Omega)$ égal à u dans $\Omega \setminus \overline{\Lambda}$ et a-harmonique dans Λ. Alors

$$||\tilde{u}||_{H^q(\Lambda)} \leq C||u||_{H^{q-\frac{1}{2}}(\partial\Lambda)} \leq C||u||_{H^q(\Omega\setminus\overline{\Lambda})}$$

et

$$||u - \tilde{u}||_{H^q(\Lambda)} \leq ||u||_{H^q(\Lambda)} + C||u||_{H^{q-\frac{1}{2}}(\partial\Lambda)} \leq C||u||_{H^q(\Lambda)}$$

où C désigne des constantes génériques. On introduit maintenant $\tilde{u}_H$ comme la projection a-orthogonale de $\tilde{u}$ sur V_{0H}. 0n a alors

$$a(\tilde{u}_H, v_H) = a(\tilde{u}, v_H) = 0, \quad \forall v_H \in V^0_H,$$

c'est-à-dire $\tilde{u}_H \in \bar{V}_H$. En introduisant $\tilde{u}_h \in V_{0h}$ comme l'interpolé de $u - \tilde{u}$ sur V_{0H} et $\tilde{u}_{Hh} = \tilde{u}_H + \tilde{u}_h$, on conclut la preuve de la même façon que la preuve du théorème 2.1. ■

L'algorithme 1.3 est particulièrement efficace dans le cas $\bar{\Lambda} = \bar{\Lambda}_H$. Cette situation sera dénommée "maillages conformes au bord" parce que le bord $\partial\Lambda$ est composé d'arêtes de triangles de $\mathcal{T}_h$ ou $\mathcal{T}_H$. Le taux de convergence de l'algorithme 1.3 peut être estimé de façon indépendante de H et h dans ce cas.

Proposition 2.7 *Si $\bar{\Lambda} = \bar{\Lambda}_H$, alors la vitesse de convergence de l'algorithme 1.3 est majorée par une constante $\kappa \in]0, 1[$ qui ne dépend que de Ω, Λ et de la régularité des maillages.*

Démonstration. Par le lemme 2.3, la vitesse de convergence asymptotique de l'algorithme 1.3 est donnée par $\tilde{\gamma}(\bar{V}_H, V_{0h})^2$. Pour estimer cette quantité, on prend n'importe quelle fonction $v_H \in \bar{V}_H, v_H \neq 0$ et on examine l'extension harmonique v de $v_H|_{\partial\Lambda}$ à l'intérieur de Λ, c'est-à-dire la fonction $v \in H^1(\Lambda)$ satisfaisant

$$\Delta v = 0 \quad \text{dans} \quad \Lambda, \quad v|_{\Omega\setminus\Lambda} = v_H|_{\Omega\setminus\Lambda}.$$

Soit $u_H \in V_H^0$ l'interpolant de Scott-Zhang [63] de v sur $\mathcal{T}_H$ à l'intérieur de Λ. On a alors

$$u_H|_{\Omega\setminus\Lambda} = v_H|_{\Omega\setminus\Lambda} \text{ et } |u_H - v|_{1,\lambda} \leq C_{SZ}|v|_{1,\Lambda}$$

avec une constante C_{SZ} qui ne dépend que de Λ et de la régularité de $\mathcal{T}_H$. Puisque

$$a(v_H - v, u_H - v_H) = 0,$$

on voit que

$$|v_H - v|_{1,\lambda}^2 = |u_H - v|_{1,\lambda}^2 - |u_H - v_H|_{1,\lambda}^2 \leq |u_H - v|_{1,\lambda}^2 \leq C_{SZ}^2|v|_{1,\Lambda}^2.$$

De même, puisque

$$a(v_H - v, v) = 0,$$

on voit que

$$|v_H|_{1,\Lambda}^2 = |v_H - v|_{1,\Lambda}^2 + |v|_{1,\Lambda}^2 \geq (1 + C_{SZ}^{-2})|v_H - v|_{1,\Lambda}^2.$$

Donc, on a pour tous $v_H \in \bar{V}_H$ et $v_h \in V_h$

$$a(v_H, v_h) = a(v_H - v, v_h) \leq \frac{1}{\sqrt{1 + C_{SZ}^{-2}}}|v_H|_{H^1(\Lambda)}|v_h|_{H^1(\Lambda)} \leq \frac{1}{\sqrt{1 + C_{SZ}^{-2}}}|v_H|_{H^1(\Lambda)}|v_h|_{H^1(\Lambda)}$$

ce qui implique

$$\tilde{\gamma}(\bar{V}_H, V_{0h}) \leq \frac{1}{\sqrt{1 + C_{SZ}^{-2}}}.$$

■

3 Sur la méthode Chimère

3.1 Convergence de l'algorithme de Schwarz sur des maillages arbitraires non-conformes

D'après l'article [32].

Bien que la convergence de l'algorithme de Schwarz soit bien établie au niveau continu depuis le début de l'histoire des méthodes de décomposition de domaine, une théorie complète manque toujours au niveau discret surtout dans le cas de superposition arbitraire de maillages. En effet, la convergence de cet algorithme sur des maillages uniformes a été démontrée seulement dans le cadre de différences finies dans [41]. La preuve est basée sur le principe du maximum et sur la décroissance exponentielle des solutions des EDP elliptiques loin de la frontière. On a utilisé des idées similaires dans [32] dans le cas des éléments finis P_1 sur des maillages triangulaires arbitraires.

On rappelle les conventions introduites dans la description de l'algorithme Chimère 1 :

$$\Omega = \Omega_H \cup \Lambda, \;\; \partial\Lambda \subset \Omega_H, \;\; \partial\Lambda \cap \Gamma = \emptyset, \;\; \partial\Omega_H \backslash \Gamma \subset \Lambda. \tag{3.1}$$

On utilise les notations $S_h = \partial\Lambda$ et $S_H = \partial\Omega_H \backslash \Gamma$. Etant données les triangulations $\mathcal{T}_H$ sur Ω_H et $\mathcal{T}_h$ sur Λ, on introduit les espaces d'éléments finis P_1 sur $\mathcal{T}_H$ et $\mathcal{T}_h$ qu'on désigne par V_H et V_h respectivement, $V_{0H} = V_H \cap H_0^1(\Omega_H)$ et $V_{0h} = V_h \cap H_0^1(\Lambda)$. Pour simplifier les notations, on considère ici seulement le cas du problème de Laplace-Poisson (1.1) avec $K = 1$. On note donc $a(u,v) = \int_\Omega \nabla u \cdot \nabla v$,

$$a_h(u,v) = \int_\Lambda \nabla u \cdot \nabla v \quad \text{et} \quad a_H(u,v) = \int_{\Omega_H} \nabla u \cdot \nabla v. \tag{3.2}$$

Hypothèse 3.1 (principe du maximum discret) *On suppose que tous les coefficients non-diagonaux des matrices de rigidité de a_h sur V_h et de a_H sur V_H sont non-positifs. Par conséquent, si $u_h \in V_h$*

(resp $u_H \in V_H$) est approximativement harmonique, c'est-à-dire

$$a_h(u_h, \phi_h) = 0, \quad \forall \phi_h \in V_{0h},$$
$$a_H(u_H, \phi_H) = 0, \quad \forall \phi_H \in V_{0H},$$

alors le maximum de u_h (resp u_H) est atteint sur la frontière de Λ (resp Ω_H).

Hypothèse 3.2 (une variante de principe du maximum fort sur Λ) *On suppose que la solution $\nu_H \in V_H$ de*

$$a_H(\nu_H, \phi_H) = 0, \ \ \forall \phi_H \in V_{0H}, \ \ \nu_H|_{S_H} = 1, \ \ \nu_H|_{\Gamma_H} = 0,$$

est telle que $|\nu_H|_{L^\infty(S_h)} := \lambda < 1$.

On peut montrer que le principe de maximum discret est validé (pour les éléments finis P_1) sur les triangulations dont tous les angles sont aigus [42]. Les hypothèses 3.1 et 3.2 impliquent

$$\|v_H\|_{L^\infty(S_h)} \leq \lambda \|v_H\|_{L^\infty(\Omega_H)} \tag{3.3}$$

pour toute fonction $v_H \in V_H$ telle que $a_H(v_H, \phi_H) = 0, \ \ \forall \phi_H \in V_{0H}$ et $v_H|_\Gamma = 0$.

Proposition 3.3 *Sous les hypothèses 3.1 et 3.2, l'algorithme de Schwarz converge vers la solution unique $(u_h^*, u_H^*) \in V_h \times V_H$ de*

$$\begin{aligned} &a_H(u_H^*, w_H) = (f, w_H), \ \ \forall w_H \in V_{0H}, \ \ u_H^*|_{S_H} = \gamma_H u_h^*, \ \ u_H^*|_{\Gamma_H} = g_H \\ &a_h(u_h^*, w_h) = (f, w_h), \ \ \forall w_h \in V_{0h}, \ \ u_h^*|_{S_h} = \gamma_h u_H^*. \end{aligned} \tag{3.4}$$

Démonstration. Considérons le problème couplé : étant donné $u_h \in V_h$, trouver $v_H \in V_H, \ v_h \in V_h$ tels que

$$\begin{aligned} &a_H(v_H, w_H) = 0, \ \ \forall w_H \in V_{0H}, \ \ v_H|_{S_H} = r_H u_h, \ \ v_H|_{\Gamma_H} = 0 \\ &a_h(v_h, w_h) = 0, \ \ \forall w_h \in V_{0h}, \ \ v_h^{m+1}|_{S_h} = r_h v_H. \end{aligned}$$

Soit $T : V_h \to V_h$ l'application affine définie par $T(u_h) = v_h$. Par le principe du maximum et du fait que les interpolations r_H et r_h n'augmentent pas la norme L^∞, on voit que

$$\|v_H\|_{L^\infty(\Omega_H)} \leq \|u_h\|_{L^\infty(S_H)}, \quad \|v_h\|_{L^\infty(\Lambda)} \leq \|v_H\|_{L^\infty(S_h)}.$$

et donc, par l'hypothèse 3.2,

$$\|v_h\|_{L^\infty(\Lambda)} \leq \|v_H\|_{L^\infty(S_h)} \leq \lambda \|v_H\|_{L^\infty(\Omega_H)} \leq \lambda \|u_h\|_{L^\infty(\Lambda)}.$$

Par conséquent, T est une contraction dans la norme $L^\infty(\Lambda)$. Par le théorème de contraction de Banach, le processus itératif $u_h^m = T u_h^{m-1}$, qui est l'algorithme de Schwarz, converge vers u_h^* le point fixe unique de T. ■

Afin d'obtenir une estimation d'erreur pour l'algorithme de Schwarz, on aura besoin des estimations d'erreur pour les éléments finis standards [61] : soit $\Pi_h u \in V_h$ la solution de

$$a_h(\Pi_h u, w_h) = a_h(u, w_h) \ \ \forall w_h \in V_{0h}, \ \ \Pi_h u = r_h u \text{ sur } S_h.$$

Alors,

$$||\Pi_h u - u||_{L^\infty(\Lambda)} \leq h^2 \log \frac{1}{h} ||u||_{H^{2,\infty}(\Lambda)},$$

On a une estimation analogue pour $\Pi_H u$ définie dans V_H au lieu de V_h.

Proposition 3.4 *Sous les hypothèses 3.1 et 3.2, la solution de Schwarz (u_h^*, u_H^*) dans (3.4) satisfait*

$$\max(||u_H^* - u||_{L^\infty(\Omega_H)}, ||u_h^* - u||_{L^\infty(\Omega_H)}) \\ \leq C(H^2 \log \frac{1}{H} ||u||_{H^{2,\infty}(\Omega_H)} + h^2 \log \frac{1}{h} ||u||_{H^{2,\infty}(\Omega_h)}), \tag{3.5}$$

avec une constante C qui ne dépend que des domaines Ω_H et Ω_h.

Démonstration. La solution exacte u du problème de Poisson satisfait $u|_\Gamma = 0$ et

$$a_H(u, w) = (f, w) \ \forall w \in H_0^1(\Lambda), \ \ u = r_H u + (u - r_H u) \text{ sur } S_H,$$
$$a_h(u, w) = (f, w) \ \forall w \in H_0^1(\Lambda), \ \ u = r_h u + (u - r_h u) \text{ sur } S_h.$$

Soit $e = u_H^* - u$ et $\varepsilon = u_h^* - u$. On choisit $w = w_H$ dans la première équation et $w = w_h$ dans la seconde :

$$a_H(e, w_H) = 0 \ \forall w_H \in V_{0H}, \ \ e = r_H \varepsilon - (u - r_H u) \text{ sur } S_H, \ \ e|_\Gamma = 0$$
$$a_h(\varepsilon, w_h) = 0 \ \forall w_h \in V_{0h}, \ \ \varepsilon = r_h e - (u - r_h u) \text{ sur } S_h.$$

Soient

$$\varepsilon_H = u_H - \Pi_H u = e + u - \Pi_H u, \quad \varepsilon_h = u_h - \Pi_h u = \varepsilon + u - \Pi_h u.$$

Alors $\varepsilon_H \in V_H$, $\varepsilon_h \in V_h$ et

$$a_H(\varepsilon_H, w_H) = 0 \ \forall w_H \in V_{0H}, \ \ \varepsilon_H = r_H(\varepsilon_h + \Pi_h u - u) \text{ sur } S_H, \ \ \varepsilon_H|_\Gamma = 0$$
$$a_h(\varepsilon_h, w_h) = 0 \ \forall w_h \in V_{0h}, \ \ \varepsilon_h = r_h(\varepsilon_H + \Pi_H u - u) \text{ sur } S_h.$$

Le principe de maximum donne

$$\|\varepsilon_H\|_{L^\infty(\Omega_H)} \leq \|\Pi_h u - u\|_{L^\infty(S_H)} + \|\varepsilon_h\|_{L^\infty(S_H)},$$
$$\|\varepsilon_h\|_{L^\infty(\Lambda)} \leq \|\Pi_H u - u\|_{L^\infty(S_h)} + \|\varepsilon_H\|_{L^\infty(S_h)},$$

et par (3.3)

$$\|\varepsilon_H\|_{L^\infty(S_h)} \leq \lambda \|\varepsilon_H\|_{L^\infty(\Omega_H)}.$$

Donc

$$\max(\|\varepsilon_h\|_{L^\infty(\Lambda)}, \|\varepsilon_H\|_{L^\infty(\Omega_H)}) \leq \frac{1}{1-\lambda}(\|\Pi_H u - u\|_{L^\infty(\Omega_H)} + \|\Pi_h u - u\|_{L^\infty(\Lambda)}).$$

D'où le résultat. ■

3.2 Une extension au cas des problèmes non stationnaires

D'après l'article [26] inclus en annexe.

Les résultats sur la convergence de l'algorithme de Schwarz peuvent être réutilisés dans le cadre des discrétisation en temps et en espace de l'équation de convection-diffusion du type

$$\frac{\partial u}{\partial t} + b \cdot \nabla u - \Delta u = f \text{ dans } \Omega \text{ avec } u = g \text{ sur } \Gamma,$$
$$u|_{t=0} = u_0. \tag{3.6}$$

Bien que cela n'entre pas dans le cadre du présent manuscrit (équations multi-échelles elliptiques), on va indiquer brièvement les résultats de [26] pour donner un autre perspective possible.

La méthode de Schwarz pour l'équation de la chaleur ($b = 0$)

Comme dans la section précédente, on choisit deux sous-domaines de Ω, à savoir Ω_H et Λ satisfaisant (3.1), ainsi que les triangulations $\mathcal{T}_H$ de Ω_H et $\mathcal{T}_h$ de Λ. On garde les notations de Fig. 1.1 et on travaille avec les éléments finis P_1 qui forment les espaces V_H sur $\mathcal{T}_H$ et et V_h sur $\mathcal{T}_h$.

L'algorithme d'Euler-Schwarz est initialisé par $u_H^0 = r_H u_0$, $u_h^0 = r_h u_0$ et on procède en cherchant $u_H^n \in V_H$, $u_h^n \in V_h$, $n = 1, \ldots, N$ tels que

$$
\begin{aligned}
&(\frac{u_H^n - u_H^{n-1}}{\Delta t}, w_H)_H + a_H(u_H^n, w_H) = (f^n, w_H), \;\; u_H^n|_{S_H} = r_H u_h^{n-1},, \;\; u_H^n|_\Gamma = r_H g^n, \\
&(\frac{u_h^n - u_h^{n-1}}{\Delta t}, w_h)_h + a_h(u_h^n, w_h) = (f^n, w_h), \;\; u_h^n|_{S_h} = r_h u_H^{n-1},
\end{aligned} \tag{3.7}
$$

pour tout $w_H \in V_(0H)$ et pour tout $w_h \in V_(0h)$. Ici $f^n = f(t^n, \cdot)$, $g^n = g(t^n, \cdot)$, $a_{H,h}(u, v)$ sont définis par (3.2) et les parenthèses $(\cdot, \cdot)$ désignent le produit scalaire L^2 et $(\cdot, \cdot)_H$ (resp $(\cdot, \cdot)_h$) désignent son approximation "mass lumping" sur $\mathcal{T}_H$ (resp $\mathcal{T}_h$).

Théorème 3.5 *Sous les hypothèses 3.1, 3.2 et à condition que u, solution de (3.6) avec $b = 0$, soit suffisamment régulière, (u_h^n, u_H^n) dans (3.7) est une solution approchée pour (3.6) avec les estimations d'erreur L^∞ suivantes*

$$
\begin{aligned}
&\max_{1 \le n \le N} (||u_H^n - u(t^n, \cdot)||_{L^\infty(\Omega_H)}, ||u_h^n - u(t^n, \cdot)||_{L^\infty(\Lambda)}) \\
&\qquad \le CT \Big\{ (\Delta t + H^2 \log \frac{1}{H}) ||u||_{C^2([0,T], C(\Omega_H)) \cap C^1([0,T], H^{2,\infty}(\Omega_H))} \\
&\qquad\qquad + (\Delta t + h^2 \log \frac{1}{h}) ||u||_{C^2([0,T], C(\Lambda)) \cap C^1([0,T], H^{2,\infty}(\Lambda))} \Big\},
\end{aligned} \tag{3.8}
$$

avec $T = N\Delta t$ le temps final et C une constante qui ne dépend que des domaines Ω_H et Λ.

Remarque 3.6 *On peut associer deux pas de temps différents pour les deux sous-domaines, par exemple Δt pour les problèmes avec u_H^n et δt pour les problèmes avec u_h^n. Une variante de l'algorithme de Schwarz peut alors être construite si $\Delta t = m\delta t$ avec un entier m. On avance en temps la solution dans le zoom m fois sur chaque pas de temps du problème grossier sur Ω_H. La preuve de convergence s'étend aisément à ce cas.*

Remarque 3.7 *L'extension de nos résultats à l'opérateur $\nabla \cdot (K\nabla)$ au lieu de Δ est immédiate.*

Convection Diffusion ($b \neq 0$)

Le terme de convection dans (3.6) peut être discrétisé par le schéma des caractéristiques. La combinaison avec la méthode de Schwarz donne l'algorithme suivant : trouver $u_H^n \in V_H, \; u_h^n \in V_h$, tels que pour tous $w_H \in V_{0H}, \;\; w_h \in V_{0h}$:

$$
\begin{aligned}
&\frac{1}{\Delta t}(u_H^n - u_H^{n-1} o X^{n-1}, w_H)_H + a_H(u_H^n, w_H) = (f, w_H), \\
&u_H^n|_{S_H} = r_H u_h^{n-1}, \;\; u_H^n|_\Gamma = g_H^n, \\
&\frac{1}{\Delta t}(u_h^n - u_h^{n-1} o X^{n-1}, w_h)_h + a_h(u_h^n, w_h) = (f, w_h), \\
&u_h^n|_{S_h} = r_h u_H^{n-1},
\end{aligned} \tag{3.9}
$$

où $X^{n-1}(x) = X(t^{n-1})$ et X est la solution de $\dot{X} = b(X(t), t)$, $X(t^n) = x$. On considère le cas où X^{n-1} peut être calculé exactement (b constante par morceaux, par exemple).

Comme le montre [60], la précision de la méthode caractéristiques-Galerkin se détériore lorsque le pas de temps Δt devient trop petit devant le pas du maillage en espace. On ne peut s'attendre donc à généraliser complètement le théorème 3.5 au cas $b \neq 0$. On a, cependant, un résultat plus faible :

Théorème 3.8 *Sous les hypothèses 3.1, 3.2 et à condition que u, solution de (3.6) avec $b = 0$, soit suffisamment régulière, (u_h^n, u_H^n) dans (3.9) est une solution approchée pour (3.6) avec les estimations d'erreur L^∞ suivantes*

$$\begin{aligned}\max_{1\leq n\leq N}(||u_H^n - u(t^n)||_{L^\infty(\Omega_H)}, ||u_h^n - u(t^n)||_{L^\infty(\Lambda)}) \\ \leq CT\Big\{(\Delta t + \frac{H^2 \log \frac{1}{H}}{\Delta t})||u||_{H^{2,\infty}(\Omega_H\times[0,T])} \\ +(\Delta t + \frac{h^2 \log \frac{1}{h}}{\Delta t})||u||_{H^{2,\infty}(\Lambda\times[0,T])}\Big\},\end{aligned} \tag{3.10}$$

avec $T = N\Delta t$ le temps final et C une constante qui ne dépend que des domaines Ω_H et Λ.

4 Une comparaison numérique des algorithmes précédents

Nous illustrons maintenant les cinq algorithmes présentés dans chapitre 1 sur l'exemple du problème de Poisson-Dirichlet (2.1) avec $K = 1$, $g = 0$ dans $\Omega = (-1, 1)^2$ dont la solution exacte est

$$u = \cos(\frac{1}{2}\pi x_1)\cos(\frac{1}{2}\pi x_2) + \eta\chi(r)\exp\left(\frac{1}{\epsilon^2} - \frac{1}{|\epsilon^2 - r^2|}\right)$$

où $r = \sqrt{x_1^2 + x_2^2}$ et $\chi(r)$ est la fonction indicatrice du disque $r \leq \epsilon$. On prend les valeurs suivantes pour les paramètres : $\eta = 20$ et $\epsilon = 0.3$. Dans une région proche de $(0, 0)$ où la solution est à son apogée, nous appliquons un patch avec une grille fine. On choisit les éléments finis P_1 pour les deux espaces V_H et V_h. Le logiciel `Freefem` [47] est utilisé pour générer les grilles et pour la mise en œuvre des algorithmes.

On considère trois situations en ce qui concerne le positionnement des maillages dans les algorithmes 1.2–1.5 :

I. Le cas des maillages emboîtés comme en Fig. 4.1. On choisit ici les deux maillages structurés cartésiens : $\mathcal{T}_H$ de 21×21 nœuds sur Ω et $\mathcal{T}_h$ de 41×41 nœuds sur $\Lambda = (-0.2, 0.2)^2$. On voit donc que chaque triangle de $\mathcal{T}_H$ à l'intérieur de Λ est décomposé en 100 petits triangles de $\mathcal{T}_h$.

II. Le cas des maillages non emboîtés mais construits de telle façon que Λ est composé de quelques triangles de $\mathcal{T}_H$, voir Fig. 4.2. On appellera de tels maillages "conformes au bord" parce que le bord du patch Λ est constitué d'arêtes de triangles grossiers de $\mathcal{T}_H$, ainsi que d'arêtes de triangles fins de $\mathcal{T}_h$. On obtient ce cas test en utilisant un maillage non structuré $\mathcal{T}_H$ et en imposant des arêtes le long du bord du zoom $\Lambda = (-0.2, 0.2)^2$. Noter que les pas des maillages sont comparables à ceux du cas présenté en Fig. 4.1. En effet, on construit le maillage $\mathcal{T}_H$ avec 20 triangles adjacents à chaque côté de $\partial\Omega$, le domaine de zoom Λ est le même dans les deux cas et les deux maillages $\mathcal{T}_h$ sont presque identiques sauf que ici on a 44×44 nœuds au lieu de 41×41.

III. Le cas le plus général des maillages non emboîtés et non conformes au bord comme en Fig. 4.3. Maintenant, même le bord du patch $\partial\Lambda$ coupe arbitrairement les triangles de $\mathcal{T}_H$. Pour que les maillages de ce cas soient semblable à ceux du cas précédent, on a choisi la même triangulation grossière $\mathcal{T}_H$ qu'avant, la triangulation fine est toujours structurée cartésienne

avec 44×44 nœuds, mais elle couvre le patch $\Lambda = (-0.22, 0.22)^2$ qui est donc un peu plus large que le patch du cas II.

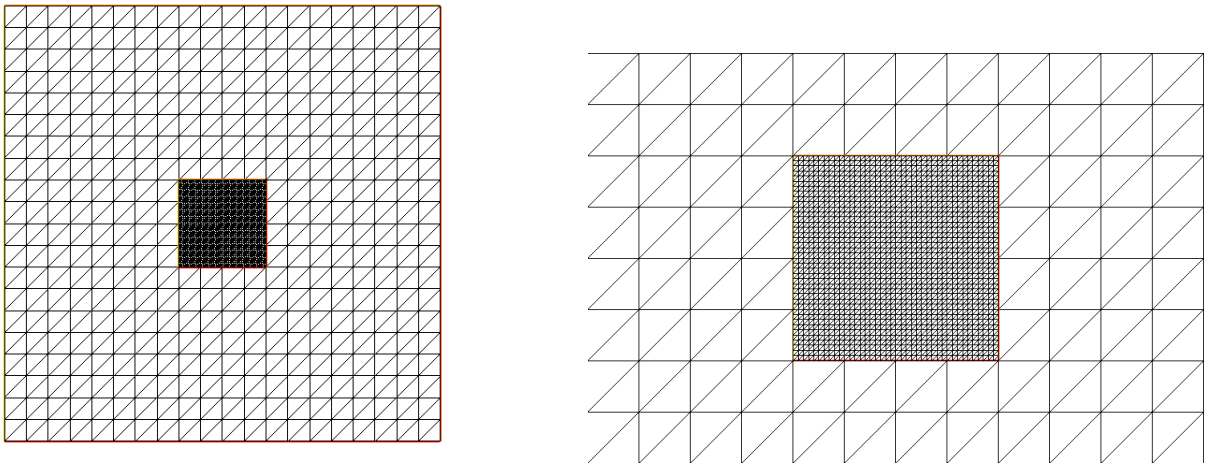


FIGURE 4.1: Les maillages emboîtés. A gauche : les deux maillages sur Ω ; à droite : un agrandissement autour de Λ.

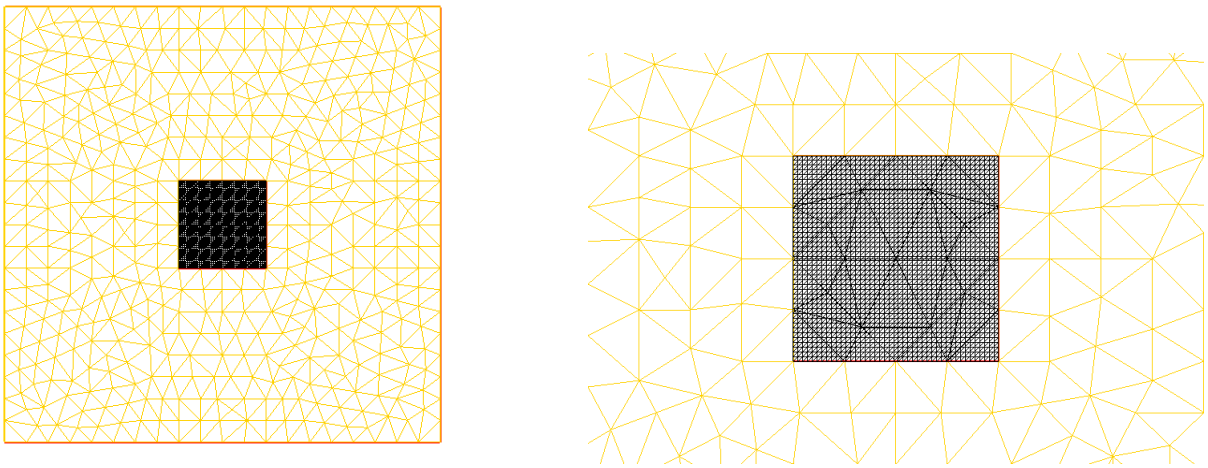


FIGURE 4.2: Les maillages non emboîtés, mais conformes au bord. A gauche : les deux maillages sur Ω ; à droite : un agrandissement autour de Λ.

La méthode Chimère 1 est testée sur des maillages comparables à ceux décrits précédemment. Comme il n'y a pas de correspondance directe entre les maillages utilisés par la méthode Chimère et ceux utilisés par les autres algorithmes 1.3–1.5, on ne présente qu'un exemple de maillages pour la méthode Chimère. On prend notamment Ω_H comme étant Ω privé du trou $(-0.2, 0.2)^2$ et $\Lambda = (-0.2, 0.2)^2$. Les maillage $\mathcal{T}_H$, $\mathcal{T}_h$ sont structurés cartésiens, $\mathcal{T}_H$ contient 21×21 nœuds (mis à part le trou) et $\mathcal{T}_h$ contient 41×41 nœuds.

Le point le plus délicat dans la mise en œuvre des algorithmes de patchs 1.2 et 1.3 est le calcul des termes "mixtes" du type $\int_\Lambda \nabla u_H \cdot \nabla v_h$ avec $u_H \in V_H$ et $u_h \in V_h$. Ce calcul ne pose pas de problème dans le cas I des maillages emboîtés. Il suffit, dans ce cas, d'employer une formule de

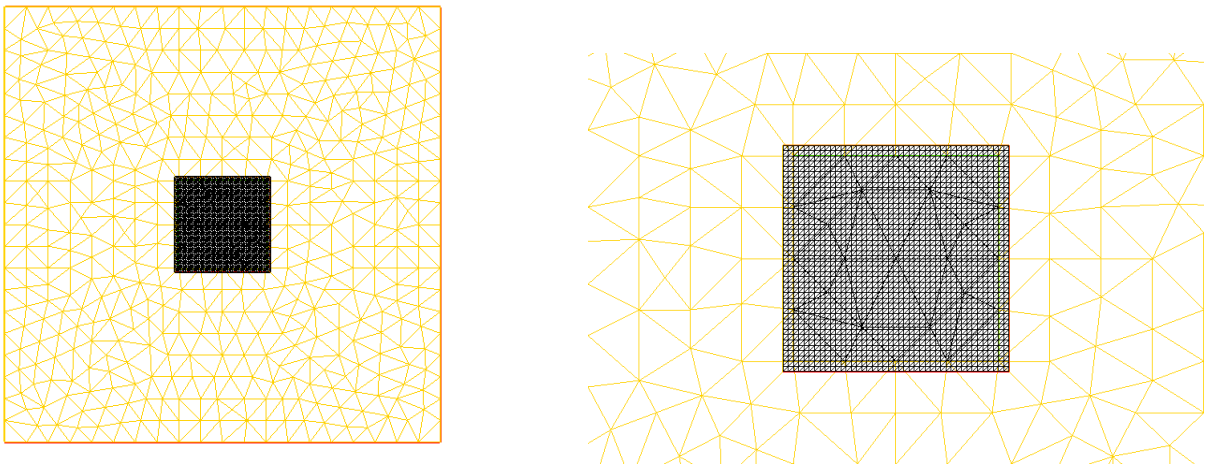


FIGURE 4.3: Les maillages non emboîtés, non conformes au bord. A gauche : les deux maillages sur Ω ; à droite : un agrandissement autour de Λ.

quadrature assez précise sur le maillage fin $\mathcal{T}_h$. Pour calculer les intégrales mixtes exactement sur des triangulations générales, on construit un maillage d'intersection entre $\mathcal{T}_H$ et $\mathcal{T}_h$, c'est-à -dire un maillage $\mathcal{T}_{Hh}$ qui est un raffinement et de $\mathcal{T}_H$ et de $\mathcal{T}_h$. Il suffit alors d'employer une formule de quadrature assez précise sur $\mathcal{T}_{Hh}$. En pratique, nous construisons $\mathcal{T}_{Hh}$ à l'aide du logiciel `Triangle`. Notons qu'un algorithme plus efficace pour la construction d'un maillage d'intersection a été proposé dans [45].

On retrouve des difficultés du même genre dans la mise en œuvre des algorithmes 1.4 et 1.5. Les termes qui posent des problèmes ici sont les intégrales du type $\int_\Lambda K\nabla u_H \cdot \nabla v_H$ ou $\int_\Lambda K\nabla w_h \cdot \nabla v_H$ avec $u_H, v_H \in V_H$, $w_h \in V_h$. Encore une fois, ces intégrales peuvent se calculer exactement à l'aide du maillage d'intersection $\mathcal{T}_{Hh}$.

Bien que possible, le calcul exact des intégrales mixtes reste un tâche difficile du point de vue algorithmique. Il est beaucoup plus facile, en pratique, d'approcher ces intégrales par des formules de quadrature sur le maillage fin $\mathcal{T}_h$. De cette façon, on introduit des erreurs d'interpolation dans les algorithmes et on ne sait pas maîtriser ces erreurs au niveau théorique pour l'instant. Néanmoins, des expériences numériques montrent que l'approximation par quadrature ne gêne pas beaucoup la convergence des algorithmes et l'erreur ainsi introduite est négligeable dans la plupart des situations.

Quelques résultats numériques sont présentés en Figs. 4.4–4.5. Dans la description de ces images, on sous-entend que les intégrales mixtes sont calculées exactement si rien n'est mentionné à ce propos. Sinon, on sous-entend par calcul inexact l'utilisation de quadratures sur $\mathcal{T}_h$. Au vu de résultats en Fig. 4.4, on peut faire les observation suivantes :

- La méthode originale des patchs 1.2 converge en général lentement, sauf dans les situations des maillages emboîtés. Sinon, cette méthode n'est pas beaucoup plus efficace que la méthode Chimère.
- La version harmonique 1.3 permet de retrouver le taux de convergence élevé du cas "emboîté" sur des maillages quelconques, même si ils ne sont pas conformes au bord.
- Les deux méthodes "semi-Schwarz" convergent au moins aussi vite que la méthode harmonique. Par contre, elles sont plus sensibles aux erreurs de quadratures dans les intégrales mixtes. Heureusement, ce petit défaut n'est pas trop important en pratique parce qu'il suffit de faire quelques itérations pour trouver la solution très proche de l'optimalité.

– La convergence très lente de la méthode Chimère est conditionnée par la minceur de la zone de recouvrement entre Λ et Ω_H. Noter néanmoins que en faisant cette zone plus large, c'est-à-dire en agrandissant le sous-domaine Λ, on aurait obtenu des problèmes sur Λ plus coûteux.

Pour étudier le comportement des erreurs par rapport au raffinement des maillages, on choisit en Fig. 4.5 la méthode des patchs harmonique 1.3, la plus robuste. On observe que l'ordre de convergence est optimal et que les erreurs ne dépendent pas beaucoup ni du positionnement des maillages (emboîtés – pas emboîtés) ni de la façon de calculer les intégrales mixtes.

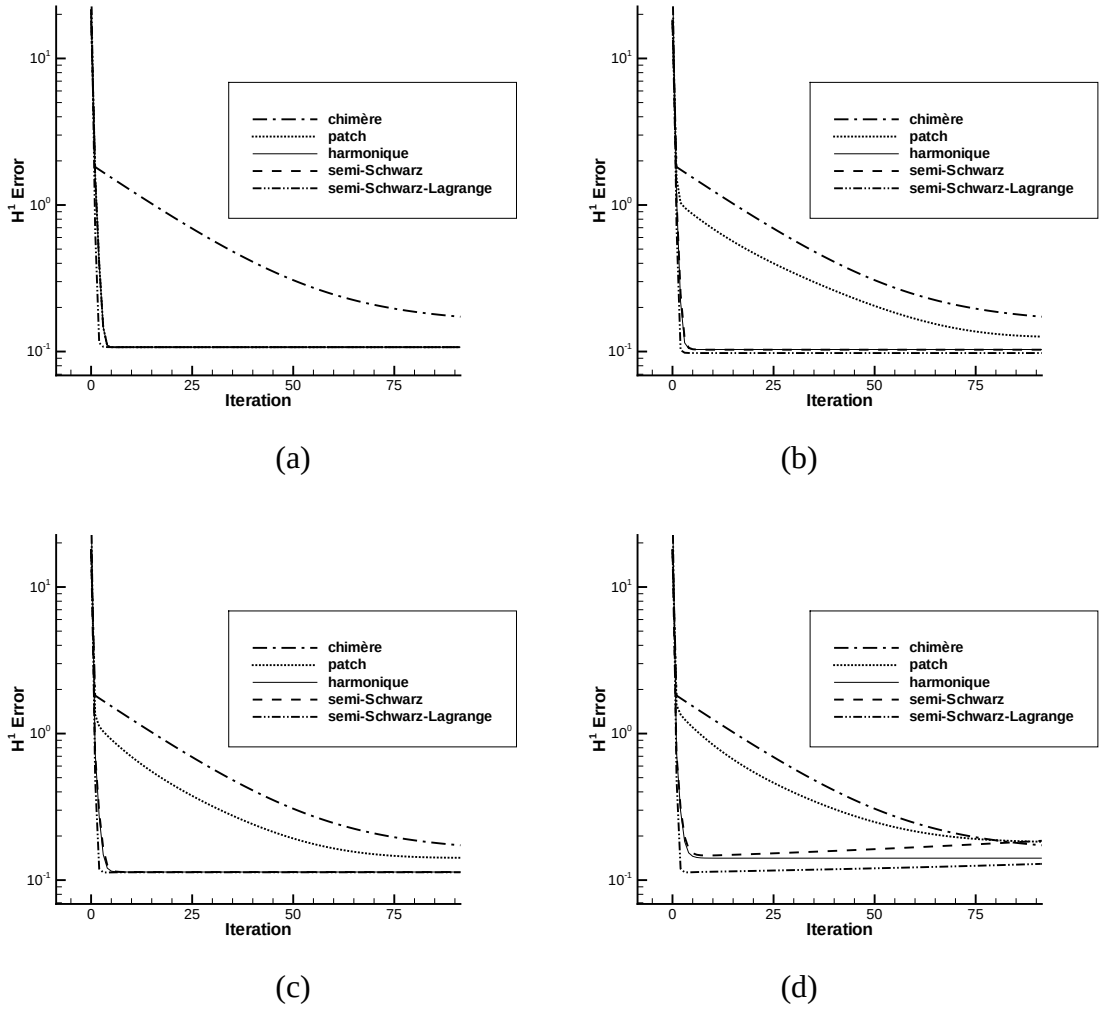


FIGURE 4.4: L'évolution de la norme $H^1(\Omega)$ de l'erreur sur les itérations. (a) – les maillages emboîtés de Fig. 4.1 ; (b) – les maillages conformes au bord de Fig. 4.2 ; (c) – les maillages non conformes au bord de Fig. 4.3 ; (d) – les mêmes maillages que dans le cas (c), mais en utilisant le calcul non exact des intégrales mixtes.

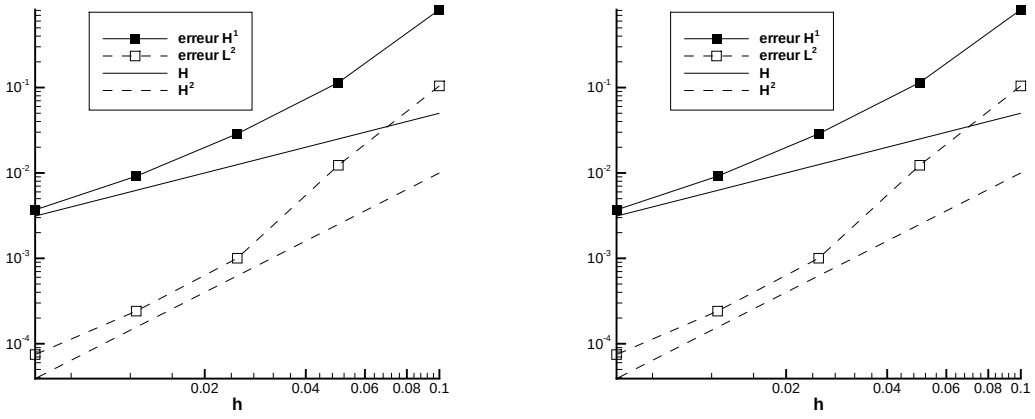


FIGURE 4.5: L'erreur par rapport au raffinement des maillages. A gauche : solutions limites obtenues par l'algorithme des patchs harmoniques sur des maillages emboîtés (situation I de la page 25 et les raffinements) ; à droite : solutions limites obtenues par le même algorithme avec le calcul non exact des intégrales mixtes sur des maillages non emboîtés, non conformes au bord (situation III de la page 25 et les raffinements).

5 Une variante de l'approche "Zoom Numérique" adaptée aux situations "multi-modèles"

Le travail en cours en collaborations avec Patrick Laborde

Toutes les méthodes décrites jusqu'à présent étaient principalement destinées aux situations qu'on peut qualifier de problèmes avec des données multi-échelles. On sous-entend par cela que le caractère multi-échelles de la solution exacte d'une EDP du type (1.1) est dû à la présence d'une forte variation du second membre f tandis que les coefficients à l'intérieur des opérateurs différentiels dans les équations sont assez régulières, ne présentant pas de variations brusques localisées. Une seule déviation de ce cadre jusqu'à ce point était l'application de la méthode des patchs aux problèmes avec des solutions singulières dues aux particularités de la frontière du domaine du calcul.

Cependant, on trouve très souvent en pratique des situations où le caractère multi-échelles apparaît dans les opérateurs différentiels eux-mêmes. L'exemple le plus simple est celui de l'équation (1.1) avec le coefficient K qui varie brusquement à l'intérieur de Λ, sous-domaine de Ω, en étant plus ou moins constant en dehors de Λ. On aimerait alors avoir une approche de Zoom Numérique qui résout au niveau grossier un problème qui ne voit pas l'échelle fine, c'est-à-dire un problème avec K plus ou moins constant partout. On peut aussi penser aux multi-échelles dans la géométrie de Ω. Par exemple, si Ω contient un trou ou une fissure d'une taille trop petite par rapport au pas du maillage qu'on envisage d'utiliser, on peut essayer de boucher le trou (ce qui donne le domaine sain $\tilde{\Omega} \supset \Omega$) et de mailler le domaine sain. La présence du trou sera alors prise en compte par le zoom $\Lambda \subset \Omega$ qui entoure le trou ou la fissure et où on construit un maillage fin et conforme à la géométrie du trou. On peut qualifier ces problèmes comme "multi-modèles" en prévoyant que leur étude est un premier pas vers des situations encore plus complexes où même les équations peuvent changer à l'intérieur du patch Λ. il pourra s'agir, par exemple, d'un couplage d'une description atomistique dans Λ avec une description macroscopique en dehors de Λ.

On ce concentre ici sur la situation la plus simple dans la classe des problèmes "multi-modèles" : l'équation (1.1) dans laquelle le coefficient K varie brusquement à l'intérieur de Λ. L'approche Chimère est applicable dans ce cas sans aucune modification. En effet, dans cette approche, on doit creuser un trou D dans Ω qui peut couvrir la zone de grandes variations de K. Le problème grossier

Entrées : fonctions $u_H^0 \in V_{gH}$ et $\lambda_h^0 \in M_h$.
pour $n = 1...N$ **faire**
Trouver $u_H^n \in V_{gH}$ tel que

$$\int_\Omega \tilde{K}\nabla u_H^n \cdot \nabla v_H = \int_{\Omega\setminus\Lambda} f v_H - \int_{\partial\Lambda} \lambda_h^{n-1} v_H + \int_\Lambda \tilde{K}\nabla u_H^{n-1}\cdot\nabla v_H, \ \ \forall v_H \in V_{0H}. \quad (5.1)$$

Trouver $w_h^n \in V_h$ et $\lambda_h^n \in M_h$ tels que

$$\begin{aligned} &\int_\Lambda K\nabla w_h^n \cdot \nabla v_h - \int_{\partial\Lambda} \lambda_h^n v_h = \int_\Lambda f v_h, \ \ \forall v_h \in V_{0h}, \quad (5.2)\\ &\int_{\partial\Lambda} w_h^n \mu_h = \int_{\partial\Lambda} u_H^n \mu_h, \quad \forall \mu_h \in M_h. \end{aligned}$$

fin

Algorithme 5.1: Semi-Schwarz-Lagrange, version pour “multi-modèles”

Entrées : fonctions $u_H^0 \in V_{gH}$ et $\lambda_h^0 \in M_h$.
pour $n = 1...N$ **faire**
Trouver $u_H^n \in V_{gH}$ tel que

$$\begin{aligned} \int_\Omega \tilde{K}\nabla u_H^n \cdot \nabla v_H = \rho \int_{\Omega\setminus\Lambda} f v_H - \rho \int_{\partial\Lambda} \lambda_h^{n-1} v_H + \rho \int_\Lambda \tilde{K}\nabla u_H^{n-1}\cdot\nabla v_H \\ + (1-\rho)\int_\Omega \tilde{K}\nabla u_H^{n-1}\cdot\nabla v_H \ \ \forall v_H \in V_{0H}. \end{aligned}$$

Trouver ($w_h^n \in V_h$, $\lambda_h^n \in M_h$), solution de (5.2).
fin

Algorithme 5.2: Semi-Schwarz-Lagrange, version relaxée pour “multi-modèles”

est alors posé sur $\Omega_H = \Omega \setminus D$ et peut être discrétisé sur un maillage de pas beaucoup plus grand que la taille caractéristique des variations de K dans D. Par contre, les algorithmes 1.2–1.5 semblent inapplicables au premier regard parce qu’ils maillent tout le domaine Ω de manière grossière. Il est intéressant de contourner ce problème parce que les algorithmes 1.2–1.5 ont montré leur efficacité par rapport à l’approche Chimère dans les situations "données multi-échelles" décrites précédemment. On peut heureusement adapter notre dernier algorithme 1.5 "Semi-Schwarz Lagrange" pour notre problème de classe "multi-modèles". L’idée de cette adaptation provient de l’observation que la seule occurrence de la partie multi-échelles de K au niveau grossier est dans les termes $\int_\Omega K\nabla u_H^n \cdot \nabla v_H$ et $\int_\Lambda K\nabla u_H^{n-1}\cdot\nabla v_H$. On peut alors remplacer dans ces termes le vrai coefficient K par une autre fonction $\tilde{K}$ telle que $\tilde{K} = K$ en dehors de Λ et qui est plus régulière que K dans Λ. Cela donne l’algorithme 5.1.

Etude de convergence des méthode itératives "multi-modèles" au niveau continu

Il convient d'étudier d'abord la convergence de l'analogue EDP de l'algorithme itératif 5.1 au niveau continu. Notons, qu'une telle étude était sans intérêt dans le cas des algorithmes 1.2–1.5 parce qu'ils convergent en une itération au niveau continu. Ce ne sera pas le cas pour l'algorithme 5.1. On verra même qu'il diverge dans certaines situations.

On rappelle les notations : Λ est un sous-domaine de Ω. Sans perte de généralité, on peut ne considérer que le cas $g = 0$. On introduit U^n comme l'analogue au niveau continu de u_H^n et w^n comme celui de w_h^n. L'analogue continue de l'algorithme 5.1 est alors : à partir des $U^0 \in H_0^1(\Omega)$ et $\lambda^0 \in H^{-1/2}(\Lambda)$ trouver successivement $U^n \in H_0^1(\Omega)$, $w^n \in H^1(\Lambda)$ et $\lambda^n \in H^{-1/2}(\Lambda)$ pour $n = 1, 2, \ldots$

$$\int_\Omega \tilde{K}\nabla U^n \cdot \nabla V = \int_{\Omega\setminus\Lambda} fV - \int_{\partial\Lambda} \lambda^{n-1} V + \int_\Lambda \tilde{K}\nabla U^{n-1}\cdot\nabla V, \qquad \forall V \in H_0^1(\Omega) \tag{5.3}$$

$$\int_\Lambda K\nabla w^n \cdot \nabla v - \int_{\partial\Lambda} \lambda^n v = \int_\Lambda fv, \qquad \forall v \in H^1(\Lambda) \tag{5.4}$$

$$\int_{\partial\Lambda} w^n \mu = \int_{\partial\Lambda} U^n \mu, \qquad \forall \mu \in H^{-1/2}(\partial\Lambda) \tag{5.5}$$

En posant

$$\hat{U}^n = \begin{cases} U^n \text{ dans } \Omega \setminus \Lambda \\ w^n \text{ dans } \quad \Lambda \end{cases} \tag{5.6}$$

on voit tout de suite que si la suite $\hat{U}^n$ converge dans $H_0^1(\Omega)$, alors la limite est égale à u, c'est-à-dire la solution de (1.1).

Dans le cas $K = \tilde{K}$, la méthode (5.3)–(5.5) est équivalente à la méthode des patchs sans relaxation (en posant $u^n = w^n - U^n|_\Lambda$, $u^n \in H_0^1(\Omega)$). Elle converge ainsi en une itération, c'est-à-dire $\hat{U}^1 = u$. Dans le cas général : $K \neq \tilde{K}$ la convergence de (5.3)-(5.5) n'est pas malheureusement garantie. On peut voir cela sur l'exemple

$$\tilde{K} = 1, \quad K = \begin{cases} \alpha, & x \in \bar{\Lambda} \\ 1, & x \in \Omega\setminus\Lambda \end{cases}$$

avec une constante α assez grande. En effet, en posant pour simplifier $f = 0$, $U^0 = 0$, $\lambda^0 \neq 0$, on montre aisément par récurrence $\Delta U^n = \Delta w^n = 0$ dans Λ pour tout n. Par conséquent, $w^n = U^n|_\Lambda$ pour tout n et l'algorithme peut être réécrit en termes de λ^n seulement :

$$\lambda^1 = -\alpha S_\Lambda \lambda^0 \qquad \text{et} \qquad \lambda^n = (1-\alpha) S_\Lambda \lambda^{n-1} \text{ pour } n \geq 2,$$

où $S_\Lambda : H^{-1/2}(\partial\Lambda) \to H^{-1/2}(\partial\Lambda)$ est l'opérateur défini pour toute fonction λ sur $\partial\Lambda$ par la relation $S_\Lambda\lambda = \frac{\partial u}{\partial \mathbf{n}}\big|_{\partial\Lambda}$ où $u \in H_0^1(\Omega)$ satisfait

$$\int_\Omega \nabla u \cdot \nabla v = \int_{\partial\Lambda} \lambda v, \ \forall v \in H_0^1(\Omega).$$

Cela implique $\lambda^n \to \infty$ lorsque $n \to \infty$ si la valeur de α est assez grande et si λ^0 est choisi, par exemple, comme une fonction propre de S_Λ.

Pour remédier au manque de convergence, on peut relaxer la méthode. Cela veut dire dans notre cas qu'on remplace U^n dans (5.3) par une combinaison convexe de l'expression non-relaxée et de

l'itération précédente. Soit ρ le paramètre de relaxation, alors la version relaxée de l'algorithme s'écrit

$$\int_\Omega \tilde{K}\nabla U^n \cdot \nabla V = \rho \int_{\Omega\setminus\Lambda} fV - \rho \int_{\partial\Lambda} \lambda^{n-1} V + \rho \int_\Lambda \tilde{K}\nabla U^{n-1}\cdot\nabla V \tag{5.7}$$
$$+ (1-\rho)\int_\Omega \tilde{K}\nabla U^{n-1}\cdot\nabla V, \qquad \forall V \in H_0^1(\Omega)$$
$$\int_\Lambda K\nabla w^n \cdot \nabla v - \int_{\partial\Lambda} \lambda^n v = \int_\Lambda fv, \qquad \forall v \in H^1(\Lambda) \tag{5.8}$$
$$\int_{\partial\Lambda} w^n \mu = \int_{\partial\Lambda} U^n \mu, \qquad \forall \mu \in H^{-1/2}(\partial\Lambda) \tag{5.9}$$

L'algorithme relaxé converge si le paramètre ρ est assez petit.

Proposition 5.1 *Il existe un nombre $\rho_0 > 0$ (dépendant des domaines Ω, Λ et des coefficients K, $\tilde{K}$) tel que l'algorithme (5.7)–(5.9) converge vers la solution exacte de (1.1) pour tout $\rho \in]0, \rho_0[$. Cela veut dire que $\hat{U}^n \to u$ dans $H^1(\Omega)$ lorsque $n \to \infty$, si on définit $\hat{U}^n$ par (5.6) à partir de U^n, w^n donnés par (5.7)–(5.9).*

Démonstration. On introduit l'opérateur $T : H_0^1(\Omega) \to H_0^1(\Omega)$ comme suit : pour tout $U \in H_0^1(\Omega)$ on trouve $w \in H^1(\Lambda)$ telle que

$$-\nabla\cdot(K\nabla w) = 0 \text{ dans } \Lambda,\ w|_{\partial\Lambda} = U|_{\partial\Lambda} \tag{5.10}$$

et on définit $TU \in H_0^1(\Omega)$ par

$$\int_\Omega \tilde{K}\nabla(TU)\cdot\nabla V = \int_{\Omega\setminus\Lambda} \tilde{K}\nabla U\cdot\nabla V + \int_\Lambda K\nabla w\cdot\nabla V,\ \forall V \in H_0^1(\Omega). \tag{5.11}$$

L'opérateur T est symétrique dans le produit scalaire $a(U,V) = \int_\Omega \tilde{K}\nabla U\cdot\nabla V$. En effet, on peut réécrire sa définition (5.11) de manière symétrique

$$\int_\Omega \tilde{K}\nabla(TU)\cdot\nabla V = \int_{\Omega\setminus\Lambda} \tilde{K}\nabla U\cdot\nabla V + \int_\Lambda K\nabla w\cdot\nabla\tilde{w},\ \forall V \in H_0^1(\Omega),$$

où $\tilde{w} \in H^1(\Lambda)$ est la même chose que w mais défini pour V au lieu de U, c'est-à-dire

$$-\nabla\cdot(K\nabla\tilde{w}) = 0 \text{ dans } \Lambda,\ \tilde{w}|_{\partial\Lambda} = V|_{\partial\Lambda}.$$

On décompose $H_0^1(\Omega)$ en somme directe $H_0^1(\Omega) = V_0 \oplus^\perp V_1$ où $V_0 = H_0^1(\Lambda)$ et V_1 est le complément orthogonal de V_0 au sens du produit scalaire a. Cela veut dire que les fonctions de V_1 satisfont $-\nabla\cdot(\tilde{K}\nabla U) = 0$ dans Λ. En étudiant l'action de T sur V_0 et V_1 séparément, on observe facilement les faits suivants :

- $TU = 0, \forall U \in V_0$.
- $TU \subset V_1, \forall U \in V_1$, c'est-à-dire T est bien défini comme un opérateur $V_1 \to V_1$.
- T est coercive et continu sur V_1. En effet, la continuité de T est évidente. Pour montrer sa coercivité on prend n'importe quelle $U \in V_1$ et on considère $w \in H^1(\Lambda)$ donné par (5.10). On a par intégration par parties

$$\int_\Lambda K\nabla w\cdot\nabla(U-w) = 0$$

Cela implique

$$\begin{aligned}\int_\Lambda K\nabla w\cdot\nabla U &= \int_\Lambda K|\nabla w|^2 \geq \alpha\int_\Lambda \tilde{K}|\nabla w|^2\\ &= \alpha\left(\int_\Lambda \tilde{K}|\nabla U|^2+\int_\Lambda \tilde{K}|\nabla(U-w)|^2\right)\geq \alpha\int_\Lambda \tilde{K}|\nabla U|^2\end{aligned}$$

avec $\alpha=\inf_{x\in\Lambda}\frac{K(x)}{\tilde{K}(x)}>0$.
En posant $V=U$ dans la définition de T on voit maintenant

$$a(TU,U)=a(T_1U,U)\geq\min(1,\alpha)a(U,U)$$

qui donne la coercivité de T sur V_1.
En notant $e^n=U^n-u$ où U^n est calculé par (5.7)–(5.9), on voit que e^n et e^{n-1} sont reliés par

$$e^n=e^{n-1}-\rho Te^{n-1}.$$

On décompose sur chaque itération e^n en $e^n=e_0^n+e_1^n$ avec $e_0^n\in V_0$ et $e_1^n\in V_1$. Les propriétés de T ci-dessus impliquent

$$\begin{aligned}e_0^n &= e_0^{n-1}\\ e_1^n &= e_1^{n-1}-\rho Te_1^{n-1}\end{aligned}$$

Comme l'opérateur $T:V_1\to V_1$ est auto-adjoint et coercif, on a $e_1^n\to 0$ lorsque $n\to\infty$ à condition que $0<\rho<2/||T||$. Cela implique $U^n\to u$ dans $H^1(\Omega\setminus\Lambda)$ et, en particulier, $U^n\to u$ sur $\partial\Lambda$ dans $H^{1/2}(\partial\Lambda)$. En combinant ce résultat avec (5.8)-(5.9), on voit que $w^n\to u$ dans $H^1(\Lambda)$. ■

La discrétisation par éléments finis de l'algorithme relaxé (5.7)–(5.9) est détaillée dans l'algorithme 5.2.

Une estimation d'erreur pour les algorithmes 5.1 et 5.2

On rappelle que les approximations de la solution exacte u sont obtenues à chaque itération de l'algorithme 5.1 ou 5.2 de la manière suivante :

$$u_{Hh}^n=\begin{cases}w_h^n, \text{ dans } \Lambda\\ u_H^n, \text{ en dehors de } \Lambda\end{cases}$$

La fonction u_{Hh}^n n'est pas continue en général. Par contre, elle appartient à l'espace combiné

$$\begin{aligned}V_{Hh} = \Big\{&v:\Omega\to\mathbb{R}, \text{ tel que } \exists v_H\in V_H^0, \exists v_h\in V_h\\ &\int_{\partial\Lambda}v_H\mu=\int_{\partial\Lambda}v_h\mu\quad\forall\mu\in M_h\\ &v=v_H \text{ sur } \Omega\setminus\Lambda \text{ et } v=v_h \text{ sur } \Lambda\Big\}\end{aligned}$$

dans lequel la continuité à travers $\partial\Lambda$ est imposée de manière faible.

Supposons que l'algorithme 5.1 (ou 5.2) converge, c'est-à-dire $u_H^n \to u_H$, $u_h^n \to u_h$, $\lambda_h^n \to \lambda_h$ lorsque $n \to \infty$. Alors $u_H \in V_H$, $u_h \in V_h$, $\lambda_h \in M_h$ satisfont

$$\begin{aligned}
&\int_{\Omega\setminus\Lambda} K\nabla u_H \cdot \nabla v_H = \int_{\Omega\setminus\Lambda} f v_H - \int_{\partial\Lambda} \lambda_h v_H, && \forall v_H \in V_{0H}\\
&\int_{\Lambda} K\nabla w_h \cdot \nabla v_h - \int_{\partial\Lambda} \lambda_h v_h = \int_{\Lambda} f v_h, && \forall v_h \in V_h\\
&\int_{\partial\Lambda} w_h \mu_h = \int_{\partial\Lambda} u_H \mu_h, && \forall \mu_h \in M_h
\end{aligned}$$

On peut combiner u_H et u_h dans

$$u_{Hh} = \begin{cases} w_h, \text{ dans } \Lambda \\ u_H, \text{ en dehors de } \Lambda \end{cases}$$

qui est l'approximation de Galerkin sur l'espace V_{Hh}, c'est-à-dire $u_{Hh} \in V_{Hh}$ satisfait

$$\int_{\Omega\setminus\partial\Lambda} K\nabla u_{Hh} \cdot \nabla v_{Hh} = \int_{\Omega} f v_{Hh}, \quad \forall v_{Hh} \in V_{Hh} \tag{5.12}$$

Cette formulation ressemble à la méthode "Mortar" sur des grilles non conformes, étudiée dans [33]. On y emprunte l'estimation *a priori* de cette méthode :

Théorème 5.2 *Soit $r = s = 1$ et supposons que la solution u de (1.1) appartient à $H^2(\Omega)$. Alors, on a l'estimation d'erreur* a priori *suivante pour la solution u_{Hh} de (5.12) :*

$$|u - u_{Hh}|_{H^1(\Omega)} \le C\left(H||u||_{H^q(\Omega\setminus\Lambda)} + h||u||_{H^q(\Lambda)}\right), \tag{5.13}$$

avec une constante C qui ne dépend pas de pas des maillages $\mathcal{T}_H$ et $\mathcal{T}_h$.

Il serait préférable de préciser la dépendance de la constante C sur les coefficient K et $\tilde{K}$. Cela fait l'objet d'un travail en cours.

Un test numérique dans un domaine "fissuré"

On finit ce chapitre par une application de l'algorithme 5.1 dans une situation où le domaine de calcul Ω contient une "fissure ouverte", voir Fig. 5.1. On prend notamment $\tilde{\Omega}$ comme le disque unité et $\Omega = \tilde{\Omega} \setminus \mathcal{F}$ avec

$$\mathcal{F} = \{(x, y) \in \mathbb{R}^2 : x \le -0.5 \text{ et } |y| \le |0.5 + x| \tan(0.01)\}.$$

Le but est de résoudre numériquement le problème

$$-\Delta u = 1 \text{ dans } \Omega, \quad u = 0 \text{ sur } \partial\Omega,$$

en utilisant le maillage grossier qui ne voit pas la fissure $\mathcal{F}$. On construit donc $\mathcal{T}_H$ sur le disque entier $\tilde{\Omega}$, on prend le zoom Λ dont la géométrie est conforme avec la fissure (voir Fig. 5.1) et on y construit la triangulation $\mathcal{T}_h$. L'algorithme 5.1 est modifié comme suit :

- On met $K = \tilde{K} = 1$ dans (5.1) et (5.2).
- On remplace Ω par $\tilde{\Omega}$ dans (5.1).

La convergence de cet algorithme est illustrée en Fig. 5.2. On y compare les itérés de l'algorithme avec la solution "exacte" de référence obtenue sur une maillage très fin conforme à la géométrie de Ω. On voit que l'algorithme donne une très bonne approximation dès la 4-ième itération.

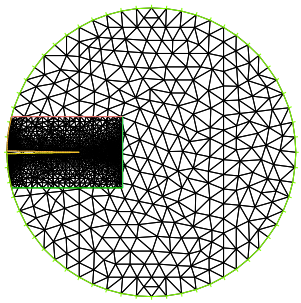
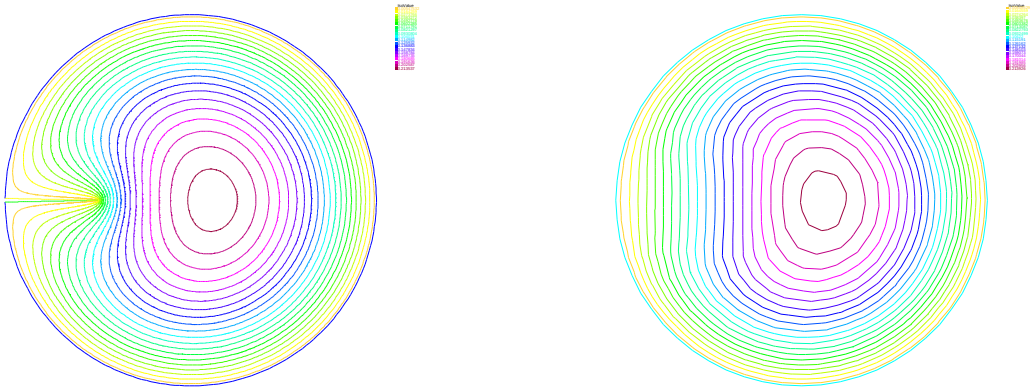
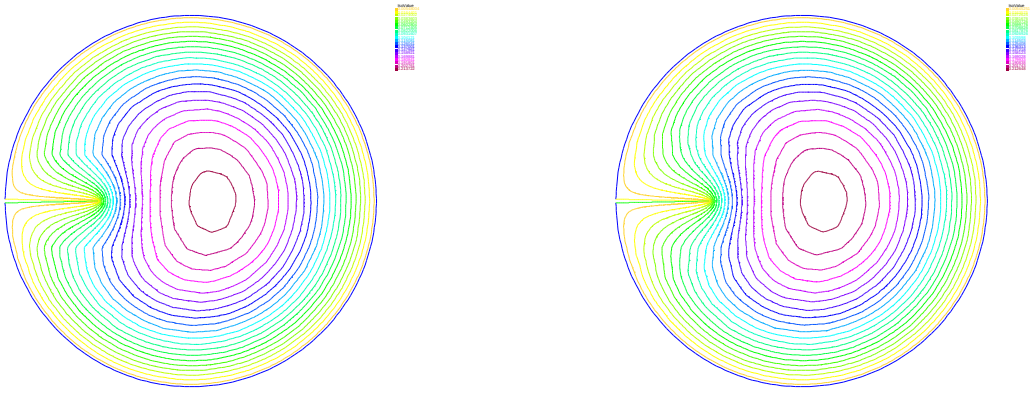


FIGURE 5.1: Le domaine Ω, la triangulation grossière $\mathcal{T}_H$, le patch Λ et la triangulation fine $\mathcal{T}_h$.



la solution exacte — la solution grossière



l'itération 1 — l'itération 4

FIGURE 5.2: La solution exacte du problème dans le domaine Ω dessiné en Fig. 5.1 ; la solution grossière qui sert à initialiser l'algorithme ; les solutions approchées sur les itérations 1 et 4.

6 Les méthodes de type "éléments finis multi-échelles - MsFEM "

Le travail en cours en collaboration avec plusieurs collègues de l'IMT

Beaucoup de problèmes fondamentaux et importants en pratique possèdent des solutions à échelles multiples. Les matériaux composites, les milieux poreux, le transport de polluant en milieu urbain sont des exemples de ce type. A la différence des problèmes étudiés plus haut, le comportement multi-échelles n'est plus localisé dans une petite région, mais il est présent partout dans le domaine du calcul. L'utilisation des idées de "Zoom Numérique" aboutirait donc aux calculs très lourds. Cependant, il peut être nécessaire de réaliser des simulations de tels phénomènes en temps réel. Par exemple, une prédiction du transport de polluant en milieu urbain est un outil d'aide à la décision pour les autorités lors de situations extrêmes. Il serait donc souhaitable d'avoir une méthode qui prenne en compte les effets des petites échelles sans nécessiter pour autant la résolution de tous les détails des petites échelles.

Dans ce but, T.Hou a développé la méthode des éléments finis multi-échelles (MsFEM), voir [44]. La particularité de cette technique est de travailler sur des mailles sous-résolues, c'est-à-dire beaucoup plus grandes que les tailles des échelles les plus petites. Les structures à l'échelle microscopique sont prises en compte par le biais de solutions locales représentatives (les fonctions de base multi-échelles), dans un travail préparatoire (coûteux en temps de calcul) qui peut être réalisé une fois pour toutes. Typiquement, ces fonctions sont calculées comme les solutions locales des problèmes homogènes. Toute l'enjeu de la méthode est alors de bien choisir les conditions aux limites pour ces problèmes.

Dans ce chapitre on va présenter quelques variantes de cette méthode sur l'exemple de l'équation de diffusion-réaction. On considéra d'abord le cas 1D où on démontrera une estimation d'erreur optimale. On tournera après au cas 2D où la théorie reste incomplète. On proposera une nouvelle variante de la méthode MsFEM qui s'avère en pratique plus robuste que les autres variantes proposées dans la littérature.

Le cas 1D

Considérons d'abord le cas 1D : on pose $\Omega = (0, L)$ et on cherche $u : \Omega \to \mathbb{R}$ telle que

$$\begin{cases} Lu = -(\nu u')' + \sigma u = f \\ u(0) = u(L) = 0 \end{cases} \tag{6.1}$$

avec $f \in L^2(\Omega)$ donnée. On suppose que les coefficients ν et σ sont des fonctions L^∞ satisfaisant $\nu(x) \geq \nu_{\min} > 0$ et $\sigma(x) \geq \sigma_{\min} \geq 0$ pour tout $x \in \Omega$. En introduisant la forme bilinéaire sur $H_0^1(\Omega)$

$$a(u, v) = \int_0^L \nu u' v' dx + \int_0^L \sigma uv dx, \tag{6.2}$$

on écrit la formulation faible de (6.1) suivante : trouver $u \in H_0^1(\Omega)$ tel que

$$a(u, v) = \int_0^L fv dx, \quad \forall v \in H_0^1(\Omega). \tag{6.3}$$

On introduit aussi un maillage sur Ω en divisant Ω en N segments $K_i = [x_{i-1}, x_i]$, $i = 1, \ldots, N$ par les nœuds $0 = x_0 < x_1 < \cdots < x_N = L$. On définit le pas de maillage H par $\max |x_i - x_{i-1}|$. Considérons l'espace d'éléments finis multi-échelles adapté à l'opérateur L

$$V_H = \{u_H \in C_0(\Omega) \text{ telle que } Lu_H = 0 \text{ sur } K_i, i = 1, \ldots, N\}. \tag{6.4}$$

L'approximation standard de Galerkin $u_H \in V_H$ de la solution de (6.1) peut alors être introduite comme

$$a(u_H, v_H) = \int_0^L f v_H dx, \quad \forall v_H \in V_H. \tag{6.5}$$

Proposition 6.1 *La solution de Galerkin u_H de (6.5) satisfait*

$$||u - u_H||_E \leq \frac{H}{\pi \sqrt{\nu_{\min}}} ||f||_{L^2(\Omega)}$$

où $|| \cdot ||_E$ est la norme d'énergie sur $H_0^1(\Omega)$: $|| \cdot ||_E^2 = a(\cdot, \cdot)$.

Démonstration. D'après le lemme de Céa

$$||u - u_H||_E = \inf_{v_H \in V_H} ||u - v_H||_E \tag{6.6}$$

On prend $v_H \in V_H$ comme l'interpolant éléments finis de u, c'est-à-dire

$$v_H(x_i) = u(x_i), \quad i = 0, 1, \ldots, N$$

et on considère l'erreur d'interpolation $e = u - v_H$. On a sur chaque maille

$$\begin{cases} -(\nu e')' + \sigma e = f \text{ on } (x_{i-1}, x_i) \\ e(x_{i-1}) = e(x_i) = 0 \end{cases}$$

Multiplication par e et intégration sur (x_{i-1}, x_i) donnent

$$\int_{x_{i-1}}^{x_i} \nu |e'|^2 dx + \int_{x_{i-1}}^{x_i} \sigma e^2 dx = \int_{x_{i-1}}^{x_i} f e dx \leq ||f||_{L^2(K_i)} ||e||_{L^2(K_i)} \tag{6.7}$$

Comme la fonction e s'annule sur la frontière de K_i, on peut appliquer l'inégalité de Poincaré avec la meilleure constante possible h_i/π. Ainsi

$$||e||_{L^2(K_i)} \leq \frac{h_i}{\pi}\left(\int_{x_{i-1}}^{x_i}|e'|^2dx\right)^{\frac{1}{2}} \leq \frac{h_i}{\pi\sqrt{\nu_{\min}}}\left(\int_{x_{i-1}}^{x_i}\nu|e'|^2dx+\int_{x_{i-1}}^{x_i}\sigma e^2dx\right)^{\frac{1}{2}} \tag{6.8}$$

En substituant dans (6.7) on voit que

$$\int_{x_{i-1}}^{x_i}\nu|e'|^2dx+\int_{x_{i-1}}^{x_i}\sigma e^2dx \leq \frac{h_i^2}{\pi^2\nu_{\min}}||f||_{L^2(K_i)}$$

La sommation sur tous les segments K_i donne le résultat désiré. ■

Remarque 6.2 *L'intérêt de la proposition 6.1 réside dans la constante devant H dans l'estimation d'erreur. En effet, pour la méthode d'éléments finis P_1 standards, l'erreur est aussi proportionnelle à H, mais la constante de la proportionnalité dépend de la norme H^2 de la solution exacte. Par conséquent, elle peut être énorme si les coefficients ν et σ de l'opérateur différentiel oscillent à une petite échelle. La méthode MsFEM permet de retrouver l'erreur, toujours proportionnelle à H, mais avec la constante qui ne dépend pas des oscillation des coefficients dans l'équation qu'on résout.*

Remarque 6.3 *Si $\sigma_{\min} > 0$ l'estimation d'erreur peut être améliorée :*

$$||u-u_H||_E \leq \frac{H}{\sqrt{\pi^2\nu_{\min}+H^2\sigma_{\min}}}||f||_{L^2(\Omega)}$$

en utilisant une majoration plus précise dans (6.8).

Remarque 6.4 *Le résultat de la proposition 6.1 peut être étendu aussi au cas d'éléments finis d'ordre plus élevé. Soit $k \geq 2$ et*

$$V_H^k = \{u_H \in C_0(\Omega) \text{ tel que } Lu_H \in \mathbb{P}_{k-2} \text{ sur } K_i,\ i = 1,\ldots,N\}.$$

On considère l'approximation de Galerkin (6.5) avec V_H remplacé par $V_H^{(k)}$. On a alors la majoration d'erreur suivante pour la solution $u_H \in V_H^{(k)}$

$$||u-u_H||_E \leq c_k\frac{H^k}{\pi\sqrt{\nu_{\min}}}|f|_{H^{k-1}(\Omega)} \tag{6.9}$$

avec $c_k = 1/(2^{k-1}(k-2)!\sqrt{(2k-3)(2k-2)})$.

En effet, il suffit de construire un interpolant $v_H \in V_H^{(k)}$ qui satisfait (6.9). Pour cela, on prend f_H comme une approximation de f qui est un polynôme d'ordre $k-2$ sur chaque segment $K_i = (x_{i-1}, x_i)$. On peut construire une telle f_h satisfaisant

$$||f-f_H||_{L^2(K_i)} \leq c_k h_i^{k-1}|f|_{H^{k-1}(K_i)}. \tag{6.10}$$

On prend $v_H \in V_H^{(k)}$ comme un interpolant de u tel que $v_H(x_i) = u(x_i)$, $i = 0,1,\ldots,N$ et $Lv_H = f_H$ sur K_i, $i = 1,\ldots,N$. On considère l'erreur d'interpolation $e = u - v_H$. On a sur chaque maille

$$\begin{cases} -(\nu e')' + \sigma e = f - f_H \text{ on } (x_{i-1}, x_i) \\ e(x_{i-1}) = e(x_i) = 0 \end{cases}$$

Comme dans la preuve de la proposition 6.1, on voit que

$$\int_{x_{i-1}}^{x_i}\nu|e'|^2dx+\int_{x_{i-1}}^{x_i}\sigma e^2dx = \int_{x_{i-1}}^{x_i}(f-f_H)edx \leq ||f-f_H||_{L^2(K_i)}||e||_{L^2(K_i)}$$

Cela implique (6.9) en vue de (6.10) et (6.8).

L'implémentation de la méthode multi-échelles dans le cas 1D

La méthode présentée ci-dessus n'offre qu'un intérêt théorique puisque la construction d'espace V_H s'appuie sur des solutions exactes des EDPs sur chaque intervalle K_i. Pour mettre cette méthode en marche en pratique, on doit encore remplacer ces solutions exactes par leurs approximations éléments finis sur une grille fine de pas $h << H$. Pour simplifier, on suppose que $x_i = iH$, $i = 0, \ldots, n$. On introduit ainsi les nœuds $\xi_i = ih$, $i = 0, \ldots, N$ où $h = L/N$ et $N >> n$ est divisible par n, de sorte que $H/h = N/n$ soit le nombre d'éléments fins dans chaque élément grossier. On définit

$$V_h = \{u_h \in C_0(\Omega) \text{ tel que } u_h \in P_1(\xi_{i-1}, \xi_i),\, i = 1, \ldots, N\}.$$

L'espace d'éléments finis multi-échelles sera un sous-espace de V_h introduit comme

$$V_{Hh} = \{u_h \in V_h \text{ tel que } a(u_h, \phi_h) = 0\ \forall \phi_h \in V_h \text{ tel que supp}\phi_h \subset K_i \text{ pour certain } i \in [1, \ldots, n]\}.$$

On considère toujours l'approximation standard de Galerkin : trouver $u_{Hh} \in X_0$ tel que

$$a(u_{Hh}, \phi_{Hh}) = \int_0^L f\phi_{Hh}dx, \quad \forall \phi_{Hh} \in V_{Hh}. \tag{6.11}$$

Théorème 6.5 *La solution de Galerkin u_{Hh} dans (6.11) satisfait*

$$||u - u_{Hh}||_E \leq \frac{H}{\pi\sqrt{\nu_{\min}}}||f||_{L^2(\Omega)} + \inf_{v_h \in V_h} ||u - v_h||_E.$$

Démonstration. Soit V l'espace fonctionnel H_0^1 muni du produit scalaire $a(\cdot, \cdot)$. On décompose V en une somme directe $V = V^0 \oplus^\perp V_H$ où V^0 contient toutes les fonctions de V qui s'annulent en tous les points x_i, $i = 1, \ldots, n$ et où V_H est son complément orthogonal. On voit immédiatement que V_H est l'espace d'éléments finis multi-échelles "théorique" introduit plus haut dans (6.4). On reproduit la même décomposition au niveau discret : $V_h = V_h^0 \oplus^\perp V_{Hh}$ où $V_h^0 = V_h \cap V^0$ et V_{Hh} est le complément orthogonal de V_h^0 dans V_h. Soit $u_h \in V_h$ la solution éléments finis sur le maillage fin

$$a(u_h, \phi_h) = \int_0^L f\phi_h dx, \quad \forall \phi_h \in V_h.$$

Considérons les décompositions uniques de u et de u_h :

$$\begin{aligned} u &= u^0 + u_H \text{ avec } u^0 \in V^0 \text{ et } u_H \in V_H, \\ u_h &= u_h^0 + u_{Hh} \text{ avec } u_h^0 \in V_h^0 \text{ et } u_{Hh} \in V_{Hh}, \end{aligned}$$

où u_H et u_{Hh} sont définis par (6.5) et (6.11) respectivement. On a

$$||u_h^0||_E^2 = a(u_h, u_h^0) = \int_0^L f u_h^0 dx = a(u, u_h^0) = a(u^0, u_h^0) \leq ||u^0||_E||\ u_h^0||_E,$$

puisque V_h^0 est orthogonal à V_H et à V_{Hh}. Par conséquent, $||u_h^0||_E \leq ||u^0||_E$ et

$$||u - u_{Hh}||_E \leq ||u - u_h||_E + ||u_h^0||_E \leq ||u - u_h||_E + ||u^0||_E = ||u - u_h||_E + ||u - u_H||_E.$$

On conclut en appliquant le lemme de Céa et la proposition 6.1. ■

MsFEM en 2D : des variantes conformes et non conformes

Soit $\Omega \subset \mathbb{R}^2$ un domaine borné polygonal. On considère le problème aux limites : trouver $u : \Omega \to \mathbb{R}$ tel que

$$\begin{cases} Lu = -\nabla \cdot (\nu \nabla u) + \sigma u = f \text{ dans } \Omega \\ u = 0 \text{ sur } \partial\Omega \end{cases} \tag{6.12}$$

avec $f \in L^2(\Omega)$ donné. On suppose que les coefficients ν et σ sont des fonctions L^∞ satisfaisant $\nu(x) \geq \nu_{\min} > 0$ et $\sigma(x) \geq \sigma_{\min} \geq 0$ pour tout $x \in \Omega$. En introduisant la forme bilinéaire sur $H_0^1(\Omega)$

$$a(u, v) = \int_\Omega \nu \nabla u \cdot \nabla v dx + \int_\Omega \sigma u v dx, \tag{6.13}$$

la formulation faible du (6.12) s'écrit : trouver $u \in H_0^1(\Omega)$ tel que

$$a(u, v) = \int_\Omega f v dx, \quad \forall v \in H_0^1(\Omega). \tag{6.14}$$

On notera $|| \cdot ||_E$ la norme d'énergie sur $H_0^1(\Omega)$ définie comme $|| \cdot ||_E^2 = a(\cdot, \cdot)$.

Soit $\mathcal{T}_H$ un maillage sur Ω. Pour la simplicité, on se restreint ici au cas où Ω est le carré unité et $\mathcal{T}_H$ est une grille cartésienne rectangulaire sur Ω. On notera $\mathcal{E}_H$ l'ensemble de tous les arêtes de rectangles de $\mathcal{T}_H$. Le pas du maillage est $H = \max_{K \in \mathcal{T}_H} diam(K)$. Comme on a vu dans le cas 1D, l'idée de la méthode MsFEM consiste à introduire un espace des éléments finis multi-échelles adapté à l'opérateur L. Au niveau théorique, les fonctions de cet espaces satisferont toujours l'équation $Lu_H = 0$ dans chaque maille K du maillage $\mathcal{T}_H$. Au niveau pratique, il faut discrétiser ces EDPs sur un maillage fin $\mathcal{T}_h$. On va ici supposer que $\mathcal{T}_h$ est une subdivision de $\mathcal{T}_H$. On notera W_h l'espace standard d'éléments fini Q_1 sur $\mathcal{T}_h$ et on posera $V_h = W_h \cap H_0^1(\Omega)$.

Tout l'enjeu de MsFEM, dans le cas 2D, consiste en un choix convenable des conditions aux limites sur les arêtes du maillage $\mathcal{T}_H$ pour les fonctions dans l'espace adapté V_H. Dans ce qui suit, on va présenter quelques options possibles. Certains de nos choix amènent à des espaces non conformes, c'est-à-dire, $V_H \not\subset H^1(\Omega)$ généralement et, par conséquent, $V_H \not\subset V_h$. Concrètement, les fonctions de V_H seront parfois discontinues à travers les arêtes de $\mathcal{T}_H$. On est obligé donc d'introduire une forme bilinéaire brisée

$$a_H(u, v) = \sum_{K \in \mathcal{T}_H} \left[\int_K \nu \nabla u \cdot \nabla v dx + \int_K \sigma u v dx \right] \tag{6.15}$$

L'approximation Galerkin $u_H \in V_H$ de la solution du problème (6.12) sera définie comme

$$a_H(u_H, v_H) = \int_\Omega f v_H dx, \quad \forall v_H \in V_H. \tag{6.16}$$

1. MsFEM classique [44] On construit V_H de manière suivante

$$\begin{aligned} V_H \;=\; \{u_h \in V_h : \quad & a(u_h, v_h) = 0 \text{ pour tout } v_h \in V_h \text{ tel que supp}(v_h) \subset K \text{ avec } K \in \mathcal{T}_H \\ & \text{et } u_h \text{ est linéaire sur tous les arêtes de} \mathcal{T}_H.\} \end{aligned}$$

Cela donne un espace H^1-conforme : $V_H \subset V_h \subset H_0^1(\Omega)$. L'inconvénient de cet approche est que l'approximation linéaire sur les arêtes du maillage n'est pas vraiment adaptée à l'opérateur L, ce qui dégrade la qualité de la solution MsFEM.

2. MsFEM avec l'oversampling [44] On agrandit d'abord les mailles : pour tout $K \in \mathcal{T}_H$, on choisit un rectangle plus grand $O(K) \supset K$ en prenant $O(K)$ comme un ensemble de mailles de $\mathcal{T}_h$. L'espace d'éléments finis V_H est défini par

$$\begin{aligned} V_H &= \{u \in L^2(\Omega) : u \text{ est continue dans les nœuds de } \mathcal{T}_h, \\ &\quad \text{pour tout } K \in \mathcal{T}_h, \text{ il existe } u_h \in V_h \text{ tel que} \\ &\quad u_h \text{ est linéaire sur toutes les arêtes de } \partial O(K), \\ &\quad a(u_h, v_h) = 0 \text{ pour tout } v_h \in V_h \text{ tel que supp}(v_h) \subset O(K), \\ &\quad \text{et } u|_K = u_h|_K\} \end{aligned}$$

Puisque $u_H \in V_H$ est continue seulement aux nœuds du maillage, on observe que $V_H \not\subset H^1(\Omega)$ et par conséquent $V_H \not\subset V_h$.

3. MsFEM non conforme Bien que la qualité de la solution MsFEM avec l'oversampling soit en général meilleur que celle sans oversampling, la deuxième méthode n'est pas complètement satisfaisant, au moins du point de vue théorique. En effet, la notion de continuité aux nœuds n'a même pas de sens pour les fonctions de $H^2(\Omega)$ qui est l'espace fonctionnel naturel de notre problème. C'est pour cette raison qu'il est intéressant d'introduire encore un autre approche pour la construction de l'espace des éléments finis non conformes (à La Crouzeix-Raviart) dans lequel la continuité est préservée au sens des valeurs moyennes sur les arêtes du maillage. Ainsi, on fera de sorte que si E est une arête interne du maillage partagée par deux mailles K_1 et K_2, alors toutes les fonctions u_H dans le nouvel espace satisferont

$$\int_E u_H|_{K_1} = \int_E u_H|_{K_2}.$$

A cet effet, on introduit d'abord un espace auxiliaire où on autorise les sauts sur les arêtes internes du maillage :

$$\begin{aligned} W_{hH} &= \{u \in L^2(\Omega) : \int_E [u] = 0 \text{ pour tout } E \in \mathcal{E}_H, \\ &\quad \text{pour tout } K \in \mathcal{T}_H \text{ il existe } u_h \in W_h \text{ avec } u|_K = u_h|_K\}. \end{aligned}$$

Ici, $[u]$ désigne le saut sur une arête si cette arête est interne, sinon $[u] = u$ sur une arête du bord. Soit

$$W^0_{hH} = \{u \in W_{hH} : \int_E u = 0 \text{ pour tout } E \in \mathcal{E}_H\}.$$

On introduit l'espace MsFEM non conforme

$$V_H = \{u_h \in W_{hH} : a_H(u_h, v_h) = 0 \text{ pour tout } v_h \in W^0_{hH}\}.$$

En d'autres termes, toute fonction $u_h \in V_H$ doit satisfaire sur n'importe quelle maille $K \in \mathcal{T}_H$

$$\int_K \nu \nabla u_h \cdot \nabla v_h dx + \int_K \sigma u_h v_h dx = 0, \quad \forall v_h \in W_h \text{ tel que } \int_{\partial \Lambda_i} v_h = 0, \;\; i = 1, \ldots, 4,$$

où $\partial\Lambda_i$, $i = 1, \ldots, 4$ sont les quatre arêtes qui forment le bord de K. Donc il existent des nombres $\lambda_1, \ldots, \lambda_4$ tels que

$$\int_K \nu \nabla u_h \cdot \nabla v_h dx + \int_K \sigma u_h v_h dx = \sum_{i=1}^{4} \lambda_i \int_{\partial\Lambda_i} v_h, \quad \forall v_h \in W_h.$$

On voit que la restriction de u_h à K soit une approximation éléments finis sur $\mathcal{T}_h$ de la solution du problème aux valeurs limites

$$Lu = 0 \text{ on } K, \quad \nu \frac{\partial u}{\partial n}\bigg|_{\partial\Lambda_i} = \lambda_i, \ i = 1, \ldots, 4.$$

Dans le cas le plus simple du laplacien ($L = -\Delta$), on retrouve les éléments finis non conformes classiques de Rannacher-Turek sur des rectangles : $u_h|_K \in \text{span}\{1, x, y, x^2 - y^2\}$ sur chaque maille $K \in \mathcal{T}_H$. Si on travaillait sur des maillages triangulaires, on retrouverait les éléments finis finis de Crouzeix-Raviart : $u_h|_K \in \text{span}\{1, x, y\}$.

Décrivons la construction d'une base pour V_H. Chaque fonction de base ϕ_E peut être associée à une arête interne $E \in \mathcal{E}_H$. La fonction ϕ_E est définie de sorte que son intégrale sur E soit égale à 1, tandis que les intégrales sur toutes les autres arêtes sont 0. Le support d'une telle fonction est dans $\overline{K^1 \cap K^2}$ où K^1 et K^2 sont deux mailles partageant l'arête E. Notons Γ_i^1, $i = 1, \ldots, 4$ les arêtes qui composent le bord ∂K^1 et Γ_i^2, $i = 1, \ldots, 4$ celles pour ∂K^2, en numérotant de telle sorte que $\Gamma_1^1 = \Gamma_1^2 = E$. Soit W_h^1 (resp W_h^2) l'espace de restrictions de fonctions de W_h sur K^1 (resp sur K^2). Soit $u_h^1 \in W_h^1$, $u_h^2 \in W_h^2$ les solutions des problèmes suivants

$$\begin{aligned}
\int_K \nu \nabla u_h^1 \cdot \nabla v_h^1 dx + \int_K \sigma u_h^1 v_h^1 dx &= \sum_{i=1}^{4} \lambda_i^1 \int_{\Gamma_i^1} v_h^1, \quad \forall v_h^1 \in W_h^1 \\
\int_{\Gamma_i^1} u_h^1 &= \delta_{i1}, \quad i = 1, \ldots, 4 \\
\int_K \nu \nabla u_h^2 \cdot \nabla v_h^2 dx + \int_K \sigma u_h^2 v_h^2 dx &= \sum_{i=1}^{4} \lambda_i^2 \int_{\Gamma_i^2} v_h^2, \quad \forall v_h^2 \in W_h^2 \\
\int_{\Gamma_i^2} u_h^2 &= \delta_{i1}, \quad i = 1, \ldots, 4
\end{aligned}$$

On voit que la fonction de base ϕ_E peut être construite comme

$$\phi_E = \begin{cases} u_h^1 \text{ sur } K^1, \\ u_h^2 \text{ sur } K^2, \\ 0 \text{ ailleurs.} \end{cases}$$

Les fonctions de base de l'espace MsFEM non conforme ont donc les mêmes supports localisés que les fonctions éléments finis non conformes habituelles. En pratique, la construction de chaque fonction de base requiert des calculs lourds sur les mailles autour de chaque arête.

Remarque 6.6 *Les fonctions de base des espaces MsFEM dans les premières deux variantes sont caractérisés aussi par leurs supports localisés. notamment, on associe les fonctions de base aux nœuds du maillage $\mathcal{T}_H$ comme pour les éléments finis conformes habituels. Les détails de construction des fonctions de base dans MsFEM classique et MsFEM-oversamling peuvent être trouvés dans [44].*

Des résultats numériques

On considère le cas test suivant : le problème (6.12) sur $\Omega = (0,1)^2$ avec les coefficients

$$\begin{aligned} \nu &= 1 + 100\, \mathbb{1}_{\cos(50x)>0} \mathbb{1}_{\sin(50y)>0} \\ \sigma &= 0, \end{aligned}$$

où $\mathbb{1}$ désigne la fonction indicatrice. Le maillage $\mathcal{T}_H$ est rectangulaire cartésien composé de $n \times n$ carrés de taille $H \times H$, $H = 1/n$. Le maillage $\mathcal{T}_h$ est une partition de Ω en triangles obtenue en divisant Ω en $N \times N$ carrés de taille $h \times h$, $h = 1/N$ et en coupant chaque carré ainsi obtenu en deux triangles. Dans MsFEM avec l'oversampling, on choisit les mailles élargies $O(K)$ 1.5 fois plus large que K et on les construit par homothétie autour du centre de K.

On présente en Fig. 6.1 le comportement de l'erreur par rapport au raffinement. Les erreurs sont calculées en prenant la solution standard éléments finis P_1 sur le maillage 512×512 comme solution de référence. On observe que les variantes oversampling et non conforme sont toutes les deux beaucoup plus précises que la variante classique. L'avantage de l'approche non conforme sur l'oversampling est illustré par le fait que ses courbe de convergences ont les mêmes pentes que celles d'une méthode éléments finis Q_1 dans le cas bien résolu. Ce n'est pas le cas pour la MsFEM-oversampling. On voit aussi que la qualité de cette dernière méthode se dégrade quand le pas de maillage devient proche de la taille caractéristique ε de la petite échelle, c'est-à-dire la période des oscillations de ν. Cette observation est consistante avec l'analyse théorique de l'erreur dans [44, Theorem 6.12]. En effet, la majoration d'erreur contient un terme proportionnel à ε/H. Cette particularité de la variante avec l'oversampling (visiblement absente de la méthode non conforme) peut avoir des conséquences négatives dans les applications pratiques où la séparation des échelles n'est pas toujours bien établie. On pense donc que la variante non conforme est un bon point de départ pour le développement des méthodes robustes pour des problèmes du type propagation des polluants dans un milieu urbain.

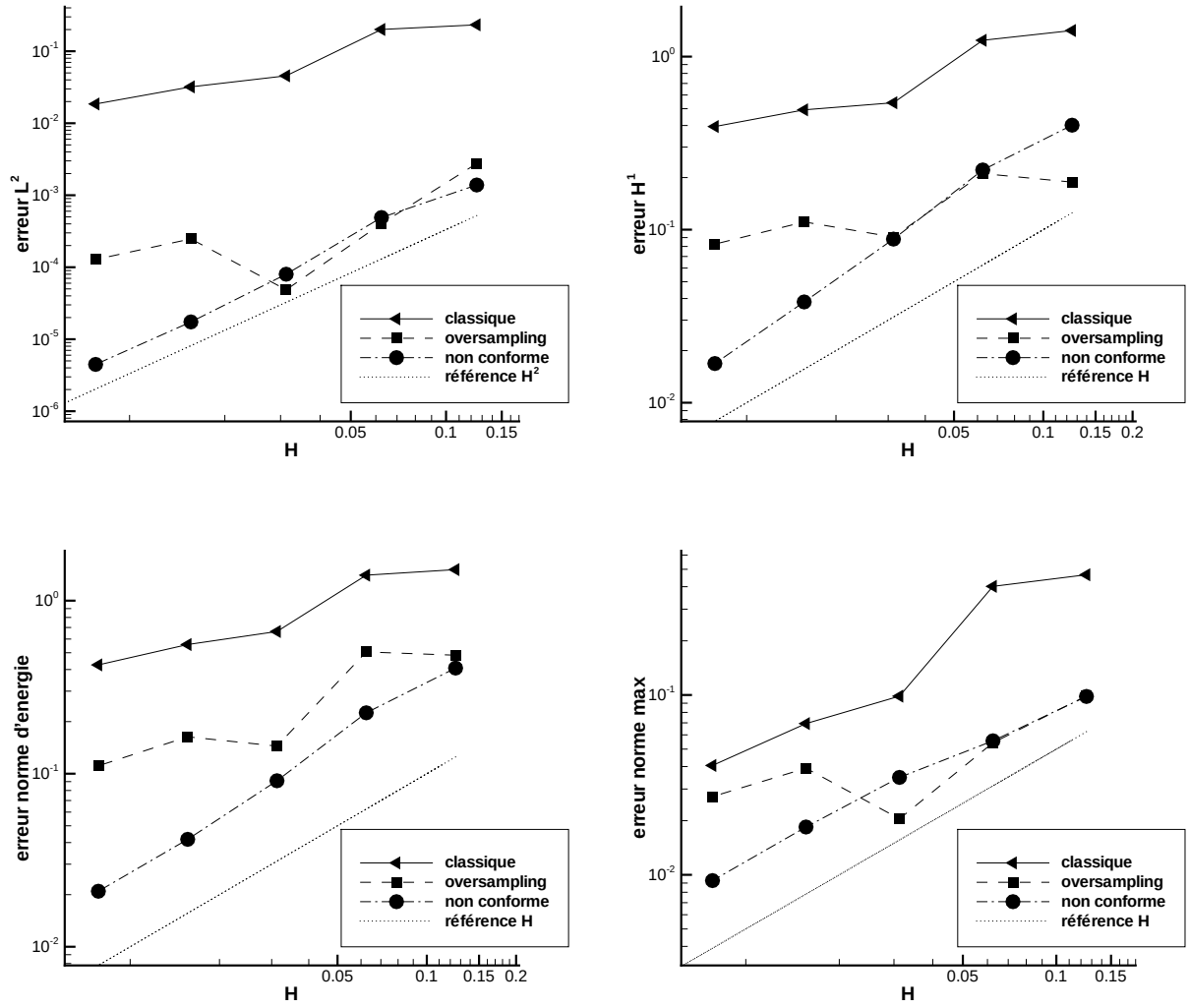


FIGURE 6.1: L'errer de MsFEM. La ligne pointillée montre la puissance appropriée de H, c'est-à-dire H^2 pour le graphe L^2 et H pour tous les autres graphes.

7 Perspectives

Les travaux présentés dans cette parties pourront être poursuivis dans plusieurs directions :

- Application des méthodes “Zoom Numérique” à des problèmes industrielles. L’algorithme 5.1 est particulièrement prometteur pour des problèmes de la mécanique de structures avec de petites inclusions.
- Extension des méthodes “Zoom Numérique” aux problèmes non symétriques. Une idée serait d’utiliser ces algorithmes comme des préconditioneurs dans GMRES.
- L’étude théorique de la nouvelle variante de la méthode MsFEM non conforme.
- Une extension de MsFEM aux problèmes du type Stokes (un travail déjà bien entamé).

Deuxième partie

Sur la modélisation des fluides viscoélastiques

8 Une construction du modèle de chaînes bille-ressort pour les solutions de polymère

D'après l'article [25].

Cette partie est consacrée à quelques questions liées à la modélisation des fluides non newtoniens. Rappelons que la modélisation des écoulements des fluides est basée sur une relation entre le champ de vitesse et le tenseur des contraintes de Cauchy $\boldsymbol{\tau}$. L'exemple le plus simple d'une telle relation est la loi de Newton, $\boldsymbol{\tau} = \eta(\nabla \boldsymbol{u} + (\nabla \boldsymbol{u})^T) - pI$, où η, $\boldsymbol{u}$ et p sont respectivement la viscosité, la vitesse et la pression du fluide. Cette loi mène aux équations de Navier-Stokes qui décrivent convenablement le comportement des fluides les plus répandus composés de petites molécules. Cependant, certains fluides de structure interne complexe (des fluides composés de macro-molécules, suspensions de particules élastiques, etc.) peuvent exiger des relations plus compliquées. On parle alors des fluides non newtoniens ou, plus spécifiquement, des fluides viscoélastiques lorsque le tenseur des contraintes dépend de l'histoire des déformations subies par un élément du fluide.

Nous considérerons ici un des exemples les plus simples de fluides viscoélastiquesé : une solution diluée de polymère. On appliquera l'approche de la théorie cinétique, aussi connue comme l'approche micro-macro. On représente donc les molécules du polymère par des objets relativement simples comme des chaînes bille-ressort (Fig. 8, à gauche) ou des dumbbells ("haltères" en français) – deux billes reliés par un ressort, voir Fig. 8, à droite. Bien que cette description ne descende pas vraiment au niveau atomistique, chaque ressort correspondant à un long segment d'une molécule, elle permet de donner de manière assez précise la distribution statistique des configurations moléculaires. Le couplage micro-macro consiste à étudier l'influence de l'écoulement du solvant (vue au niveau macroscopique) sur l'évolution de la distribution statistique de chaînes ou de dumbbells et la rétroaction des polymères sur le solvant. Bien que les expressions mathématiques qui décrivent l'interaction entre les polymères et le solvant soient bien connues et disponibles dans plusieurs manuels, dont [37] et [57], nous en présentons ici une nouvelle construction qui diffère des approches traditionnelles en plusieurs aspects :

– Notre point du départ est une modélisation complète qui prend en compte les effets d'inertie des billes dans les chaînes représentant les molécules du polymère. Le cas sans inertie est

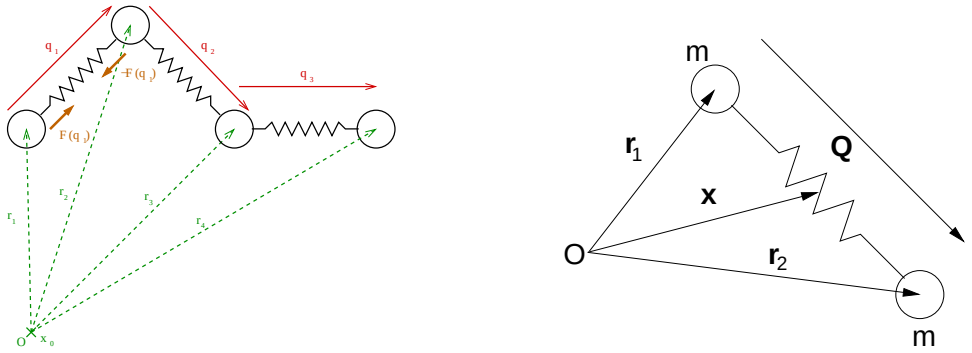


FIGURE 8.1: Une chaîne bille-ressort (à gauche) et une dumbbell (à droite).

obtenu alors à la limite quand la masse de la bille tend vers zéro, en utilisant les techniques de développements asymptotiques. Par contre, dans les constructions que l'on peut trouver dans la littérature, les termes d'inertie dans les équations de mouvement des billes sont négligés tout de suite avec des justifications plus ou moins obscures, voir par exemple paragraphe 13.1 du livre [37]. Bien que le modèle sans inertie soit bien approprié à l'étude de la plupart des phénomènes physiques, le modèle complet avec l'inertie pourrait apporter de nouvelles informations dans le cas de brusque changement de vitesse, par exemple.

– Notre expression de la force exercée sur le solvant par le polymère est une conséquence directe des équations du mouvement des billes et du principe des actions réciproques. Nous retrouvons l'expression traditionnelle – la formule de Kramers – dans la limite quand la masse de la bille tend vers zéro, par le même procédé des développements asymptotiques que celui utilisé pour déduire l'équation de Fokker-Planck pour la distributions statistiques de chaînes. Par contre, traditionnellement, l'expression de Kramers est obtenue par un raisonnement qui est logiquement indépendant de la construction de l'équation de Fokker-Planck, voir paragraphe 13.3 de [37] par exemple.
– Notre construction est effectué dans l'hypothèse d'un écoulement non-homogènes de la solution de polymère, c'est-à-dire sans aucune restriction sur le champ de vitesse. Cela inclut, entre autre chose, une description du phénomène important associé à de tels écoulements qui est la migration moléculaire. Cela veut dire que la vitesse du centre de masse d'une molécule est en général différente de celle du solvant ambiant. De tels phénomènes sont hors de portée de modélisation traditionnelle qui est effectuée à la base dans l'hypothèse d'un écoulement homogène, c'est-à-dire en supposant que la vitesse du solvant (et la vitesse moyenne de chaînes) dépend linéairement de la position en espace.

On va présenter le modèle seulement dans le cas des dumbbells. Une extension au cas de chaînes avec un nombre de ressorts quelconque ne présent aucune difficulté.

8.1 Le modèle prenant en compte l'inertie des billes

On considère une solution diluée de polymères qui remplit un domaine $\Omega \subset \mathbb{R}^d$. La solution est composée de deux phases, du polymère et du solvant, avec deux vitesses macroscopiques distinctes $\boldsymbol{v}_p$ et $\boldsymbol{v}_s$, respectivement. La fraction volumique du polymère φ (et, par conséquent, la fraction volumique du solvant $1-\varphi$) peut varier en espace et en temps. On suppose que le solvant est un fluide incompressible newtonien. Sa vitesse $\boldsymbol{v}_s$ est donc régie par les équation macroscopiques de

conservation du moment linéaire et de la masse (Navier-Stokes), auxquelles on ajoutera la force d'interaction entre les deux phases. Par contre, la phase du polymère est modélisée au niveau mésoscopique, tandis que les quantités macroscopiques $\boldsymbol{v}_p$ et φ sont définies comme des moyennes sur les grands ensembles de dumbbells.

On choisit représenter les molécules du polymère par les dumbbells comprenant deux billes (plus précisément, des points matériels), de masse m chacune, réliées par un ressort qui ne porte pas de masse, voir Fig. 8, à gauche. Pour écrire l'équation du mouvement d'une dumbbell, on note d'abord les positions de ses billes numéros 1 et 2 par les vecteurs $\boldsymbol{r}_1(t)$ et $\boldsymbol{r}_2(t)$ et leur vitesses pat $\boldsymbol{V}_i = \dot{\boldsymbol{r}}_i$, $i = 1, 2$. La vitesse du fluide ambiant, c'est-à-dire celle du solvant, au point $\boldsymbol{r}_i$ est notée par $\boldsymbol{v}_i = \boldsymbol{v}_s(\boldsymbol{r}_i)$. Les équations du mouvement de la bille i s'écrivent alors ainsi :

$$m d\boldsymbol{V}_i = (-\zeta(\boldsymbol{V}_i - \boldsymbol{v}_i) + \boldsymbol{F}_i)\ dt + \sqrt{2\zeta k_B T} d\boldsymbol{W}_i, \tag{8.1}$$

$$d\boldsymbol{r}_i = \boldsymbol{V}_i\ dt. \tag{8.2}$$

On a pris en compte ici les trois forces qui agissent sur une bille d'une dumbbell :

1. La force du ressort $\boldsymbol{F}_i$ agissant sur la bille i. On suppose que $\boldsymbol{F}_1 = -\boldsymbol{F}_2 = \boldsymbol{F}$, où $\boldsymbol{F}$ est une fonction donnée du vecteur bout-à-bout $\boldsymbol{Q} = \boldsymbol{r}_2 - \boldsymbol{r}_1$. La justification de cette force se trouve dans une étude de la distribution du vecteur aléatoire $\boldsymbol{q}$ à l'équilibre thermique. Elle doit assurer notamment que cette distribution est proche de la distribution du vecteur bout-à-bout d'une vraie molécule de polymère en équilibre.
2. La force de frottement qui est proportionnelle à la vitesse relative de la bille par rapport au solvant d'après l'hydrodynamique classique (on peut supposer que le nombre de Reynolds est 0 à l'échelle mésoscopique). Le coefficient de traînée est noté ζ.
3. La force brownienne, provenant des impacts aléatoires des petites molécules du solvant sur la bille. $\boldsymbol{W}_1$ et $\boldsymbol{W}_2$ désignent donc deux mouvement browniens (processus de Wiener) mutuellement indépendants. Le coefficient $\sqrt{2\zeta k_B T}$ avec la constante de Boltzmann k_B et la température T provient du principe d'équipartition de l'énergie.

Remarque 8.1 *La vitesse $\boldsymbol{v}_i = \boldsymbol{v}_s(\boldsymbol{r}_i)$ dans (8.1) doit être comprise comme celle du solvant moyennée sur une boîte de taille plus grande que la taille de la bille mais plus petite que la distance entre les deux extrémités d'une dumbbell. Les fluctuations dans la vitesse du solvant autour de $\boldsymbol{v}_s$ sont donc négligées à l'échelle de cette boîte. Cela implique que le mécanisme de l'interaction hydrodynamique n'est pas inclus dans notre modélisation, c'est-à-dire les fluctuations de la vitesse du solvant induites par les vibrations d'une bille s'effacent sur une distance plus petite que la distance à l'autre bille. Dans une modélisations plus complète, on devrait tenir compte du bruit dans la vitesse du solvant. Le champ $\boldsymbol{v}_s$ satisferait alors une équation stochastique aux dérivées partielles. Une telle approche est esquissée dans [56] sous le nom de Coupled Langevin Equations.*

Les équations (8.1)–(8.2) sont des équations différentielles stochastiques (EDS) pour les processus aléatoires $\boldsymbol{r}_i$, $\boldsymbol{V}_i$. Il est bien connu que la densité de probabilité d'un processus stochastique satisfaisant à une EDS, évolue selon l'équation de Fokker-Planck, dont les coefficients sont faciles à déterminer à partir de coefficients de l'EDS, voir [57] par exemple. Pour alléger les notations on écrira $\boldsymbol{r} = (\boldsymbol{r}_1^T, \boldsymbol{r}_2^T)^T$, $\boldsymbol{V} = (\boldsymbol{V}_1^T, \boldsymbol{V}_2^T)^T$, etc. On introduit la fonction de distribution des dumbbells $\Psi(\boldsymbol{r}, \boldsymbol{V}, t)$ telle que $\Psi(\boldsymbol{r}, \boldsymbol{V}, t)\ d\boldsymbol{r}\ d\boldsymbol{V}$ donne l'espérance du nombre des dumbbells dont les billes sont dans l'intervalle $[\boldsymbol{r}, \boldsymbol{r} + d\boldsymbol{r}]$ avec les vitesses dans l'intervalle $[\boldsymbol{V}, \boldsymbol{V} + d\boldsymbol{V}]$ au temps t. La fonction de distribution $\Psi(\boldsymbol{r}, \boldsymbol{V}, t)$ diffère de la densité de probabilité de processus $(\boldsymbol{r}, \boldsymbol{V})$ juste par une

constante de normalisation. Cette fonction évolue donc selon l'équation de Fokker-Planck suivante

$$\frac{\partial \Psi}{\partial t} + \nabla_{\boldsymbol{r}} \cdot (\boldsymbol{V}\Psi) + \frac{1}{m}\nabla_{\boldsymbol{V}} \cdot ((-\zeta(\boldsymbol{V} - \boldsymbol{v}) + \boldsymbol{F})\Psi) = \frac{\zeta k_B T}{m^2}\nabla^2_{\boldsymbol{V}}\Psi. \tag{8.3}$$

On se tourne maintenant vers la modélisation de la phase de solvant. En utilisant les hypothèses que le tenseur de contrainte est newtonien et que le fluide est incompressible, on peut décrire la conservation de moment linéaire et de la masse par

$$\rho_s(1-\varphi)\frac{D\boldsymbol{v}_s}{Dt} = \eta_s\nabla^2_{\boldsymbol{x}}\boldsymbol{v}_s + \eta_s\nabla_{\boldsymbol{x}}(\nabla_{\boldsymbol{x}} \cdot \boldsymbol{v}_s) - \nabla_{\boldsymbol{x}}p_s + \boldsymbol{f}, \tag{8.4}$$

$$\frac{\partial(1-\varphi)}{\partial t} + \nabla_{\boldsymbol{x}} \cdot ((1-\varphi)\boldsymbol{v}_s) = 0. \tag{8.5}$$

Ici, p_s est la pression dans le solvant et $\boldsymbol{f} = \boldsymbol{f}(\boldsymbol{x}, t)$ est la densité de force exercée en point $\boldsymbol{x}$ et en temps t sur le solvant par les dumbbells. La dérivée matérielle est associée à la vitesse du solvant, c'est-à-dire $\frac{D}{Dt} = \frac{\partial}{\partial t} + \boldsymbol{v}_s \cdot \nabla_{\boldsymbol{x}}$.

L'idée cruciale qui permet définir $\boldsymbol{f}$ consiste à observer que toutes les interactions entre le polymère et le solvant sont décrites par l'équation du mouvement des dumbbells (8.1). Mise à part la force du ressort (qui est une force interne de la phase polymérique), l'action du solvant sur le polymère est représentée par la force de frottement $-\zeta(\boldsymbol{V}_i - \boldsymbol{v}_i)$ et la force brownienne. Cette dernière ne doit pas entrer dans le calcul de la force moyenne $\boldsymbol{f}$. En effet, les force Browniennes représentent des impacts aléatoires qu'une molécule du polymère reçoit de la part du solvant ambiant. Mais $\boldsymbol{f}$ est une quantité macroscopique, et donc son calcul entraîne la moyennisation de grandes ensembles de chaînes. Les impacts aléatoires doivent donc s'annuler mutuellement dans cette procédure puisque la probabilité d'un tel impact est la même que celle d'un impact de la même intensité mais dans la direction opposée. On voit donc que la formule pour $\boldsymbol{f}$ ne doit contenir que la force de frottement. Par le principe de l'action réciproque, cette force est de signe opposé dans l'expression de $\boldsymbol{f}$ par rapport à l'équation du mouvement des billes (8.1) :

$$\boldsymbol{f}(\boldsymbol{x}, t) = \int\int_\Omega \zeta(\boldsymbol{V}_1 - \boldsymbol{v}_1)\ \Psi|_{\boldsymbol{r}_1=\boldsymbol{x}}\ d\boldsymbol{r}_2 d\boldsymbol{V} + \int\int_\Omega \zeta(\boldsymbol{V}_2 - \boldsymbol{v}_2)\ \Psi|_{\boldsymbol{r}_2=\boldsymbol{x}}\ d\boldsymbol{r}_1 d\boldsymbol{V}. \tag{8.6}$$

La première intégrale ici prend en compte les dumbbells dont la bille numéro 1 est à la position $\boldsymbol{x}$ et la deuxième intégrale prend en compte les billes numéro 2 en $\boldsymbol{x}$.

Pour fermer le système, on doit encore fournir une équation pour la fraction volumique du polymère φ. On postule qu'elle est proportionnelle à la densité numérique n des dumbbells

$$\varphi = nV_d, \tag{8.7}$$

où V_d peut être interprétée comme le volume occupé par une dumbbell. La densité numérique est donnée par

$$n(\boldsymbol{x}, t) := \frac{1}{2}\left(\int\int_\Omega \Psi|_{\boldsymbol{r}_1=\boldsymbol{x}}\ d\boldsymbol{r}_2 d\boldsymbol{V} + \int\int_\Omega \Psi|_{\boldsymbol{r}_2=\boldsymbol{x}}\ d\boldsymbol{r}_1 d\boldsymbol{V}\right). \tag{8.8}$$

Cela achève la modélisation de la solution des dumbbells. Le système des équations régissant l'écoulement est composé donc de l'équation de Fokker-Planck (8.3) pour la phase de polymère et des équations du type Navier-Stokes (8.5)–(8.4) pour la phase de solvant qui sont couplées par les expressions de la densité numérique, de la fraction volumique et de la force inter-phasique, c'est-à-dire (8.8), (8.7) et (8.6), respectivement. On voit que ces équations ne font pas intervenir la vitesse

du polymère $\boldsymbol{v}_p$ ce qui est consistant avec notre choix de modéliser le polymère à l'échelle mésoscopique. Néanmoins, la vitesse du polymère $\boldsymbol{v}_p$ ainsi que la vitesse moyenne de la solution $\boldsymbol{u}$, peuvent être calculées dans un post-processing :

$$\boldsymbol{v}_p = \frac{1}{2n}\left(\int\int_\Omega \boldsymbol{V}_1\Psi|_{\boldsymbol{r}_1=\boldsymbol{x}}\, d\boldsymbol{r}_2 d\boldsymbol{V} + \int\int_\Omega \boldsymbol{V}_2\Psi|_{\boldsymbol{r}_2=\boldsymbol{x}}\, d\boldsymbol{r}_1 d\boldsymbol{V}\right), \tag{8.9}$$

$$\boldsymbol{u} = \varphi\boldsymbol{v}_p + (1-\varphi)\boldsymbol{v}_s. \tag{8.10}$$

En comparant (8.6) avec les définitions (8.8) et (8.9) et en observant que $\boldsymbol{v}_i = \boldsymbol{v}_s(\boldsymbol{x})$ lorsque $\boldsymbol{r}_i = \boldsymbol{x}$, on obtient encore une formule utile pour la force $\boldsymbol{f}$:

$$\boldsymbol{f} = 2n\zeta(\boldsymbol{v}_p - \boldsymbol{v}_s). \tag{8.11}$$

Notons encore que l'équation de Fokker-Planck (8.3) implique la loi de conservation pour la densité numérique n

$$\frac{\partial n}{\partial t} + \nabla_{\boldsymbol{x}} \cdot (\boldsymbol{v}_p n) = 0. \tag{8.12}$$

En multipliant cette équation par V_d et en ajoutant (8.5), on a

$$\nabla_{\boldsymbol{x}} \cdot \boldsymbol{u} = 0. \tag{8.13}$$

La solution des dmbbells est donc un fluide incompressible.

8.2 La limite sans inertie des billes

L'équation de Fokker-Planck

On procède maintenant à la dérivation d'une variante sans inertie de notre modèle, c'est-à-dire de la limite $m \to 0$. On introduit les scalings suivants :

$$m = \varepsilon^2,\ \boldsymbol{p} = \varepsilon\boldsymbol{V}. \tag{8.14}$$

Ces scalings sont motivés par le théorème de l'équipartition d'énergie qui donne la relation

$$\left\langle mV_i^2 \right\rangle = 3k_BT$$

dans l'état d'équilibre thermique. En supposant que la distribution des vitesse n'est pas trop éloignée de celle en équilibre, on obtient donc la valeur caractéristiques de vitesse $\sqrt{k_BT/m}$, ce qui implique que $\boldsymbol{V}$ se comporte comme $1/\varepsilon$ lorsque $\varepsilon \to 0$.

Remarque 8.2 *On traitera ε comme un petit paramètre dans la suite, ce qui n'a pas vraiment de sens puisque ε est une quantité dimensionnelle. Cependant, avec les non-dimensionalisations standards, on peut voir facilement que la contre-partie adimensionnelle de ε est $\sqrt{\lambda_B/\lambda_H}$ où $\lambda_B = m/\zeta$ est l'échelle de temps caractéristique pour les fluctuations de vitesse et $\lambda_H = \zeta/4H$ est la période de relaxation caractéristique d'une dumbbell, H étant la rigidité typique du ressort. L'hypothése $\lambda_B << \lambda_H$ est assez bien vérifiée dans la plupart des expériences, cf. [62], ce qui justifie la limite $\varepsilon \to 0$. Ici, nous gardons un ε dimensionnel pour alléger les notations.*

On introduit le développement asymptotique en ε pour Ψ

$$\Psi = \Psi_0 + \varepsilon\Psi_1 + \varepsilon^2\Psi_2 + \dots. \tag{8.15}$$

On va établir une équation directement pour le terme principal Ψ_0 et pour sa distribution marginale en $\boldsymbol{r}$: $\psi(\boldsymbol{r},t) = \int \Psi_0 \, d\boldsymbol{p}$.

L'équation de Fokker-Planck (8.3) se réécrit dans les nouvelle variables comme

$$\frac{\partial\Psi}{\partial t} + \frac{1}{\varepsilon}\nabla_{\boldsymbol{r}}\cdot(\boldsymbol{p}\Psi) + \frac{1}{\varepsilon}\nabla_{\boldsymbol{p}}\cdot((\zeta\boldsymbol{v}+\boldsymbol{F})\Psi) = \frac{1}{\varepsilon^2}\zeta Q(\Psi), \tag{8.16}$$

où

$$Q(\Psi) := \nabla_{\boldsymbol{p}}\cdot(\boldsymbol{p}\Psi) + k_BT\nabla_{\boldsymbol{p}}^2\Psi. \tag{8.17}$$

$$= k_BT\nabla_{\boldsymbol{p}}\cdot\left(M\nabla_{\boldsymbol{p}}\left(\frac{\Psi}{M}\right)\right) \tag{8.18}$$

et M est la distribution de Maxwell

$$M(\boldsymbol{p}) = C\exp(-p^2/2k_BT), \tag{8.19}$$

avec la constante de normalisation C. En substituant le développement asymptotique (8.15) dans (8.16), on obtient

$$\frac{\partial\Psi_{k-1}}{\partial t} + \nabla_{\boldsymbol{r}}\cdot(\boldsymbol{p}\Psi_k) + \nabla_{\boldsymbol{p}}\cdot((\zeta\boldsymbol{v}+\boldsymbol{F})\Psi_k) = \zeta Q(\Psi_{k+1}), \tag{8.20}$$

pour $k = -1, 0, 1, \dots$ avec $\Psi_{-1} = \Psi_{-2} = 0$.

Posons d'abord $k = -1$ dans (8.20). Cela donne $Q(\Psi_0) = 0$, ce qui implique que $\Psi_0(\boldsymbol{r},\boldsymbol{p},t)$ est proportionnel à $M(\boldsymbol{p})$ pour tout $\boldsymbol{r}$, t et donc $\Psi_0(\boldsymbol{r},\boldsymbol{p},t) = \psi(\boldsymbol{r},t)M(\boldsymbol{p})$.

Posons maintenant $k = 0$ dans (8.20). En multipliant (8.20) par $\boldsymbol{p}/\zeta$ et en intégrant par rapport à $\boldsymbol{p}$ on arrive à

$$\int \boldsymbol{p}Q(\Psi_1)\, d\boldsymbol{p} = \frac{1}{\zeta}\left(\nabla_{\boldsymbol{r}}\cdot\int \boldsymbol{p}\boldsymbol{p}\Psi_0\, d\boldsymbol{p} + \int \boldsymbol{p}\nabla_{\boldsymbol{p}}\cdot((\zeta\boldsymbol{v}+\boldsymbol{F})\Psi_0)\, d\boldsymbol{p}\right). \tag{8.21}$$

En utilisant l'identité $\boldsymbol{p} + \frac{k_BT}{M}\nabla_{\boldsymbol{p}}M = \boldsymbol{0}$ et en faisant quelques intégrations par parties, on voit que

$$\int \boldsymbol{p}Q(\Psi_1)\, d\boldsymbol{p} = k_BT\int \boldsymbol{p}\nabla_{\boldsymbol{p}}\cdot\left(M\nabla_{\boldsymbol{p}}\left(\frac{\Psi_1}{M}\right)\right)\, d\boldsymbol{p} = -k_BT\int M\nabla_{\boldsymbol{p}}\left(\frac{\Psi_1}{M}\right)\, d\boldsymbol{p},$$

$$= k_BT\int \frac{\Psi_1}{M}\nabla_{\boldsymbol{p}}M\, d\boldsymbol{p} = -\int \boldsymbol{p}\Psi_1\, d\boldsymbol{p}. \tag{8.22}$$

et

$$\int \boldsymbol{p}\boldsymbol{p}\Psi_0\, d\boldsymbol{p} = \psi\int \boldsymbol{p}\boldsymbol{p}M\, d\boldsymbol{p} = -k_BT\psi\int \boldsymbol{p}\,(\nabla_{\boldsymbol{p}}M)\, d\boldsymbol{p}$$

$$= k_BT\psi\boldsymbol{\delta}\int M\, d\boldsymbol{p} = k_BT\psi\boldsymbol{\delta}. \tag{8.23}$$

En substituant ces identités dans (8.21), on obtient la relation entre le premier moment de Ψ_1 et ψ

$$\int \boldsymbol{p}\Psi_1 \, d\boldsymbol{p} = \frac{1}{\zeta}\left((\zeta \boldsymbol{v} + \boldsymbol{F})\psi - k_B T \nabla_{\boldsymbol{r}}\psi\right). \tag{8.24}$$

D'autre part, en posant $k = 1$ dans (8.20) et en intégrant par rapport à $\boldsymbol{p}$, on arrive à

$$\frac{\partial \psi}{\partial t} = -\nabla_{\boldsymbol{r}} \cdot \int \boldsymbol{p}\Psi_1 \, d\boldsymbol{p}. \tag{8.25}$$

L'équation (8.24) donne une fermeture pour $\psi(\boldsymbol{r}, t)$ dans (8.25) :

$$\begin{aligned} \frac{\partial \psi}{\partial t} &= \frac{1}{\zeta}\nabla_{\boldsymbol{r}} \cdot (k_B T \nabla_{\boldsymbol{r}}\psi - (\zeta \boldsymbol{v} + \boldsymbol{F})\psi). \\ &= \frac{1}{\zeta}\sum_{i=1}^{2} \nabla_{\boldsymbol{r}_i} \cdot (k_B T \nabla_{\boldsymbol{r}_i}\psi - (\zeta \boldsymbol{v}_i + \boldsymbol{F}_i)\psi). \end{aligned} \tag{8.26}$$

C'est l'équation bien connue de Fokker-Planck pour la solution diluée de dumbbells avec l'inertie négligée.

Expression de la force d'interaction entre le polymère et le solvant

Il reste à déterminer, dans la même limite de l'inertie négligée, une expression pour la force $\boldsymbol{f}$ qui apparaît dans (8.4). On écrit $\boldsymbol{V}_i = \boldsymbol{p}_i/\sqrt{m}$ $(= \boldsymbol{p}_i/\varepsilon)$ dans (8.6) et on garde seulement les termes d'ordre $O(1)$ en ε. Cela donne la formule pour la force

$$\begin{aligned} \boldsymbol{f}(\boldsymbol{x}, t) = \zeta \int \int_{\Omega} \boldsymbol{p}_1 \Psi_1|_{\boldsymbol{r}_1 = \boldsymbol{x}} \, d\boldsymbol{r}_2 d\boldsymbol{p} - \zeta \int_{\Omega} \boldsymbol{v}_1 \psi|_{\boldsymbol{r}_1 = \boldsymbol{x}} \, d\boldsymbol{r}_2 \\ + \zeta \int \int_{\Omega} \boldsymbol{p}_2 \Psi_1|_{\boldsymbol{r}_2 = \boldsymbol{x}} \, d\boldsymbol{r}_1 d\boldsymbol{p} - \zeta \int_{\Omega} \boldsymbol{v}_2 \psi|_{\boldsymbol{r}_2 = \boldsymbol{x}} \, d\boldsymbol{r}_1. \end{aligned} \tag{8.27}$$

En utilisant (8.24) elle peut être réécrite ainsi

$$\begin{aligned} \boldsymbol{f}(\boldsymbol{x}, t) = &\left(\int_{\Omega} \boldsymbol{F}_1 \psi|_{\boldsymbol{r}_1 = \boldsymbol{x}} \, d\boldsymbol{r}_2 + \int_{\Omega} \boldsymbol{F}_2 \psi|_{\boldsymbol{r}_2 = \boldsymbol{x}} \, d\boldsymbol{r}_1\right) - k_B T \left(\int_{\Omega} \nabla_{\boldsymbol{r}_1} \psi|_{\boldsymbol{r}_1 = \boldsymbol{x}} \, d\boldsymbol{r}_2 + \int_{\Omega} \nabla_{\boldsymbol{r}_2} \psi|_{\boldsymbol{r}_2 = \boldsymbol{x}} \, d\boldsymbol{r}_1\right) \\ =: \boldsymbol{f}^s(\boldsymbol{x}, t) \qquad & + \boldsymbol{f}^t(\boldsymbol{x}, t). \end{aligned} \tag{8.28}$$

On identifie ici $\boldsymbol{f}^s$ comme la partie de force due aux ressorts et $\boldsymbol{f}^t$ comme la partie de force due aux fluctuations thermiques.

En passant aux coordonnées du centre de masse de la dumbbell $\mathbf{x}$ et du vecteur bout-à-bout $\mathbf{Q}$

$$\boldsymbol{x} := \frac{1}{2}(\boldsymbol{r}_1 + \boldsymbol{r}_2), \; \boldsymbol{Q} := \boldsymbol{r}_2 - \boldsymbol{r}_1, \tag{8.29}$$

on peut représenter $\boldsymbol{f}^s$ comme la divergence d'un tenseur

$$\begin{aligned} \boldsymbol{f}^s(\mathbf{x}, t) &= \int \mathbf{F}(\psi(\mathbf{x} + \frac{1}{2}\mathbf{Q}, \mathbf{Q}, t) - \psi(\mathbf{x} - \frac{1}{2}\mathbf{Q}, \mathbf{Q}, t))dq \\ &= \nabla_{\boldsymbol{x}} \cdot \int_{-1/2}^{1/2} \int \psi(\boldsymbol{x} + s\boldsymbol{Q}, \boldsymbol{Q}, t)\boldsymbol{Q} \otimes \boldsymbol{F} \, d\boldsymbol{Q} ds. \end{aligned}$$

Pour le deuxième terme de la force $\boldsymbol{f}^t$, on obtient immédiatement

$$\boldsymbol{f}^t(\boldsymbol{x},t) = -k_BT\nabla_{\boldsymbol{x}}\left(\int_\Omega \psi|_{\boldsymbol{r}_1=\boldsymbol{x}}\, d\boldsymbol{r}_2 + \int_\Omega \psi|_{\boldsymbol{r}_2=\boldsymbol{x}}\, d\boldsymbol{r}_1\right) = -2k_BT\nabla_{\boldsymbol{x}} n, \tag{8.30}$$

où n est toujours la densité numérique des dumbbells calculée à la limite $\varepsilon \to 0$ par

$$n(\boldsymbol{x},t) = \frac{1}{2}\left(\int_\Omega \psi|_{\boldsymbol{r}_1=\boldsymbol{x}}\, d\boldsymbol{r}_2 + \int_\Omega \psi|_{\boldsymbol{r}_2=\boldsymbol{x}}\, d\boldsymbol{r}_1\right). \tag{8.31}$$

La force $\boldsymbol{f}$ dans (8.4) peut être donc représentée comme la divergence d'un tenseur

$$\boldsymbol{f} = \nabla_{\boldsymbol{x}} \cdot \boldsymbol{\tau}_p, \tag{8.32}$$

où $\boldsymbol{\tau}_p$ est la contribution polymérique dans le tenseur de contraintes

$$\boldsymbol{\tau}_p(\boldsymbol{x},t) := \int_{-1/2}^{1/2}\int_{\mathcal{Q}_s(\boldsymbol{x})} \psi(\boldsymbol{x}+s\boldsymbol{Q},\boldsymbol{Q},t)\boldsymbol{Q}\otimes\boldsymbol{F}\, d\boldsymbol{Q}ds - 2n(\boldsymbol{x},t)k_BTI \tag{8.33}$$

On obtient donc l'expression de Kramers pour le tenseur des contraintes dans la forme généralisée pour les écoulements non homogènes, que l'on trouve par exemple dans [36].

8.3 Le cas des écoulements localement homogènes

On va retrouver maintenant les équations habituelles du modèle des dumbbells dans le cas des écoulements localement homogènes. Cela veut dire que la vitesse du fluide dépend linéairement de position sur l'échelle d'une dumbbell. Autrement dit, c'est le cas où $\ell_0/L << 1$ avec ℓ_0 la taille caractéristique d'une dumbbell et L l'échelle macroscopique de l'écoulement.

On note V la vitesse caractéristique du fluide, n_{av} la valeur moyen de la densité numérique n. L'équation de Fokker-Planck (8.26) donne la longueur typique d'une dumbbell $\ell_0 = \sqrt{k_BT/H}$ où H est la rigidité typique du ressort. On introduit des variables adimensionnelles :

$$\boldsymbol{x}^* = \frac{\boldsymbol{x}}{L},\ \boldsymbol{v}_s^* = \frac{\boldsymbol{v}_s}{V},\ t^* = \frac{tV}{L},\ \boldsymbol{Q}^* = \frac{\boldsymbol{Q}}{\ell_0},\ \boldsymbol{F}^* = \frac{\boldsymbol{F}}{H\ell_0}. \tag{8.34}$$

On réécrit l'équation de Fokker-Planck (8.26) dans les coordonnées du centre de masse de la dumbbell $\mathbf{x}$ et du vecteur bout-à-bout $\mathbf{Q}$ (8.29) et on passe aux variables adimensionelles. Cela donne

$$\begin{aligned}\frac{D\psi}{Dt^*} = \nabla_{\boldsymbol{Q}^*}\cdot\left(\frac{1}{2De}\nabla_{\boldsymbol{Q}^*}\psi + \frac{\boldsymbol{F}^*}{2De}\psi - \left\{(\boldsymbol{Q}^*\cdot\nabla_{\boldsymbol{x}^*})\boldsymbol{v}_s^* + \frac{1}{24}\left(\frac{\ell_0}{L}\right)^2(\boldsymbol{Q}^*\cdot\nabla_{\boldsymbol{x}^*})^3\boldsymbol{v}_s^* + \dots\right\}\psi\right)\\ + \nabla_{\boldsymbol{x}^*}\cdot\left(\frac{1}{8De}\left(\frac{\ell_0}{L}\right)^2\nabla_{\boldsymbol{x}^*}\psi - \left\{\frac{1}{4}\left(\frac{\ell_0}{L}\right)^2(\boldsymbol{Q}^*\cdot\nabla_{\boldsymbol{x}^*})^2\boldsymbol{v}_s^* + \cdots\right\}\psi\right),\end{aligned} \tag{8.35}$$

où $De = \lambda V/L$ est le nombre adimensionnel de Deborah qui mesure l'importance des effets viscoélastiques et $\lambda = \zeta/(4H)$ est le temps de relaxation du polymère. En négligeant les termes d'ordre $(\ell_0/L)^2$ ou plus petit et en revenant aux variables dimensionnelles $\boldsymbol{x}$ et t, mais en gardant la forme

adimensionnelle pour $\boldsymbol{Q}$, on arrive à l'équation de Fokker-Planck classique pour les écoulements localement homogènes :

$$\frac{D\psi}{Dt} = \nabla_{\boldsymbol{Q}} \cdot \left(\frac{1}{2\lambda} \nabla_{\boldsymbol{Q}} \psi + \frac{1}{2\lambda} \boldsymbol{F}\psi - (\nabla_{\boldsymbol{x}} \boldsymbol{v}_s)\boldsymbol{Q}\psi \right). \tag{8.36}$$

Comme le laplacien par rapport à $\boldsymbol{x}$ a maintenant disparu, on voit que l'intégrale $\int \psi(\boldsymbol{x}, \boldsymbol{Q}, t) d\boldsymbol{Q}$ est conservée le long des lignes du courant, ce qui implique la conservation de la densité numérique n. On peut donc supposer que $n = n_{av} = const$. De même, la formule (8.11) implique que

$$\boldsymbol{v}_p - \boldsymbol{v}_s = O\left(\frac{k_B T}{2\zeta L}\right) = O\left(\frac{V}{De}\left(\frac{\ell_0}{L}\right)^2\right)$$

et donc on peut supposer $\boldsymbol{v}_p = \boldsymbol{v}_s = \boldsymbol{u}$ dans ce régime.

Dans le cas des écoulements localement homogènes, il est plus commode de travailler avec la densité de probabilité au lieu de densité de distribution numérique de dumbbells. On fait donc le changement de variable $\psi \to \psi n_{av}/\ell_0^d$ de manière que $\psi(t, \boldsymbol{x}, \boldsymbol{Q}) d\boldsymbol{Q}$ désigne maintenant la probabilité de trouver une dumbbell ayant le vecteur bout-à-bout dans l'intervalle $[\boldsymbol{Q}, \boldsymbol{Q} + d\boldsymbol{Q}]$ parmi toutes les dumbbells dont le centre de masse est en $\boldsymbol{x}$ au temps t. Ce changement de ψ ne modifie pas (8.36) car cette équation est linéaire en ψ, mais il va introduire un coefficient dans la formule pour le tenseur des contraintes (8.33). En effectuant un développement limité en $\boldsymbol{Q}$ dans cette formule, comme on a déjà fait dans (8.35), et en négligeant des termes d'ordre $(\ell_0/L)^2$, on arrive finalement à l'expression suivante pour la contribution polymérique au tenseur des contraintes au cas localement homogène

$$\boldsymbol{\tau}_p(\boldsymbol{x}, t) = n k_B T \left(\int \psi(\boldsymbol{x}, \boldsymbol{Q}, t) \boldsymbol{Q} \otimes \boldsymbol{F}\, d\boldsymbol{Q} - I \right) \tag{8.37}$$

qui est *l'expression de Kramers* habituelle. On a fait implicitement ici un petit changement dans le dernier terme de (8.37) en ajoutant nk_BTI. Cela ne modifie en rien le modèle complet parce que ce terme est récupéré comme pression dans l'équation du moment linéaire (8.4). L'avantage de la définition (8.37), par rapport à (8.33), réside dans le fait que $\boldsymbol{\tau}_p$ dans (8.37) s'annule à l'équilibre.

Pour donner le système complet des équations dans le cas localement homogène, on rappelle que les équations (8.36)–(8.37) doivent être couplées avec les équations Navier-Stokes qui prennent la forme

$$\begin{aligned} \rho \frac{D\boldsymbol{v}_s}{Dt} &= \eta_s \nabla_{\boldsymbol{x}}^2 \boldsymbol{v}_s - \nabla_{\boldsymbol{x}} p + \nabla_{\boldsymbol{x}} \cdot \boldsymbol{\tau}_p, \\ \nabla_{\boldsymbol{x}} \cdot \boldsymbol{v}_s &= 0 \end{aligned}$$

avec $\rho = \rho_s(1 - \varphi)$ et $p = p_s + nk_BT$.

9 Sur le choix de la force de ressort

D'après l'article [24].

9.1 Motivation physiques

On se pose dans ce chapitre la question du choix raisonnable de la fonction $\mathbf{F}(\mathbf{Q})$ qui donne la loi de force de ressort dans le modèle des dumbbells. On se restreint aux écoulements homogènes, c'est-à-dire on suppose que l'écoulement macroscopique est imposé et que le champ de vitesse est linéaire par rapport à la position dans le fluide. On considère toujours une solution diluée des dumbbells et on rappelle que la distribution du vecteur bout-à-bout $\mathbf{Q}$ satisfait l'équation de Fokker-Planck (8.36) avec le gradient de vitesse $\nabla_{\boldsymbol{x}} \boldsymbol{v}_s = \kappa$ supposé être une matrice donnée. Comme on a une équivalence entre les équations de Fokker-planck et des EDS, on voit que le processus stochastique $\mathbf{Q}_t$ évolue selon l'EDS

$$d\mathbf{Q}_t = \left(\kappa \mathbf{Q}_t - \frac{1}{2\lambda} \mathbf{F}(\mathbf{Q}_t) \right) dt + \sqrt{\frac{1}{\lambda}} d\mathbf{B}_t, \tag{9.1}$$

où $\mathbf{B}_t$ est le mouvement brownien. La contribution polymérique $\boldsymbol{\tau}$ au tenseur des contraintes peut être calculé par (8.37), c'est-à-dire

$$\boldsymbol{\tau}(t) = \frac{\eta_p}{\lambda} \left(-I + \mathbb{E}(\mathbf{Q}_t \otimes \mathbf{F}(\mathbf{Q}_t)) \right), \tag{9.2}$$

où $\mathbb{E}(\cdot)$ désigne l'espérance. Le paramètre $\eta_p = n k_B T \lambda$ est connu sous le nom de la viscosité polymérique.

La fonction $\mathbf{F}(\cdot)$ devrait être choisie, en pratique, sur la base d'une description plus détaillée des forces à l'intérieur des molécules du polymère. A cette fin, on peut d'abord représenter le polymère par une chaîne composée d'un grand nombre de liens, chacun de longueur fixée, qui peuvent être soit indépendants l'un de l'autre (Freely Jointed Chain), soit avoir des directions aléatoires en préservant

les angles fixés entre les liens adjacents (Freely Rotating Chain). Dans les deux cas, le vecteur aléatoire $\mathbf{Q}_t$ dans (9.1) représente approximativement le vecteur qui va d'un bout à l'autre de la chaîne. En considérant la distribution d'équilibre et en appliquant le théorème central limite, on peut montrer au la loi de la force est alors linéaire : $\mathbf{F}(\mathbf{Q}) = \mathbf{Q}$ pour des petite extensions $\mathbf{Q}$, c'est-à-dire pour des petites déviations de l'équilibre. Toutefois, cette loi n'est certainement pas réaliste dans la limite d'extension complète de la chaîne, parce qu'elle permet aux dumbbells d'être étirées infiniment. Une alternative mathématiquement attrayante pour introduire l'extensibilité maximum dans la modélisation est de définir la force comme le sous-differentiel du potentiel suivant

$$\Pi(X) = \begin{cases} \frac{1}{2}|X|^2, \text{ if } |X| < R \\ +\infty, \text{ sinon} \end{cases} \tag{9.3}$$

avec un R fixé. Physiquement, cela veut dire que la force $\mathbf{F}(.)$ reste linéaire $\mathbf{F}(\mathbf{Q}) = \mathbf{Q}$ quand la chaîne n'est pas étirée au maximum ($|\mathbf{Q}| < R$). Mais, dès que $|\mathbf{Q}|$ atteint la valeur maximale R, la force prend momentanément l'amplitude nécessaire pour empêcher la dumbbell d'être étirée encore plus. Les équations différentielles stochastiques avec un sous-différentiel d'un potentiel infini comme le terme de dérive sont bien connues dans la littérature mathématique sous le nom de problème de Skorohod. L'écriture rigoureuse de (9.1) avec la force dérivée de (9.3) est établie de manière suivante : étant donnée la condition initiale $\mathbf{q} \in B(0, R)$, trouver les processus $\mathbf{Q}_t$ et μ_t à valeurs dans $\mathbb{R}^d$, continus et progressivement mesurables, tels qu'ils vérifient :

– l'équation

$$d\mathbf{Q}_t = \left(\kappa\mathbf{Q}_t - \frac{1}{2\lambda}\mathbf{Q}_t\right)dt - d\mu_t + \sqrt{\frac{1}{\lambda}}d\mathbf{B}_t, \tag{9.4}$$

– la condition initiale $\mathbf{Q}_0 = \mathbf{q}$,
– $\mathbf{Q}_t \in \overline{B(0, R)}$,
– μ_t est de variation bornée sur $[0, T]$ et $\mu_0 = 0$,
– pour tout processus $\mathbf{Z}_t$ continu et progressivement mesurable à valeurs dans $\overline{B(0, R)}$, on a

$$\int_0^T (\mathbf{Q}_t - \mathbf{Z}_t)d\mu_t \geq 0, \ \forall T > 0.$$

où μ_t est an processus stochastique qui force le processus à rester dans la boule $\overline{B(0, R)}$.

La formule pour le tenseur des contraintes (9.2) n'est pas facile à interpréter dans ce cas puisque la force $\mathbf{F}$ n'est plus définie comme une fonction univoque de $\mathbf{Q}$. Pour contourner ce problème, on propose d'approcher la force exacte $\partial\Pi$ par une expression pénalisée

$$F^\varepsilon(\mathbf{Q}) = \mathbf{Q} + \frac{1}{\varepsilon}\beta(\mathbf{Q}), \tag{9.5}$$

où $\beta(\mathbf{Q}) = \mathbf{Q} - \pi(\mathbf{Q})$ et $\pi(\mathbf{Q})$ est la projection de $\mathbf{Q}$ sur $\overline{B(0, R)}$

$$\pi(\mathbf{Q}) = \begin{cases} \mathbf{Q}, \text{ if } |\mathbf{Q}| \leq R \\ \frac{R}{|\mathbf{Q}|}\mathbf{Q}, \text{ sinon.} \end{cases}$$

La solution de (9.4) est donc approchée par le processus $\mathbf{Q}_t^\varepsilon$ qui satisfait

$$d\mathbf{Q}_t^\varepsilon = \left(\kappa\mathbf{Q}_t^\varepsilon - \frac{1}{2\lambda}\mathbf{Q}_t^\varepsilon - \frac{1}{2\lambda\varepsilon}\beta(\mathbf{Q}_t^\varepsilon)\right)dt + \sqrt{\frac{1}{\lambda}}d\mathbf{B}_t. \tag{9.6}$$

La pénalisation des équations stochastiques réfléchies a été étudiée dans [55] dans le but de montrer l'existence de diffusions réfléchies dans le cas général d'un domaine convexe borné. On peut donc admettre déjà que les processus $\mathbf{Q}_t^\varepsilon$ et $\mathbf{Q}_t$ sont bien définis et, en plus, $\mathbf{Q}_t^\varepsilon$ converge vers $\mathbf{Q}_t$ lorsque $\varepsilon \to 0$ dans le sens précisé plus bas. Néanmoins, ces résultats connus ne permettent pas d'assurer l'existence de la limite du tenseur de contrainte pénalisé

$$\tau^\varepsilon(t) = \frac{\eta_p}{\lambda}\left(\mathbb{E}\left(\mathbf{Q}_t^\varepsilon \otimes \mathbf{Q}_t^\varepsilon + \mathbf{Q}_t^\varepsilon \otimes \frac{1}{\varepsilon}\beta(\mathbf{Q}_t^\varepsilon)\right) - I\right) \tag{9.7}$$

lorsque ε tend vers 0. Le résultat principal de ce chapitre sera la preuve de l'existence de cette limite $\tau(t) = \lim_{\varepsilon\to 0}\tau^\varepsilon(t)$. On donnera aussi une formule pour calculer cette limite à partir de la solution de (9.4) elle-même. On finira par quelques simulations numériques illustrant notre approche.

Remarque 9.1 *Il existe un autre choix usuel de $\mathbf{F}(\cdot)$ – la loi FENE (Finitely Extensible Non-linear Elastic) [65]*

$$F_{FENE}(\mathbf{Q}) = \frac{\mathbf{Q}}{1 - |\mathbf{Q}|^2/R^2}. \tag{9.8}$$

L'existence et unicité des solutions $\mathbf{Q}_t$ à (9.1) avec la force FENE est montrée dans [49] sous l'hypothèse $R \geq \sqrt{2}$. Le processus $\mathbf{Q}_t$ ne quitte jamais la boule $B(0,R)$ dans ce cas. L'objectif de restreindre l'extension des dumbbells à une longueur maximum est ainsi atteint avec la force FENE. De plus, cette loi donne une bonne approximation de la fonction de répartition pour une "Freely Jointed Chain" en équilibre. Nous pensons néanmoins que la force dérivée du potentiel (9.3) mérite également une attention comme une alternative à la loi FENE qui pourrait être plus appropriée pour d'autres modélisations sous-jacentes des macro-molécules.

9.2 Des résultats théoriques sur le modèle avec réflexion

On énonce d'abord le résultat principal.

Théorème 9.2 *Soit $R > 0, T > 0, \lambda > 0, \eta_p > 0, \mathbf{q} \in B(0,R)$ la condition initiale, et κ une matrice $\mathbb{R}^{d\times d}$ donnée. Si $(\mathbf{Q}_t, \mu_t)$ est la solution de (9.4), $\mathbf{Q}_t^\varepsilon$ est la solution de (9.6) et τ^ε est donné par (9.7), alors la limite $\tau := \lim_{\varepsilon\to 0}\tau_\varepsilon$ existe pour tout $t \in [0,T]$. De plus, $\tau(t)$ est un fonction continue de t qui peut être calculée par la formule suivante*

$$\tau = -\eta_p \frac{\partial \mathbb{E}(\mathbf{Q}_t \otimes \mathbf{Q}_t)}{\partial t} + \eta_p \mathbb{E}((\mathbf{Q}_t \otimes \mathbf{Q}_t)\kappa^T + \kappa(\mathbf{Q}_t \otimes \mathbf{Q}_t)). \tag{9.9}$$

Pour simplifier les notations, on va écrire $\Gamma = \kappa - \frac{1}{2\lambda}I$ et $\sigma^2 = 1/\lambda$. Le problème pénalisé (9.6) devient

$$d\mathbf{Q}_t^\varepsilon = \left(\Gamma\mathbf{Q}_t^\varepsilon - \frac{1}{\varepsilon}\beta(\mathbf{Q}_t^\varepsilon)\right)dt + \sigma d\mathbf{B}_t, \quad \mathbf{Q}_0^\varepsilon = \mathbf{q}. \tag{9.10}$$

Rappelons les résultats de [55] :

– Pour chaque $1 \leq p < \infty$, il existe une constant C indépendante de ε telle que

$$\mathbb{E}\left(\sup_{t\in[0,T]} |\mathbf{Q}_t^\varepsilon|^p\right) \leq C \tag{9.11}$$

– Pour chaque $1 \leq p < \infty$ et $0 < T < \infty$, on a

$$\lim_{\varepsilon \to 0} \mathbb{E}\left(\sup_{0 \leq t \leq T} |\mathbf{Q}_t - \mathbf{Q}_t^\varepsilon|^p \right) = 0. \tag{9.12}$$

On peut noter qu'une forme faible du théorème 9.2, c'est-à-dire la formule (9.9) intégrée en temps sur un intervalle $(0, T)$, est une conséquence immédiate de la formule d'Itô appliquée à $\mathbf{Q}_t^\varepsilon \otimes \mathbf{Q}_t^\varepsilon$ et de (9.12). La question à laquelle notre théorème répond est de savoir si il est possible d'échanger l'ordre de la limite et de la somme dans cette formule.

La preuve du théorème 9.2 sera présentée comme une suite de lemmes, le moins évident entre eux étant celui d'equi-continuité en t de $\frac{1}{\varepsilon}\mathbb{E}(\mathbf{Q}_t^\varepsilon \otimes \beta(\mathbf{Q}_t^\varepsilon))$ par rapport à ε. On montre, en effet, dans le lemme 9.6 (avec $\phi(\mathbf{Q}) = \mathbf{Q} \cdot \beta(\mathbf{Q})$) que

$$\lim_{h \to 0^+} \sup_{\varepsilon > 0} \sup_{0 \leq t \leq T-h} \mathbb{E}(\mathbf{Q}_{t+h}^\varepsilon \otimes \frac{1}{\varepsilon}\beta(\mathbf{Q}_{t+h}^\varepsilon) - \mathbf{Q}_t^\varepsilon \otimes \frac{1}{\varepsilon}\beta(\mathbf{Q}_t^\varepsilon)) = 0. \tag{9.13}$$

Remarque 9.3 *Il est intéressant d'observer que ce résultat serait faux sans le mouvement brownien. Pour voir un contre-exemple, considérons le cas uni-dimensionnel* $(d = 1)$ *de (9.10) avec* $R = 1$, $\Gamma = \gamma \in \mathbb{R}$, $\gamma > 0$, $\sigma = 0$ *et* $x \in]0, 1[$. *La solution pour* $\varepsilon < 1/\gamma$ *prend la forme*

$$\mathbf{Q}_t^\varepsilon = \begin{cases} xe^{\gamma t}, & t < \tilde{t}, \\ \left(1 - \frac{1}{1-\varepsilon\gamma}\right) e^{(\gamma - 1/\varepsilon)(t-\tilde{t})} + \frac{1}{1-\varepsilon\gamma}, & t \geq \tilde{t}, \end{cases}$$

où $\tilde{t} = \frac{1}{\gamma} \ln\left(\frac{1}{x}\right)$. *Fixons* $h > 0$, *alors*

$$\frac{1}{\varepsilon}\mathbf{Q}_{\tilde{t}+h}^\varepsilon \beta(\mathbf{Q}_{\tilde{t}+h}^\varepsilon) - \frac{1}{\varepsilon}\mathbf{Q}_{\tilde{t}}^\varepsilon \beta(\mathbf{Q}_{\tilde{t}}^\varepsilon) = \frac{1}{\varepsilon}(\mathbf{Q}_{\tilde{t}+h}^\varepsilon - \mathbf{Q}_{\tilde{t}}^\varepsilon)(\mathbf{Q}_{\tilde{t}+h}^\varepsilon + \mathbf{Q}_{\tilde{t}}^\varepsilon - 1)$$

Puisque $\lim_{\varepsilon \to 0} \frac{1}{\varepsilon}(\mathbf{Q}_{\tilde{t}+h}^\varepsilon - \mathbf{Q}_{\tilde{t}}^\varepsilon) = \gamma$, $\lim_{\varepsilon \to 0}(\mathbf{Q}_{\tilde{t}+h}^\varepsilon + \mathbf{Q}_{\tilde{t}}^\varepsilon) = 2$, *on voit que*

$$\lim_{h \to 0^+} \sup_{\varepsilon > 0} \sup_{0 \leq t \leq T-h} \left| \frac{1}{\varepsilon}\mathbf{Q}_{t+h}^\varepsilon \beta(\mathbf{Q}_{t+h}^\varepsilon) - \frac{1}{\varepsilon}\mathbf{Q}_t^\varepsilon \beta(\mathbf{Q}_t^\varepsilon) \right| \geq \gamma > 0.$$

Sans perdre de généralité, on va supposer dorénavant $R = 1$ et $\sigma = 1$.

On établit d'abord un principe de comparaison pour les solutions de (9.10). On considère

$$d\mathbf{R}_t^\varepsilon = (\gamma \mathbf{R}_t^\varepsilon - \frac{1}{\varepsilon}\beta(\mathbf{R}_t^\varepsilon))dt + d\mathbf{B}_t, \qquad \mathbf{R}_0^\varepsilon = \mathbf{r}, \tag{9.14}$$

avec $\Gamma = \gamma I$, $\gamma \in \mathbb{R}$, $y \in B(0, 1)$ et

$$\gamma = \sup_{|\mathbf{Q}|=1} \mathbf{Q} \cdot \Gamma \mathbf{Q}.$$

Lemme 9.4 *Si* $\mathbf{Q}_t^\varepsilon$ *et* $\mathbf{R}_t^\varepsilon$ *sont les solutions de (9.10) et (9.14) respectivement, avec* $|\mathbf{q}| \leq |\mathbf{r}|$ *et* $g : [0, \infty[\to \mathbb{R}$ *est une fonction croissante, alors*

$$\mathbb{E}(g(|\mathbf{Q}_t^\varepsilon|)) \leq \mathbb{E}^y(g(|\mathbf{R}_t^\varepsilon|)), \ \forall 0 \leq t \leq T. \tag{9.15}$$

Démonstration. Comme le lemme porte sur le comportement des composantes radiales des solutions, l'idée de sa preuve est de bien choisir les réalisations des mouvements browniens dans les équations stochastiques pour synchroniser les composantes du bruit dans les directions de $\mathbf{Q}_t^\varepsilon$ et $\mathbf{R}_t^\varepsilon$. On obtiendra ainsi des équation quasi unidimensionnelles pour $|\mathbf{Q}_t^\varepsilon|$ et $|\mathbf{R}_t^\varepsilon|$ et il sera facile de comparer leurs solutions. Pour concrétiser cette idée, introduisons pour tout $\mathbf{Q} \in \mathbb{R}^d$ la matrice orthogonale $S(\mathbf{Q})$ telle que $S\mathbf{Q} = |\mathbf{Q}|\boldsymbol{e}_1$ et $S\mathbf{X} = \mathbf{X}$ pour tous les vecteurs $\mathbf{X}$ orthogonaux au plan $(\mathbf{Q}, \boldsymbol{e}_1)$.

Considérons les mouvements Browniens $\widetilde{\boldsymbol{B}}_t = \int_0^t S(\mathbf{Q}_s)d\boldsymbol{B}_s$ et $\widetilde{\widetilde{\boldsymbol{B}}}_t = \int_0^t S(\mathbf{R}_s)d\boldsymbol{B}_s$. La substitution dans (9.10) et (9.14) donne

$$d\mathbf{Q}_t^\varepsilon = (\Gamma\mathbf{Q}_t^\varepsilon - \frac{1}{\varepsilon}\beta(\mathbf{Q}_t^\varepsilon))dt + S^T(\mathbf{Q}_t)d\widetilde{\boldsymbol{B}}_t, \tag{9.16}$$

$$d\mathbf{R}_t^\varepsilon = (\gamma\mathbf{R}_t^\varepsilon - \frac{1}{\varepsilon}\beta(\mathbf{R}_t^\varepsilon))dt + S^T(\mathbf{R}_t)d\widetilde{\widetilde{\boldsymbol{B}}}_t. \tag{9.17}$$

Comme l'énoncé du lemme ne parle que des espérances de $g(|\mathbf{Q}_t^\varepsilon|)$ et $g(|\mathbf{R}_t^\varepsilon|)$, on peut changer librement les réalisations des mouvements Browniens dans les équations ci-dessus. On choisit donc $\widetilde{\boldsymbol{B}}_t = \widetilde{\widetilde{\boldsymbol{B}}}_t = \mathbf{B}_t = (\mathbf{B}_t^1, \dots, \mathbf{B}_t^d)$. La formule d'Itô donne

$$d|\mathbf{Q}_t^\varepsilon|^2 = 2(\mathbf{Q}_t^\varepsilon \cdot \Gamma\mathbf{Q}_t^\varepsilon - \mathbf{Q}_t^\varepsilon \cdot \frac{1}{\varepsilon}\beta(\mathbf{Q}_t^\varepsilon) + \frac{d}{2})dt + 2|\mathbf{Q}_t^\varepsilon|dB_t^1, \tag{9.18}$$

$$d|\mathbf{R}_t^\varepsilon|^2 = 2(\gamma|\mathbf{R}_t^\varepsilon|^2 - \mathbf{R}_t^\varepsilon \cdot \frac{1}{\varepsilon}\beta(\mathbf{R}_t^\varepsilon) + \frac{d}{2})dt + 2|\mathbf{R}_t^\varepsilon|dB_t^1. \tag{9.19}$$

Il est facile de voir maintenant que chaque trajectoire de $|\mathbf{Q}_t^\varepsilon|^2$ reste toujours à gauche de celle de $|\mathbf{R}_t^\varepsilon|^2$, prise sur la même trajectoire du mouvement brownien, ce qui prouve (9.15). En effet, si $|\mathbf{Q}_t^\varepsilon|^2 = |\mathbf{R}_t^\varepsilon|^2$ en un certain moment $t = t_0$, le coefficient de dérive dans (9.19) est plus grand que celui dans (9.18) à ce moment-là, et donc on va avoir $|\mathbf{Q}_t^\varepsilon|^2 < |\mathbf{R}_t^\varepsilon|^2$ immédiatement après t_0. Pour une démonstration plus rigoureuse, on renvoie à [50, par. 5, proposition 2.18] et [24]. ■

Ce principe de comparaison permet à donner une majoration uniforme en ε pour le tenseur pénalisé.

Lemme 9.5 *Pour tout $r \in [0, 1)$ il existe existe une constante $C(r)$ telle que si $\mathbf{Q}_t^\varepsilon$ est la solution de (9.10) avec la condition initiale $\mathbf{q} \in B(0, r)$, alors pour tout $t \in [0, T]$*

$$\mathbb{E}^x(\mathbf{Q}_t^\varepsilon \cdot \frac{1}{\varepsilon}\beta(\mathbf{Q}_t^\varepsilon)) \le C(r),$$

La constante $C(r)$ ne dépend que de r et γ.

Démonstration. Soit π_ε la distribution stationaire de (9.14) et $\pi_\varepsilon \cap I_+$ la distribution π_ε conditionnée sur l'ensemble $I_+ = \{z \in \mathbb{R}^d : |z| > r\}$. Le lemme 9.4 avec $g(\mathbf{Q}) = \mathbf{Q} \cdot \frac{1}{\varepsilon}\beta(\mathbf{Q})$ implique

$$\mathbb{E}^x(\mathbf{Q}_t^\varepsilon \cdot \frac{1}{\varepsilon}\beta(\mathbf{Q}_t^\varepsilon)) \le \mathbb{E}^{\pi_\varepsilon \cap I_+}(\mathbf{R}_t^\varepsilon \cdot \frac{1}{\varepsilon}\beta(\mathbf{R}_t^\varepsilon)) \le \frac{\mathbb{E}^{\pi_\varepsilon}(\mathbf{R}_t^\varepsilon \cdot \frac{1}{\varepsilon}\beta(\mathbf{R}_t^\varepsilon))}{\pi_\varepsilon(I_+)} \tag{9.20}$$

La règle d'Itô pour $|\mathbf{R}_t^\varepsilon|^2$ donne

$$\mathbb{E}^{\pi_\varepsilon}(\mathbf{R}_t^\varepsilon \cdot \frac{1}{\varepsilon}\beta(\mathbf{R}_t^\varepsilon)) = \gamma\mathbb{E}^{\pi_\varepsilon}(|\mathbf{R}_t^\varepsilon|^2) + \frac{d}{2}. \tag{9.21}$$

Il est facile de trouver une formule explicite pour la distribution stationnaire $\pi_\varepsilon(\mathbf{R}) = \frac{1}{C_\varepsilon}\phi_\varepsilon(\mathbf{R})$ avec

$$\phi_\varepsilon(\mathbf{R}) = \begin{cases} e^{\gamma|\mathbf{R}|^2}, \ |\mathbf{R}| < 1 \\ e^{\gamma|\mathbf{R}|^2 - \frac{(|\mathbf{R}|-1)^2}{\varepsilon}}, \ \text{sinon} \end{cases}$$

et $C_\varepsilon = \int_{\mathbb{R}^d} \phi_\varepsilon(\mathbf{R})d\mathbf{R}$. En appliquant le théorème de la convergence dominée, on voit que les limites $\lim_{\varepsilon\to 0}\mathbb{E}^{\pi_\varepsilon}(|\mathbf{R}_t^\varepsilon|^2)$ et $\lim_{\varepsilon\to 0}\pi_\varepsilon(I^+)$ existent et sont finis. Cela implique le résultat du lemme en combinant les inégalités (9.20) et (9.21). ■

On peut maintenant montrer le résultat d'équi-continuité (9.13) que l'on énonce de manière un peu plus générale :

Lemme 9.6 *Soit $\mathbf{Q}_t^\varepsilon$ la solution de (9.10) avec $\mathbf{q} \in B(0,1)$ et soit $\phi : \mathbb{R}^d \to \mathbb{R}^+$ telle que $\phi(z) = 0$ for all $z \in B(0,1)$. On suppose que pou tout $t \in [0,T]$*

$$\mathbb{E}\left|\frac{1}{\varepsilon}\phi(\mathbf{Q}_t^\varepsilon)\right| \leq C(|x|) \tag{9.22}$$

avec $C(|x|)$ indépendante de ε. Alors,

$$\lim_{h\to 0^+}\sup_{\varepsilon>0}\sup_{0\leq t\leq T-h}\left|\mathbb{E}\left(\frac{1}{\varepsilon}(\phi(\mathbf{Q}_{t+h}^\varepsilon) - \phi(\mathbf{Q}_t^\varepsilon))\right)\right| = 0. \tag{9.23}$$

Démonstration. Fixons un nombre r tel que $|\mathbf{q}| < r < 1$ et $\alpha \in (0,1)$. D'après la propriété de Markov, on a pour tout $h > 0$, $t \geq h^\alpha$

$$\begin{aligned}\mathbb{E}\left(\frac{1}{\varepsilon}\phi(\mathbf{Q}_t^\varepsilon)\right) &= \int_{B(0,r)} p_{h^\alpha}^{\mathbf{Q}^\varepsilon}(z)\mathbb{E}^z\left(\frac{1}{\varepsilon}\phi(\mathbf{Q}_{t-h^\alpha}^\varepsilon)\right)dz \\ &\quad + \int_{\mathbb{R}^d\setminus B(0,r)} p_{h^\alpha}^{\mathbf{Q}^\varepsilon}(z)\mathbb{E}^z\left(\frac{1}{\varepsilon}\phi(\mathbf{Q}_{t-h^\alpha}^\varepsilon)\right)dz \end{aligned} \tag{9.24}$$

où p_a^Z désigne la densité de probabilité du processus Z au temps a et $\mathbb{E}^z$ désigne l'espérance en supposant que le processus concerné est parti en $t = 0$ du point z. Pour estimer la seconde intégrale ci-dessus, on considère le temps d'arrêt $\sigma = \inf\{s > 0 : |\mathbf{Q}_s^\varepsilon| = r\}$. Cette intégrale peut alors être majorée en utilisant l'hypothèse (9.22) :

$$\begin{aligned}\int_{\mathbb{R}^d\setminus B(0,r)} p_{h^\alpha}^{\mathbf{Q}^\varepsilon}(z)\mathbb{E}^z\left(\frac{1}{\varepsilon}\phi(\mathbf{Q}_{t-h^\alpha}^\varepsilon)\right)dz &= \mathbb{E}\left(\frac{1}{\varepsilon}\phi(\mathbf{Q}_t^\varepsilon)1_{|\mathbf{Q}_{h^\alpha}^\varepsilon|\geq r}\right) \leq \mathbb{E}\left(\frac{1}{\varepsilon}\phi(\mathbf{Q}_t^\varepsilon)1_{\sigma\leq h^\alpha}\right) \\ &= \mathbb{E}^x\left(\mathbb{E}^{\mathbf{Q}_\sigma^\varepsilon}\left(\frac{1}{\varepsilon}\phi(\mathbf{Q}_{t-\sigma}^\varepsilon)\right)1_{\sigma\leq h^\alpha}\right) \leq C(r)\mathcal{P}(\sigma\leq h^\alpha). \end{aligned} \tag{9.25}$$

Puisque l'événement $\sigma\leq h^\alpha$ ne concerne que les partie initiales de trajectoires de $\mathbf{Q}_t$ avant qu'il sorte de la boule unité, la probabilité $P(\sigma\leq h^\alpha)$ ne dépend pas de ε et, en plus, elle tend vers zéro lorsque $\varepsilon \to 0$.

On écrira $o(1)$ pour toutes les quantités qui peuvent être majorées par une fonction de h qui est indépendante de ε et qui tend vers zéro lorsque $h \to 0$. On a montré donc

$$\mathbb{E}\left(\frac{1}{\varepsilon}\phi(\mathbf{Q}_t^\varepsilon)\right) = \int_{B(0,r)} p_{h^\alpha}^{\mathbf{Q}^\varepsilon}(z)\mathbb{E}^z\left(\frac{1}{\varepsilon}\phi(\mathbf{Q}_{t-h^\alpha}^\varepsilon)\right)dz + o(1).$$

Par des raisonnements similaires, on peut aussi montrer que

$$\mathbb{E}\left(\frac{1}{\varepsilon}\phi(\mathbf{Q}_t^\varepsilon)\right) \;=\; \int_{B(0,r)} p_{h^\alpha}^{\mathbf{Q}}(z)\mathbb{E}^z\left(\frac{1}{\varepsilon}\phi(\mathbf{Q}_{t-h^\alpha}^\varepsilon)\right)dz + o(1), \tag{9.26}$$

où $\mathbf{Q}_t$ est le processus qui ne voit pas la boule :

$$d\mathbf{Q}_t \;=\; \Gamma\mathbf{Q}_t dt + dB_t : \qquad \mathbf{Q}_0 \;=\; x.$$

On peut maintenant écrire (9.26) en temps t et en temps $t+h$ et prendre la différence entre les deux :

$$\begin{aligned}
&\left|\mathbb{E}\left(\frac{1}{\varepsilon}\phi(\mathbf{Q}_{t+h}^\varepsilon)\right) - \mathbb{E}\left(\frac{1}{\varepsilon}\phi(\mathbf{Q}_t^\varepsilon)\right)\right| \\
&\quad= \left|\int_{B(0,r)} (p_{h^\alpha+h}^{\mathbf{Q}}(z) - p_{h^\alpha}^{\mathbf{Q}}(z))\mathbb{E}^z\left(\frac{1}{\varepsilon}\phi(\mathbf{Q}_{t-h^\alpha}^\varepsilon)\right)dz\right| + o(1) \\
&\quad\le C(r)\int_{B(0,r)} |p_{h^\alpha+h}^{\mathbf{Q}}(z) - p_{h^\alpha}^{\mathbf{Q}}(z)|dz + o(1).
\end{aligned}$$

La preuve du lemme est ainsi réduite à une estimation de $|p_{h^\alpha+h}^{\mathbf{Q}}(z) - p_{h^\alpha}^{\mathbf{Q}}(z)|$ (noter que cette expression ne dépend pas de ε). Pour cela, on peut montrer par un calcul direct, détaillé en [24], que

$$|p_{h^\alpha+h}^{\mathbf{Q}}(z) - p_{h^\alpha}^{\mathbf{Q}}(z)| \le Ch^\delta \;\text{ pour } z\in B(0,1),\; 0\le h\le 1 \tag{9.27}$$

à condition que $0<\alpha<\frac{2}{d+2}$ et $\delta = 1-\frac{\alpha(d+2)}{2} > 0$. La constante C ne dépend que de Γ et de α. ■

Démonstration du théorème 9.2 Prenons $\mathbf{q}\in B(0,1)$ et $t\in[0,T]$. On a les estimations suivantes pour la solution $\mathbf{Q}_t^\varepsilon$ de (9.10) pour i,j tels que $1\le i,j\le d$

$$\mathbb{E}\left(\left|(\mathbf{Q}_t^\varepsilon)_i\frac{1}{\varepsilon}\beta(\mathbf{Q}_t^\varepsilon)_j\right|\right) \le C(|\mathbf{q}|)$$

avec $C(|\mathbf{q}|)$ indépendante de ε. Soit $(\varepsilon_n)_{n\ge1}\subset\mathbb{R}$ une suite avec $\lim_{n\to\infty}\varepsilon_n = 0$ et $1\le i,j\le d$. On désigne par $f_n(t) = \frac{1}{\varepsilon_n}\mathbb{E}\left((\mathbf{Q}_t^{\varepsilon_n})_i(\beta(\mathbf{Q}_t^{\varepsilon_n}))_j\right)$ la composante ij du tenseur $\frac{1}{\varepsilon_n}\mathbb{E}\left(\mathbf{Q}_t^{\varepsilon_n}\otimes\beta(\mathbf{Q}_t^{\varepsilon_n})\right)$. La formule d'Itô montre que $\forall t\in[0,T]$ et $\forall n\ge1$

$$\int_0^t f_n(s)ds = -\frac{1}{2}\big(\mathbb{E}\big((\mathbf{Q}_t^{\varepsilon_n})_i(\mathbf{Q}_t^{\varepsilon_n})_j\big) - x_ix_j\big) + \frac{1}{2}\left(\int_0^t \mathbb{E}(\mathbf{Q}_s^{\varepsilon_n}\otimes\mathbf{Q}_s^{\varepsilon_n})\Gamma^T + \Gamma(\mathbf{Q}_s^{\varepsilon_n}\otimes\mathbf{Q}_s^{\varepsilon_n}))ds\right)_{ij} + \frac{1}{2}t\delta_{ij}.$$

Soit

$$g(t) = -\frac{1}{2}\big(\mathbb{E}\big((\mathbf{Q}_t)_i(\mathbf{Q}_t)_j\big) - x_ix_j\big) + \frac{1}{2}\left(\int_0^t \mathbb{E}(\mathbf{Q}_s\otimes\mathbf{Q}_s)\Gamma^T + \Gamma(\mathbf{Q}_s\otimes\mathbf{Q}_s))ds\right)_{ij} + \frac{1}{2}t\delta_{ij}.$$

En utilisant (9.12), on obtient

$$\lim_{n\to\infty}\int_0^t f_n(s)ds = g(t). \tag{9.28}$$

De l'autre côté, d'après les lemmes 9.5 et 9.6, la suite $(f_n(t))_{n\ge1}$ est bornée et équi-continue. Par le théorème d'*Ascoli-Arzelà*, il existe une sous-suite qui converge dans $\mathcal{C}^0([0,T];\mathbb{R})$ vers la limite $f\in\mathcal{C}^0([0,T];\mathbb{R})$. La convergence uniforme implique $\int_0^t f(s)ds = g(t)$ et donc $f(t) = g'(t)\;\forall t\in[0,T]$. Comme ce raisonnement ne dépend pas du choix de la suite $(\varepsilon_n)_{n\ge1}$, cela prouve (9.9) sur $[0,T]$.

9.3 Simulations numériques

Dans cette section, nous présentons quelques schémas numériques pour le modèle (9.4) conduisant à une approximation du tenseur τ donné par (9.9) ou comme la limite $\lim_{\varepsilon\to 0}\tau^{\varepsilon}$. Comme aucune analyse théorique n'a été fait pour les approximations numériques stochastique, on va utiliser comme solution de référence celle de la méthode EDP basée sur l'équation de Fokker-Planck.

Les résultats présentés ci-dessous correspondent à la situation physiques suivante : on considère l'écoulement de cisaillement en 2D avec

$$\kappa = \begin{pmatrix} 0 & \dot{\gamma} \\ 0 & 0 \end{pmatrix}.$$

La distribution initiale de $\mathbf{Q}$ est prise, dans toutes nos expériences, comme la distribution stationnaire du modèle (9.10) avec $\kappa = 0$, autrement dit la distribution normale $\mathcal{N}(0,1)$ restreinte à la boule $\overline{B(0,R)}$.

La méthode Fokker-Planck

Soit $\psi(t,\mathbf{Q})$, $\mathbf{Q}\in B$ ($B = B(0,R)$) la densité de probabilité du processus $\mathbf{Q}_t$. Il est bien connu que ψ vérifie l'équation de Fokker-Planck

$$\frac{\partial\psi}{\partial t} + \mathrm{div}_{\mathbf{Q}}\left(\left(\kappa\mathbf{Q} - \frac{1}{2\lambda}\mathbf{Q}\right)\psi\right) = \frac{1}{2\lambda}\Delta_{\mathbf{Q}}\psi \tag{9.29}$$

avec les condition aux limites de flux zero :

$$\left(-\kappa\mathbf{Q}\psi + \frac{1}{2\lambda}\mathbf{Q}\psi + \frac{1}{2\lambda}\nabla_{\mathbf{Q}}\psi\right)\cdot\mathbf{n}|_{\partial B} = 0. \tag{9.30}$$

L'expression pour le tenseur τ (9.9) peut être réécrite en termes de densité de probabilité et en utilisant (9.29)–(9.30)

$$\begin{aligned}\tau &= -\eta_p\frac{\partial}{\partial t}\int_B \mathbf{Q}\otimes\mathbf{Q}\psi d\mathbf{Q} + \eta_p\int_B\left(\kappa\mathbf{Q}\otimes\mathbf{Q} + \mathbf{Q}\otimes\mathbf{Q}\kappa^T\right)\psi d\mathbf{Q}\\ &= \frac{\eta_p}{\lambda}\int_B \mathbf{Q}\otimes\mathbf{Q}\psi d\mathbf{Q} + \frac{\eta_p}{\lambda R}\int_{\partial B}\mathbf{Q}\otimes\mathbf{Q}\psi ds - \frac{\eta_p}{\lambda}I\end{aligned}$$

On utilise une méthode numérique spectrale en passant aux coordonnées polaires $r\in[0,R]$, $\theta\in[0,2\pi)$ et en approchant ψ par les polynômes en r (sur une grille des points de collocation de Gauss-Legendre-Radau) multipliés par des polynômes trigonométriques en θ. Le système d'équations linéaires est obtenue par la méthode de Galerkin. Il est ensuite discrétisé en temps par le schéma implicite d'Euler. Les résultats numériques sont présentés en Fig. 9.1. Ils montrent que la méthode converge par rapport au raffinement dans l'espace-temps. On prend donc la solution numérique au niveau le plus raffiné comme référence pour les suite simulations suivantes.

	N_R	N_F	Δt
Level 3	8	4	0.1×2^{-3}
Level 4	12	8	0.1×2^{-4}
Level 5	20	12	0.1×2^{-5}
Level 6	26	16	0.1×2^{-6}

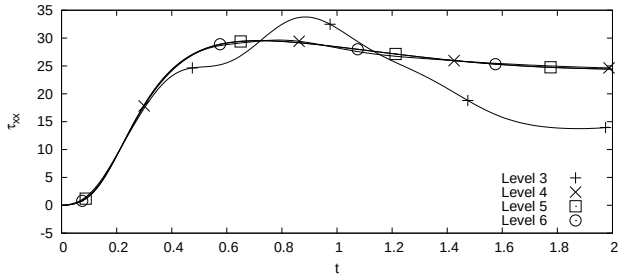


FIGURE 9.1: La partie gauche de la figure décrit les paramètres numériques correspondant aux niveaux de discrétisation différents. La partie droite de la figure présente l'évolution des τ_{xx} (Top : $t \in [0, 2]$, Bottom : $t \in [1.4, 2]$).

Simulations stochastiques

On considère deux méthodes de discrétisation de (9.4) en temps. Introduisons la grille uniforme en temps $t_n = n\Delta t$ et désignerons une approximation de $\mathbf{Q}_{t_n}$ par $\mathbf{Q}_n^{\Delta t}$. Le nombre de réalisations aléatoires utilisée pour calculer l'espérance est noté M. Pour toutes les simulations, nous imposons la relation $M \propto (\Delta t)^2$ pour assurer que le bruit stochastique n'affecte pas l'ordre de convergence en ce qui concerne le pas de temps.

La méthode de pénalisation

La première méthode est l'algorithme par pénalisation qui est une discrétisation directe de l'EDS pénalisée (9.6)

$$\mathbf{Q}_{n+1}^{\Delta t} - \mathbf{Q}_n^{\Delta t} = \left(\kappa(t_n)\mathbf{Q}_n^{\Delta t} - \frac{1}{2\lambda}\mathbf{Q}_n^{\Delta t} - \frac{1}{2\lambda\varepsilon}\beta(\mathbf{Q}_n^{\Delta t}) \right) \Delta t + \sqrt{\frac{1}{\lambda}}(B_{t_{n+1}} - B_{t_n}). \tag{9.31}$$

Le tenseur τ au temps t_n est alors approché par

$$\tau_n^{\Delta t} = \frac{\eta_p}{\lambda}\mathbb{E}\left(\mathbf{Q}_n^{\Delta t} \otimes \mathbf{Q}_n^{\Delta t} + \frac{1}{\varepsilon}\mathbf{Q}_n^{\Delta t} \otimes \beta(\mathbf{Q}_n^{\Delta t}) \right) - \frac{\eta_p}{\lambda} I. \tag{9.32}$$

Le choix du paramètre de pénalisation ε est une question délicate. D'après [59], le schéma (9.31) converge faiblement à l'ordre $\sqrt{\Delta t}$ si $\varepsilon \geq \Delta t$. Ce résultat n'est pas suffisant pour assurer la convergence de l'approximation du tenseur des contraintes (9.32). Toutefois, on a observé expérimentalement que la convergence de τ est d'ordre Δt en norme $L^2(0, T)$ si on pose $\varepsilon = \Delta t$. On présente en Fig. 9.2 l'évolution du tenseur des contraintes et de sa variance qui est estimée sur l'ensemble de 100 simulations indépendantes. On observe clairement la convergence de la solution par pénalisation vers la solution de référence en Fig. 9.2.

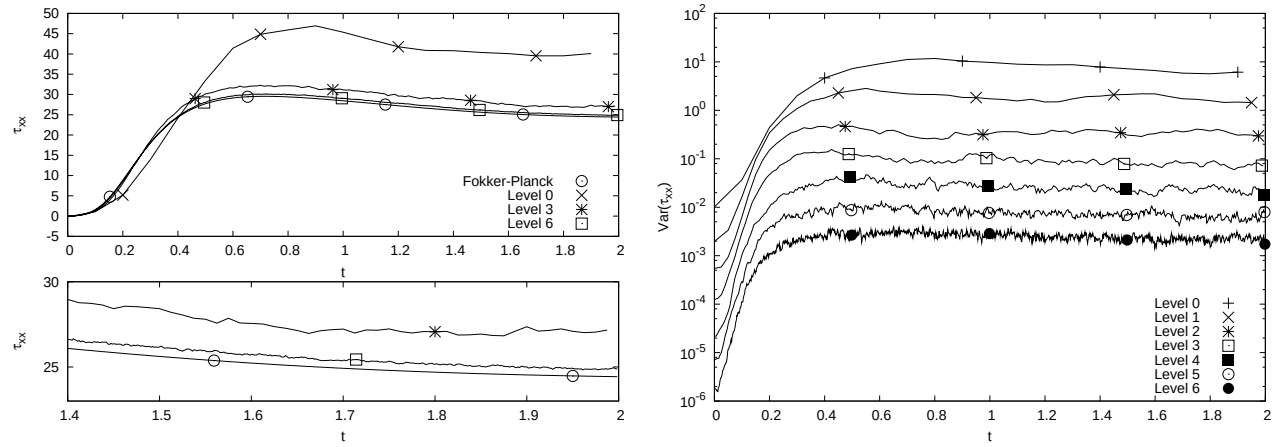


FIGURE 9.2: La méthode de pénalisation dans l'écoulement de cisaillement avec $\dot{\gamma} = 10$. On considère sept niveaux de discrétisation $L \in \{0, 1, 2, 3, 4, 5, 6\}$. Les paramètres numériques à chaque niveau L sont $\Delta t_L = 0.1 \times 2^{-L}$ et $M_L = 250 \times 4^L$. A gauche : l'évolution de $\tau_{x\mathbf{Q}}$; en bas : un zoom sur l'intervalle $t \in [1.4, 2]$) ; à droite : l'évolution de la variance.

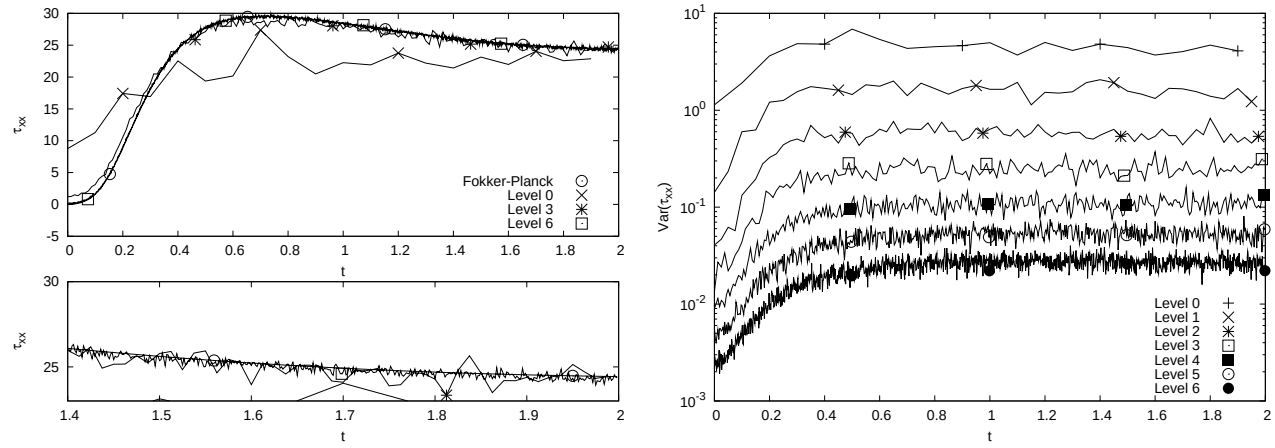


FIGURE 9.3: La méthode de réflexion symétrique dans l'écoulement de cisaillement avec $\dot{\gamma} = 10$ et les mêmes paramètres de discrétisation qu'en Fig. 9.2.

La méthode de réflexion symétrique

La seconde méthode est basée directement sur l'EDS réfléchie (9.4). L'idée est de faire une réflexion par rapport à la frontière ∂B lorsque le processus la touche [38]. Chaque pas de cet algorithme est composé de deux sous-pas suivants :

– On fait avancer d'abord le processus stochastique sans prendre en compte la force de réflexion sur la frontière $\partial B(0, R)$:

$$\tilde{\mathbf{Q}}_{n+1}^{\Delta t} - \mathbf{Q}_n^{\Delta t} = \left(\kappa(t_n)\mathbf{Q}_n^{\Delta t} - \frac{1}{2\lambda}\mathbf{Q}_n^{\Delta t}\right)\Delta t + \sqrt{\frac{1}{\lambda}}(B_{t_{n+1}} - B_{t_n}) \tag{9.33}$$

– S'il n'y a pas de réflexion, c'est-à-dire $|\tilde{\mathbf{Q}}_{n+1}^{\Delta t}| \leq R$, on pose $\mathbf{Q}_{n+1}^{\Delta t} = \tilde{\mathbf{Q}}_{n+1}^{\Delta t}$. Sinon, on prend $\mathbf{Q}_{n+1}^{\Delta t}$ comme l'image du miroir de $\tilde{\mathbf{Q}}_{n+1}^{\Delta t}$ par rapport à $\partial B(0, R)$, c'est-à-dire

$$\mathbf{Q}_{n+1}^{\Delta t} = \frac{2R - |\tilde{\mathbf{Q}}_{n+1}^{\Delta t}|}{|\tilde{\mathbf{Q}}_{n+1}^{\Delta t}|}\tilde{\mathbf{Q}}_{n+1}^{\Delta t}$$

Ce schéma converge faiblement à l'ordre Δt [38]. Le tenseur des contraintes est définie par l'expression (9.9). La discrétisation de la dérivée en temps par une différence finie donne l'approximation

$$\begin{aligned}\tau_n^{\Delta t} = &-\frac{\eta_p}{\Delta t}\mathbb{E}\left(\mathbf{Q}_{n+1}^{\Delta t} \otimes \mathbf{Q}_{n+1}^{\Delta t} - \mathbf{Q}_n^{\Delta t} \otimes \mathbf{Q}_n^{\Delta t}\right) \\ &+ \eta_p \mathbb{E}(\kappa(t_n)\mathbf{Q}_n^{\Delta t} \otimes \mathbf{Q}_n^{\Delta t} + \mathbf{Q}_n^{\Delta t} \otimes \mathbf{Q}_n^{\Delta t}\kappa^T(t_n)).\end{aligned} \tag{9.34}$$

Les résultats obtenus par cette méthode sont présentés en Fig. 9.3. On observe toujours la convergence vers la solution de référence.

En résumé, les deux méthodes stochastiques donnent des approximation de τ avec une erreur d'ordre Δt. Les résultats de la méthode de réflexion symétrique sont plus précis que ceux de la méthode de pénalisation. Par contre, la méthode de pénalisation donne une approximation de τ beaucoup moins bruitée que l'autre méthode.

10 Perspectives

Les travaux que nous venons de présenter dans cette partie pourront être poursuivis dans plusieurs directions :

- Inclusion de la balance d'énergie dans le modèle original avec inertie de billes. En effet, pour l'instant, nous avons considéré seulement la balance du moment linéaire en supposant que le solvant est isotherme. L'inclusion d'énergie permettra de considérer rigoureusement les situation où la température (et donc l'intensité des impacts browniens) peut varier en espace et en temps. Le passage à la limite $m \to 0$ nous obligera alors à considérer des termes d'ordre ε^2, mais la technique des développements asymptotiques sera toujours applicable.
- Expériences numériques plus poussés pour étudier les prévisions rhéologiques des modèles avec inertie et avec la force réfléchie.
- Applications possibles de ces deux modèles pur raffiner la modélisation de globules rouges par des dumbbells en écoulements sanguins, comme proposé dans [58]

Troisième partie

Autres travaux

11 Méthodes numériques pour les écoulements biphasiques

11.1 Sur la discrétisation en temps en simulation des écoulements avec des particules

D'après l'article [27] inclus en annexe.

Une de difficultés liées aux simulations des écoulements des fluides avec des particules rigides consiste à prendre en compte les configurations où les deux particules (ou une particule et la paroi) sont presque en contact. En effet, en raison de forces de lubrification, une particule lisse enserrée dans un fluide visqueux ne peut pas toucher le mur en un temps fini. Un résultat similaire est valable aussi dans le cas de particules multiples. Une attention particulière doit être donc portée à la discrétisation pour bien prendre en compte les forces de lubrification d'une amplitude énorme et empêcher le contact artificiel entre les particules. En effet, en présence de force de lubrification, les équations régissant le mouvement d'une particule nécessitent des simulations avec un pas de temps minuscule afin d'obtenir une solution acceptable physiquement. Pour contourner ce problème, on a présenté dans [27] une idée qui consiste à interdire à la particule de s'approcher trop près de la paroi lors d'une simulation. On choisit donc un seuil q_s pour la distance q entre la particule et la paroi et on remplace la vraie trajectoire de la particule, dès qu'elle touche le seuil, par une approximation dans laquelle la distance q est maintenue égale à q_s jusqu'à un possible rebond de la particule sur la paroi. Le moment du rebond est prédit en utilisant une quantité auxiliaire (une approximation grossière de la vitesse) qui est calculée tout au long de l'intervalle de temps où la particule est bloquée à la distance q_s de la paroi. Cette approche rappelle le modèle de "gluey particles" de [54, 52] où q_s est effectivement mis à zéro et on considère la limite de viscosité évanescente. Cependant, les motivations dans [27] sont tout à fait différente de celles derrière le modèle de "gluey particles". Ce dernier est conçu comme une alternative simple aux équations standards de type Navier-Stokes, éventuellement corrigées pour tenir compte de la rugosité de surface de la particule. Par contre, notre approche est de prendre les équations classiques telles quelles et de fournir un outil pour leur discrétisation en temps robuste.

Cela signifie, en particulier, que l'on construit une méthode assez précise et valable pour n'importe quelle valeur de la viscosité, pas nécessairement petite. A noter également, que le seuil q_s dépend généralement du pas de discrétisation en temps. C'est donc un dispositif purement numérique qui n'a aucun sens physique.

11.2 Une méthode du type domaines fictifs pour des écoulements avec des bulles de gaz

D'après l'article [20] inclus en annexe.

Ce travail concerne la modélisation et la simulation numérique du mouvement des bulles de gaz dans un fluide. On représente les bulles par des corps déformables de masse négligeable dont la position et la forme sont décrites par un petit nombre de paramètres qui évoluent au cours du temps. Le mouvement de ces corps est régi par les forces hydrodynamiques provenant du fluide ambiant et des forces extérieures telles que la gravitation. Pour chaque bulle, les équations d'équilibre des forces sont obtenues en exigeant que le travail des forces appliquées sur la surface des bulles soit nul pour tous les déplacements virtuels, c'est-à-dire pour ceux qui sont compatibles avec les contraintes imposées par la forme des bulles. Pour tenir compte numériquement de la présence d'une quantité importante de bulles, on utilise la technique des domaines fictifs [46], c'est-à-dire qu'on remplit le domaine occupé par les bulles par le fluide (fictif) identique au fluide ambiant et on résout les équations de Navier-Stokes pour une densité et une viscosité constantes dans le domaine tout entier. On impose l'égalité des vitesses du fluide et d'un corps déformable sur sa frontière par l'introduction de multiplicateurs de Lagrange. La méthode de jauge [43] est utilisée pour résoudre les équations de Navier-Stokes, ce qui permet de ne résoudre que des équations du type de Laplace à chaque pas de temps dans un schéma évolutif. Les calculs sont effectués sur une grille cartésienne, de sorte que les systèmes linéaires peuvent être résolus par FFT. On inclut la possibilité d'une coalescence des bulles en considérant celle-ci comme un processus se déroulant pendant un petit intervalle de temps. Pour le traitement numérique, on suppose que la coalescence commence lorsque deux bulles sont suffisamment proches et que la bulle qui est plus petite est "avalée" par la plus grande au cours de quelques pas de temps. On a considéré trois exemples de familles de forme de bulles, soit en 2D - des ellipses rigides (l'équilibre des forces se réduit dans ce cas aux équations d'Euler pour le corps rigide), des ellipses déformables dont l'aire est constante, ce qui permet d'avoir une contribution non triviale des forces de la tension superficielle dans les équations des forces, et une famille de formes déterminés par des polynômes d'ordre 4. Le logiciel développé permet de simuler des centaines des bulles dans ces deux situations. La méthode a été aussi appliquée à des ellipsoïdes rigides en 3D.

12 Méthodes numériques pour des problèmes fortement anisotropes

12.1 Un schéma du type "Asymptotic-Preserving" pour une équation de diffusion anisotrope

D'après l'article [28] inclus en annexe.

Des problèmes fortement anisotropes appariassent dans plusieurs applications physiques et industrielles. C'est le cas, par exemple, dans la physique des plasmas lorsqu'on doit traiter de fortes valeurs du champ magnétique. Dans le contexte d'un plasma faiblement ionisé quasi-neutre, l'équation de continuité, permettant de calculer le potentiel, contient une matrice de conductivité qui possède une très forte anisotropie dans la direction du champ magnétique. Noter que la direction peut varier en espace et en temps. Elle peut être donnée aussi par une résolution numérique d'autres équation du modèle. D'où la nécessité de développer des méthodes robustes qui fonctionnent bien pour n'importe quelle grandeur d'anisotropie et qui n'ont pas besoin d'adapter le maillage à la direction d'anisotropie.

On propose dans [28] une méthode pour une équation de diffusion anisotrope qui s'inscrit dans le cadre des schémas "Asymptotic Preserving" (AP). Cette technique est conçue pour des problèmes avec un petit paramètre ε, ce qui est notre cas, où la diffusion dans la direction d'anisotropie est d'ordre $1/\varepsilon$ et elle est d'ordre 1 dans les directions orthogonales. La méthode AP donne une solution précise pour toutes les valeurs de ε sur un maillage fixé. La dérivation d'une telle méthode exige d'identifier le modèle limite quand $\varepsilon \to 0$. On reformule alors le problème original de telle sorte que le système d'équations obtenu contient, à la fois, le modèle limite et le problème d'origine avec une transition continue par rapport à ε. Dans le cas de l'équation de diffusion anisotrope, la solution du problème limite est constante le long des lignes de champ (direction d'anisotropie). L'idée est donc de décomposer la solution ϕ pour $\varepsilon > 0$ en deux parties ($\phi = p + q$) : une partie moyenne p qui est constante le long des lignes de champ et une partie de fluctuation q qui donne une correction de la moyenne p. On utilise des maillages qui ne sont pas nécessairement adaptés à la direction

d'anisotropie. Cela est possible grâce à l'introduction de multiplicateurs de Lagrange assurant que la partie moyenne p est bien constante le long des lignes de champ.

Des cas plus complexes, quand l'anisotropie n'intervient que dans une partie du domaine du calcul, sont maintenant à l'étude.

12.2 Méthodes adaptatives sur des maillages anisotropes pour les problèmes non stationnaires

D'après l'article [23] inclus en annexe.

L'article [23] présente des nouveaux estimateurs d'erreur a posteriori pour l'équation de la chaleur $\partial u/\partial t - \Delta u = f$ discrétisée en temps par le schéma de Crank-Nicolson et en espace par des éléments finis continues linéaires sur des maillages hautement anisotropes. L'indicateur d'erreur due à la discrétisation en espace est obtenu à l'aide d'estimations d'interpolation anisotrope, tandis que l'indicateur d'erreur due à la discrétisation en temps est obtenu à l'aide d'une reconstruction quadratique par morceaux de la solution discrète, à partir d'une approximation par différences finies pour $\partial^2 u/\partial t^2$. L'utilisation de cette reconstruction dans l'estimation d'erreur est originale, même dans le cadre des éléments finis isotropes. Elle s'avère plus efficace qu'une autre reconstruction proposée antérieurement dans la littérature [34] qui était basée sur une approximation par différences finies pour $\partial f/\partial t + \Delta(\partial u/\partial t)$. Des expériences numériques confirment que l'indicateur d'erreur et la méthode adaptative proposés dans [23] sont d'ordre optimal par rapport aux pas de maillage en espace et en temps.

13 Méthodes numériques non probabilistes pour le modèle des chaînes

De nombreux phénomènes physiques, chimiques, biologiques et autres sont décrits mathématiquement par des EDP de très haute dimension. On peut penser, par exemples, à des réactions chimiques impliquant des dizaines des composants, à des fluctuations des marchés financiers etc. Un exemple typique est le modèle bille-ressort dans la théorie cinétiques des polymères (voir Fig. 8, à gauche). Une molécule de polymère linéaire y est représentée par une chaîne constituée de $(d+1)$ billes rejointes successivement par d ressorts sans masse. Notons que le modèle des dumbmells (Fig. 8, à droite) n'est qu'un cas particulier de celui-ci avec $d = 1$. Les chaînes de bille-ressort sont représentées par d vecteurs aléatoires $\boldsymbol{q}_j$. Leur dynamique dans un écoulement homogène est donnée soit par l'EDS du type (9.1) soit par l'équation de Fokker-Planck pour la fonction de densité de probabilité $\psi(t, \boldsymbol{q}_1, \ldots, \boldsymbol{q}_d)$:

$$\frac{\partial \psi}{\partial t} = \mathcal{L}^d_{FP}(\psi) = \sum_{k=1}^{d} L^k \psi - \frac{1}{4\lambda} \sum_{k=2}^{d} (N_x^{k-1} M_x^k + N_y^{k-1} M_y^k)\psi - \frac{1}{4\lambda} \sum_{k=1}^{d-1} (N_x^{k+1} M_x^k + N_y^{k+1} M_y^k)\psi. \tag{13.1}$$

où

$$L^k = \nabla_k \cdot (-\nabla \boldsymbol{v} \cdot \boldsymbol{q}_k + \frac{1}{2\lambda} \boldsymbol{F}(\boldsymbol{q}_k) + \frac{1}{2\lambda} \nabla_k),\ M_x^k = \frac{\partial}{\partial q_{k,x}},\ M_y^k = \frac{\partial}{\partial q_{k,y}},$$
$$N_x^k = F_x(\boldsymbol{q}_k) + \frac{\partial}{\partial q_{k,x}} \text{ et } N_y^k = F_y(\boldsymbol{q}_k) + \frac{\partial}{\partial q_{k,y}}.$$

avec

$$\boldsymbol{q}_k = \begin{pmatrix} q_{k,x} \\ q_{k,y} \end{pmatrix} \text{ et } \nabla_k = \begin{pmatrix} \frac{\partial}{\partial q_{k,x}} \\ \\ \frac{\partial}{\partial q_{k,y}} \end{pmatrix}.$$

Les méthodes numériques traditionnelles sont pratiquement inapplicables dans ces cas (pour $d > 3$) car leur coût croît exponentiellement avec la dimension. Cependant, depuis quelques années, plusieurs méthodes ont été proposées pour de tels problèmes. De manière très globale, on peut

les résumer en disant qu'on cherche une approximation

$$\psi \approx \sum_{i=1}^{m} \alpha_1^{(i)}(\boldsymbol{q}_1) \cdots \alpha_d^{(i)}(\boldsymbol{q}_d) \tag{13.2}$$

avec certaines fonctions uni-directionnelles $\alpha_i^{(1)}(\boldsymbol{q}) \in V_N$ et un nombre de termes m pas trop élevé. V_N désigne ici un espace de dimension finie utilisé pour l'approximation des fonctions d'une variable $\boldsymbol{q}$. Dans la première approche, dite "Sparse Grids" [40] ou "Sparse Tensor Products", on choisit *a priori* les espaces pour des combinaisons linéaires dans (13.2). L'idée de cette approche consiste à éliminer de l'espace de l'approximation les fonctions qui oscillent beaucoup dans toutes les variables $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_d)$. On a développé dans [17] (un des article annexés à la fin) une variante de cette approche adaptée à l'équation de Fokker-Planck (13.1).

Dans la deuxième approche, dite "Proper Genralized Decomposition", les fonctions $\alpha_i^{(1)}(\boldsymbol{q}) \in V_N$ sont choisis *a posteriori* de sorte que la somme ci-dessus donne une approximation optimale ou quasi-optimale de ψ entre toutes les sommes de m termes. Noter que, contrairement à l'approche des Sparse Grids, on n'impose pas des restrictions *a priori* sur le choix de fonctions $\alpha_1^{(i)}(\boldsymbol{q}_1) \cdots \alpha_d^{(i)}(\boldsymbol{q}_d)$. L'ensemble des sommes à m termes dans (13.2) ne forme pas un espace linéaire, de sorte que pour trouver la meilleure approximation, on doit résoudre un problème d'optimisation non linéaire. Un tel procédé peut être lié aux algorithmes gloutons [64] de l'approximation dans un espace de Hilbert d'un élément donné par des sommes des éléments d'un dictionnaire prédéfini [51]. Quelques résultats numériques utilisant cette approche ont été inclus dans [1].

14 Travaux pendant la thèse et avant

L'objectif de mon travail de thèse était de contribuer à la construction de méthodes numériques pour la simulation des écoulements des fluides viscoélastiques en utilisant les modèles issus de la théorie cinétique des polymères. Les contributions principales étaient les suivantes :

1. Dans le cadre du modèle classique Oldroyd B, exprimé normalement par une équation différentielle constitutive pour le tenseur des contraintes ? mais qui peut être interprété comme une suspension de dumbbells de Hooke dans un solvant Newtonien (cf. le début de Chapitre 8 et Fig. 8, à droite), on a proposé des nouvelles méthodes numériques qui respectent certaines propriétés importantes des équations différentielles concernées. Dans [8], on a développé un schéma numérique basé directement sur la description stochastique du modèle ? qui respecte la propriété de la définition positive du tenseur de conformation ? contrairement aux discrétisations directes de l'équation constitutive Oldroyd-B. Ces méthodes ont été implémentées à l'aide de discrétisations spectrales pour l'écoulement bidimensionnel autour d'un cylindre placé dans un canal. Des expériences numériques démontrent que nos méthodes sont plus stables que celles utilisant l'équation constitutive dans une plage de nombres de Deborah modérés. Dans [11], on a continué à explorer cette voie en démontrant une estimation a priori de l'énergie et en proposant un schéma numérique pour lequel l'estimation est satisfaite automatiquement.
2. Une alternative courante aux dumbbells de Hooke consiste à considérer les dumbbells FENE (Finitely Extensible Non-linear Elastic) où la loi de force de ressort est choisie de telle façon que la longueur du ressort ne puisse pas excéder une certaine limite conformément à la physique (voir remarque 9.1). Ce modèle ne possède plus d'équations constitutives et peut être écrit de deux façons formellement équivalentes : des équations différentielles stochastiques pour le vecteur $\boldsymbol{Q}(t, \boldsymbol{x})$ ou une EDP de Fokker-Planck pour la densité de probabilité $\psi(t, \boldsymbol{x}, \boldsymbol{Q})$. La première option peut être utilisée pour construire des méthodes numériques stochastiques, qui sont devenues très populaires dans la communauté de rhéologie computationnelle. La seconde option est relativement peu exploitée. Dans ma thèse, j'ai démontré que les méthodes numériques basées sur l'équation de Fokker-Planck peuvent être une alternative intéressante aux simulations stochastiques. On a implémenté une méthode spectrale efficace pour l'équation de

Fokker-Planck correspondant au modèle des dumbbells FENE [9, 12]. Des expériences numériques confirment que cette méthode est jusqu'à 50 fois plus rapide que des simulations stochastiques au cas de l'écoulement autour d'un cylindre placé dans un canal. Elle est aussi plus précise en raison de l'absence du bruit stochastique. La même approche a été appliquée aux écoulement non homogènes dans [14], ainsi qu'à un modèle de "reptation" pour les solutions concentrées de polymère [10].

Pendant les trois années passées au Centre de Calcul de l'Académie des Sciences de Russie (1997-2000), j'ai participé aux deux projets suivants :

1. Développement des nouvelles méthodes numériques pour un problème du type Stokes. L'idée de ces méthodes, proposées par Prof. B.V. Pal'tsev, consiste à utiliser l'équation de Poisson pour la pression avec des conditions de bord qui proviennent d'un processus itératif. J'ai effectué une étude théorique de ces méthodes discrétisées par des éléments finis dans le cas d'une bande avec conditions périodiques. J'ai proposé également des modifications des processus itératifs ayant un taux de convergence plus élevé [5]. Les méthodes ont été testés numériquement dans le cas des anneaux non concentriques [6].
2. Dans le même temps, j'ai travaillé avec Prof. A.A. Mayer sur une méthode d'éléments finis pour des EDP non linéaires qui décrivent la propagation de la lumière dans des fibres optiques couplées par l'effet tunnel. Des phénomènes physiques importants, comme l'auto-transition et l'amplification des signaux quasi-soliton, ont été étudiés dans [2, 3, 4].

Bibliographie

[1] A. LOZINSKI, R. OWENS, AND T. PHILLIPS, *The Langevin and Fokker-Planck Equations in Polymer Rheology,* Special Volume of *Handbook of Numerical Analysis,* Numerical Methods for Non-Newtonian Fluids, Vol. XVI (edited by R. Glowinski and J. Xu), 2010.

• Publication dans les ouvrages avec un comité des lecteurs

[2] A. A. MAIER AND A. S. LOZINSKII, *Self-switching of fundamental solitons in tunneling-coupled optical waveguides*, Physics - Doklady, 42 (1997), pp. 483–487.

[3] ——, *Self-switching and amplification of unidirectional distributed-coupling solitons of orthogonal polarizations*, Physics - Doklady, 43 (1998), pp. 74–79.

[4] ——, *Self-switching of solitons in quadratically nonlinear, tunneling-coupled optical waveguides*, Physics - Doklady, 43 (1998), pp. 333–339.

[5] A. S. LOZINSKII, *On the acceleration of finite-element realizations of iterative processes with splitting of boundary conditions for a Stokes-type system*, Zh. Vychisl. Mat. Mat. Fiz., 40 (2000), pp. 1339–1363.

[6] ——, *Finite-element realization of iterative processes with splitting of boundary conditions for a Stokes-type system in nonconcentric annuli*, Zh. Vychisl. Mat. Mat. Fiz., 41 (2001), pp. 1203–1216.

[7] A. LOZINSKI, R. OWENS, AND A. QUARTERONI, *On the simulation of unsteady flow of an Oldroyd-B fluid by spectral methods*, J. Sci. Comput., 17 (2002), pp. 375–383.

[8] C. CHAUVIÈRE AND A. LOZINSKI, *An efficient technique for simulations of viscoelastic flows, derived from the Brownian configuration field method*, SIAM J. Sci. Comput., 24 (2003), pp. 1823–1837.

[9] A. LOZINSKI AND C. CHAUVIÈRE, *A fast solver for Fokker-Planck equation applied to viscoelastic flows calculations : 2D FENE model*, J. Comput. Phys., 189 (2003), pp. 607–625.

[10] A. LOZINSKI, C. CHAUVIÈRE, J. FANG, AND R. G. OWENS, *Fokker-planck simulations of fast flows of melts and concentrated polymer solutions in complex geometries*, Journal of Rheology, 47 (2003), pp. 535–561.

[11] A. LOZINSKI AND R. G. OWENS, *An energy estimate for the Oldroyd B model : Theory and applications*, J. Non-Newtonian Fluid Mech., 112 (2003), pp. 161–176.

[12] C. CHAUVIÈRE AND A. LOZINSKI, *Simulation of complex viscoelastic flows using the Fokker-Planck equation : 3D FENE model*, J. Non-Newtonian Fluid Mech., 122 (2004), pp. 201–214.

[13] ——, *Simulation of dilute polymer solutions using a Fokker-Planck equation*, Comput. Fluids, 33 (2004), pp. 687–696.

[14] A. LOZINSKI, R. G. OWENS, AND J. FANG, *A Fokker-Planck-based numerical method for modelling non-homogeneous flows of dilute polymeric solutions*, J. Non-Newtonian Fluid Mech., 122 (2004), pp. 273–286.

[15] J. FANG, A. LOZINSKI, AND R. G. OWENS, *Towards more realistic kinetic models for concentrated solutions and melts*, J. Non-Newtonian Fluid Mech., 122 (2004), pp. 79–90.

[16] R. GLOWINSKI, J. HE, A. LOZINSKI, J. RAPPAZ, AND J. WAGNER, *Finite element approximation of multi-scale elliptic problems using patches of elements*, Numer. Math., 101 (2005), pp. 663–687.

[17] P. DELAUNAY, A. LOZINSKI, AND R. G. OWENS, *Sparse tensor-product Fokker-Planck-based methods for nonlinear bead-spring chain models of dilute polymer solutions*, in High-dimensional partial differential equations in science and engineering, vol. 41 of CRM Proc. Lecture Notes, Amer. Math. Soc., Providence, RI, 2007, pp. 73–89.

[18] J. HE, A. LOZINSKI, AND J. RAPPAZ, *Accelerating the method of finite element patches using approximately harmonic functions*, C. R. Math. Acad. Sci. Paris, 345 (2007), pp. 107–112.

[19] A. LOZINSKI, J. RAPPAZ, AND J. WAGNER, *Finite element method with patches for Poisson problems in polygonal domains*, ESAIM Proceedings, 21 (2007), pp. 45–64.

[20] A. LOZINSKI AND M. V. ROMERIO, *Motion of gas bubbles, considered as massless bodies, affording deformations within a prescribed family of shapes, in an incompressible fluid under the action of gravitation and surface tension*, Math. Models Methods Appl. Sci., 17 (2007), pp. 1445–1478.

[21] V. REZZONICO, A. LOZINSKI, M. PICASSO, J. RAPPAZ, AND J. WAGNER, *Multiscale algorithm with patches of finite elements*, Math. Comput. Simulation, 76 (2007), pp. 181–187.

[22] F. HECHT, A. LOZINSKI, A. PERRONNET, AND O. PIRONNEAU, *Numerical zoom for multiscale problems with an application to flows through porous media*, Discrete Contin. Dyn. Syst., 23 (2009), pp. 265–280.

[23] A. LOZINSKI, M. PICASSO, AND V. PRACHITTHAM, *An anisotropic error estimator for the Crank-Nicolson method : application to a parabolic problem*, SIAM J. Sci. Comput., 31 (2009), pp. 2757–2783.

[24] A. BONITO, A. LOZINSKI, AND T. MOUNTFORD, *Modeling viscoelastic flows using reflected stochastic differential equations*, Comm. Math. Sci., 8 (2010), pp. 655–670.

[25] P. DEGOND, A. LOZINSKI, AND R. G. OWENS, *Kinetic models for dilute solutions of dumbbells in non-homogeneous flows revisited*, J. Non-Newtonian Fluid Mech., 165 (2010), pp. 509–518.

[26] A. LOZINSKI AND O. PIRONNEAU, *Numerical zoom for localized multiscales*, Numerical Methods for Partial Differential Equations, sous presse (2010) .

[27] M. HILLAIRET, A. LOZINSKI, AND M. SZOPOS, *On discretization in time in simulations of particulate flows*, Discrete Contin. Dyn. Syst. B, sous presse (2010).

• Articles soumis à publication

[28] P. DEGOND, F. DELUZET, A. LOZINSKI, J. NARSKI, AND C. NEGULESCU, *Duality-based asymptotic-preserving method for highly anisotropic diffusion equations*. Soumis à Comm. Math. Sci. Un preprint sur http ://arxiv.org/abs/1008.3405, 2010.

• Proceedings

[29] C. CHAUVIÈRE, J. FANG, A. LOZINSKI, AND R. G. OWENS, *On the numerical simulation of flows of polymer solutions using high-order methods based on the Fokker-Planck equation*, Int. J. Mod. Phys. B, 17 (2003), pp. 9–14.

[30] M. ROMERIO, A. LOZINSKI, AND J. RAPPAZ, *A new modelling for simulating bubble motions in a smelter*, in Light Metals 2005 as held at the 134 th TMS Annual Meeting ; San Francisco, CA, 2005, pp. 547–552.

[31] R. GLOWINSKI, J. HE, A. LOZINSKI, M. PICASSO, J. RAPPAZ, V. REZZONICO, AND J. WAGNER, *Finite element methods with patches and applications*, in Domain decomposition methods in science and engineering XVI, vol. 55 of Lect. Notes Comput. Sci. Eng., Springer, Berlin, 2007, pp. 77–89.

[32] F. HECHT, A. LOZINSKI, AND O. PIRONNEAU, *Numerical zoom and the Schwarz algorithm*, in Domain decomposition methods in science and engineering XVIII, vol. 70 of Lect. Notes Comput. Sci. Eng., Springer, Berlin, 2009, pp. 63–74.

• **Citations externes**

[33] Y. ACHDOU AND Y. MADAY, *The mortar element method with overlapping subdomains,* SIAM J. Numer. Anal., 40 (2002), pp. 601–628 (electronic).

[34] G. AKRIVIS, C. MAKRIDAKIS, AND R. H. NOCHETTO, *A posteriori error estimates for the Crank-Nicolson method for parabolic equations,* Math. Comp., 75 (2006), pp. 511–531 (electronic).

[35] J.-B. APOUNG KAMGA AND O. PIRONNEAU, *Numerical zoom for multiscale problems with an application to nuclear waste disposal,* J. Comput. Phys., 224 (2007), pp. 403–413.

[36] P. BILLER AND F. PETRUCCIONE, *The flow of dilute polymer solutions in confined geometries : A consistent numerical approach,* Journal of Non-Newtonian Fluid Mechanics, 25 (1987), pp. 347–364.

[37] R. BIRD, C. CURTISS, R. ARMSTRONG, AND O. HASSAGER, *Dynamics of polymeric liquids, vol. 1 and 2,* John Wiley & Sons, New-York, 1987.

[38] M. BOSSY, E. GOBET, AND D. TALAY, *A symmetrized Euler scheme for an efficient approximation of reflected diffusions,* J. Appl. Probab., 41 (2004), pp. 877–889.

[39] F. BREZZI, J.-L. LIONS, AND O. PIRONNEAU, *Analysis of a chimera method,* C. R. Acad. Sci. Paris Sér. I Math., 332 (2001), pp. 655–660.

[40] H-J. BUNGARTZ AND M. GRIEBEL, *Sparse grids,* Acta Numerica, 13 (2004), pp. 147–269.

[41] X.-C. CAI, T. P. MATHEW, AND M. V. SARKIS, *Maximum norm analysis of overlapping nonmatching grid discretizations of elliptic equations,* SIAM J. Numer. Anal., 37 (2000), pp. 1709–1728 (electronic).

[42] P. G. CIARLET AND P.-A. RAVIART, *Maximum principle and uniform convergence for the finite element method,* Comput. Methods Appl. Mech. Engrg., 2 (1973), pp. 17–31.

[43] W. E AND J.-G. LIU, *Gauge method for viscous incompressible flows,* Commun. Math. Sci., 1 (2003), pp. 317–332.

[44] Y. EFENDIEV AND T. Y. HOU, *Multiscale finite element methods,* vol. 4 of Surveys and Tutorials in the Applied Mathematical Sciences, Springer, New York, 2009. Theory and applications.

[45] M. GANDER AND C. JAPHET, *A compact algorithm with optimal complexity for non-matching grid projections,* in Domain decomposition methods in science and engineering XVIII, vol. 70 of Lect. Notes Comput. Sci. Eng., Springer, Berlin, 2009, pp. ?– ?

[46] R. GLOWINSKI, T. W. PAN, T. I. HESLA, D. D. JOSEPH, AND J. PÉRIAUX, *A fictitious domain approach to the direct numerical simulation of incompressible viscous flow past moving rigid bodies : application to particulate flow,* J. Comput. Phys., 169 (2001), pp. 363–426.

[47] F. HECHT, O. PIRONNEAU, A. L. HYARIC, AND K. OHTSUKA, `Freefem++`, *ver. 2.23,* http ://www.freefem.org, (2007).

[48] F. D. J.L. STEGER AND J. BENEK, *A chimera grid scheme,* in Advances in Grid Generation. K. N. Chis and U. Ghia ed., ASME FED-Vol. 5, June 1983.

[49] B. JOURDAIN AND T. LELIÈVRE, *Mathematical analysis of a stochastic differential equation arising in the micro-macro modelling of polymeric fluids*, in Probabilistic methods in fluids, World Sci. Publishing, River Edge, NJ, 2003, pp. 205–223.

[50] I. KARATZAS AND S. SHREVE, *Brownian motion and stochastic calculus*, vol. 113 of Graduate Texts in Mathematics, Springer-Verlag, New York, 1991.

[51] C. LE BRIS, T. LELIÈVRE, AND Y. MADAY, *Results and questions on a nonlinear approximation approach for solving high-dimensional partial differential equations*, Constr. Approx., 30 (2009), pp. 621–651.

[52] A. LEFEBVRE, *Numerical simulation of gluey particles*, M2AN Math. Model. Numer. Anal., 43 (2009), pp. 53–80.

[53] J.-L. LIONS AND O. PIRONNEAU, *Domain decomposition methods for CAD*, C. R. Acad. Sci. Paris Sér. I Math., 328 (1999), pp. 73–80.

[54] B. MAURY, *A gluey particle model*, ESAIM : Proceedings, 18 (2007), pp. 133–142.

[55] J.-L. MENALDI, *Stochastic variational inequality for reflected diffusion*, Indiana Univ. Math. J., 32 (1983), pp. 733–744.

[56] H. ÖTTINGER AND Y. RABIN, *Diffusion equation versus coupled Langevin equations approach to hydrodynamics of dilute polymer solutions*, Journal of rheology, 33 (1989), p. 725.

[57] H. C. ÖTTINGER, *Stochastic processes in polymeric fluids*, Springer-Verlag, Berlin, 1996. Tools and examples for developing simulation algorithms.

[58] R.G. OWENS, *A new microstructure-based constitutive model for human blood*, J. Non-Newtonian Fluid Mech., 140 (2006), pp. 57–70.

[59] R. PETTERSSON, *Penalization schemes for reflecting stochastic differential equations*, Bernoulli, 3 (1997), pp. 403–414.

[60] O. PIRONNEAU, *On the transport-diffusion algorithm and its applications to the navier-stokes equations*, Num. Math., 38 (1982), pp. 309–312.

[61] A. H. SCHATZ AND L. B. WAHLBIN, *On the quasi-optimality in L_∞ of the $\dot{H}^1$-projection into finite element spaces*, Math. Comp., 38 (1982), pp. 1–22.

[62] J. SCHIEBER AND H. ÖTTINGER, *The effects of bead inertia on the Rouse model*, The Journal of Chemical Physics, 89 (1988), p. 6972.

[63] L. R. SCOTT AND S. ZHANG, *Finite element interpolation of nonsmooth functions satisfying boundary conditions*, Math. Comp., 54 (1990), pp. 483–493.

[64] V. N. TEMLYAKOV, *Greedy approximation*, Acta Numer., 17 (2008), pp. 235–409.

[65] H. WARNER, *Kinetic theory and rheology of dilute suspensions of finitely extendible dumbbells*, Ind. Eng. Chem. Fund, 11 (1972), pp. 379–387.

[66] J. XU, *Iterative methods by space decomposition and subspace correction*, SIAM Rev., 34 (1992), pp. 581–613.

Quatrième partie

Articles annexés

NUMERICAL ZOOM FOR ADVECTION DIFFUSION PROBLEMS WITH LOCALIZED MULTISCALES

par

ALEXEI LOZINSKI AND OLIVIER PIRONNEAU

paru dans

Numerical Methods for Partial Differential Equations

Numerical Zoom for Advection Diffusion Problems with Localized Multiscales

Alexei Lozinski,[1] **Olivier Pironneau**[2]
[1] *Institut de Mathématiques de Toulouse, Université Paul Sabatier, France*

[2] *Laboratoire Jacques-Louis Lions, Université Pierre et Marie Curie, France*

Received 21 July 2010; accepted 26 July 2010
Published online 1 October 2010 in Wiley Online Library (wileyonlinelibrary.com).
DOI 10.1002/num.20642

We investigate the Schwarz' domain decomposition algorithm for numerical zooms. Error estimates are given for nonmatching meshes for elliptic and parabolic partial differential equations. The method is applied to the security assessment of the burial of nuclear waste which is a typical multiscale problem for porous media flow. © 2010 Wiley Periodicals, Inc. Numer Methods Partial Differential Eq 27: 197–207, 2011

Keywords: domain decomposition; finite elements; multiscale; numerical zoom; porous media flow

I. INTRODUCTION

Security assessment of nuclear waste repository sites requires the numerical simulations of elliptic and parabolic partial differential equations, namely Darcy's law for the hydrostatic potential and the convection diffusion equation of the radio-nucleides. These problems have multiscales because the simulations are done over half a million years while the canister leaks over less than 5,000 years; also the repository sites could pollute as much as 10 sq km of area while the length scales of a canister is of the order of the meter.

Strictly speaking the computations should be done with billions of mesh points on supercomputers; however often enough engineers prototype their applications with a coarse calculation and then a finer one on a subset (zoom) D of the whole domain Ω.

We wish here to justify this approach, i.e., to study convergence and errors when the strategy is implemented in the time loop of the convection diffusion equations and when the meshes used for the zooms are not divided submeshes of the coarse mesh. One may argue that it is unnecessarily complex to do so; however many engineers use the shelves solvers which do not have a numerical zoom capacity built in and so it is quite complex—if even possible—to use them with a zoom mesh made of divided triangles of the coarse mesh.

For Darcy's law the situation is as follows: a coarse calculation is done in Ω, then another one in a zoom $D \subset \Omega$; a correction to the coarse problem must be found which takes into account the

Correspondence to: Olivier Pironneau, Laboratoire Jacques-Louis Lions, Université Pierre et Marie Curie, France (e-mail: olivier.pironneau@upmc.fr)

© 2010 Wiley Periodicals, Inc.

calculation in the zoom. In [1] and [2] two algorithms have been shown to converge: a subspace correction method (or Hilbert space Decomposition, equivalent to Schwarz domain decomposition on the continuous level) and the Harmonic Patch Iterator.

It is very natural to work with composite meshes, a coarse one for most of the domain and a fine one in the region where the data are irregular; engineers do it spontaneously without even bothering about convergence issues; this idea seems to be as old as the finite element method itself and bears many names such as Chimera [3] in fluid mechanics, Global-local [4], composite grids [5], Arlequin [6] in structure mechanics, domain decomposition [7], Hilbert Space Decomposition [8], Mortars [9], etc. Therefore it is the analysis, more than the method, which is possibly new in this article.

For nonmatching meshes the error estimations started with [8] and [1] for Hilbert space Decomposition but stumbled on the problem of quadrature errors for integrals involving products of functions on coarse and fine meshes [10]. The problem was solved in [11] by working with the sup-norm. Here we present an extension of the same argument for parabolic problems. For the heat equation and the convection diffusion problems we will analyze the error for Schwarz' when only one iteration is done at each time step.

II. CONVERGENCE OF SCHWARZ' ALGORITHM FOR DARCY'S LAW ON ARBITRARY NONMATCHING MESHES

In simple fully saturated cases the hydrostatic pressure H of the flow of water underground is given by

$$-\nabla\cdot(K\nabla H) = f \text{ in } \Omega, \quad H \text{ or } (K\nabla H)\cdot n \text{ given on } \Gamma \tag{1}$$

where K is the permeability tensor, n the normal to $\Gamma := \partial\Omega$. Assume that K has multiple scales or is irregular in $D \subset \Omega$ and consider the following Schwarz algorithm to zoom in D numerically.

A. The Schwarz-Zoom Method

Let Ω_H be a triangulated subdomain of Ω such that $\partial\Omega_H = \Gamma_H \cup S_H$ with the property that S_H is strictly inside D and non empty. Let Ω_h be a triangulated approximation of D and call $S_h = \partial\Omega_h$ (see Fig. 1). The triangumation of Ω_H (resp Ω_h) will be denoted $\mathcal{T}_H$ (resp $\mathcal{T}_h$).

Let

$$V_H = \left\{v \in C^0(\Omega_H) \ : \ v|_K \in P^1, \ \forall K \in \mathcal{T}_H\right\}, \quad V_{0H} = \{v \in V_H \ : \ v|_{\partial\Omega_H} = 0\},$$

and similarly with h. Denote by γ_H (resp γ_h) the interpolation operator on V_H (resp V_h).

To simplify the notation, we choose the variant of (1) with the boundary condition $H|_\Gamma = g$ and set K to the identity matrix so that (1) becomes the Dirichlet problem for the Poisson equation. The solution of this problem will be denoted by u from now on. Let $g_H = \gamma_H g$ (to avoid further technical difficulties we suppose here that all the boundary nodes of $\mathcal{T}_H$ lying on Γ_H also belong to Γ so that $\gamma_H g$ is well defined). Starting from any $u_h^0 \in V_h$, the Schwarz-Zoom algorithm proceeds as follows: find $u_H^m \in V_H$, $u_h^m \in V_h$, $m = 1, 2, \ldots$, such that $\forall w_H \in V_{0H}$, $\forall w_h \in V_{0h}$

$$\begin{aligned} &a_H\left(u_H^m, w_H\right) = (f, w_H), \ u_H^m|_{S_H} = \gamma_H u_h^{m-1}, \ u_H^m|_{\Gamma_H} = g_H, \\ &a_h\left(u_h^m, w_h\right) = (f, w_h), \ u_h^m|_{S_h} = \gamma_h u_H^m. \end{aligned} \tag{2}$$

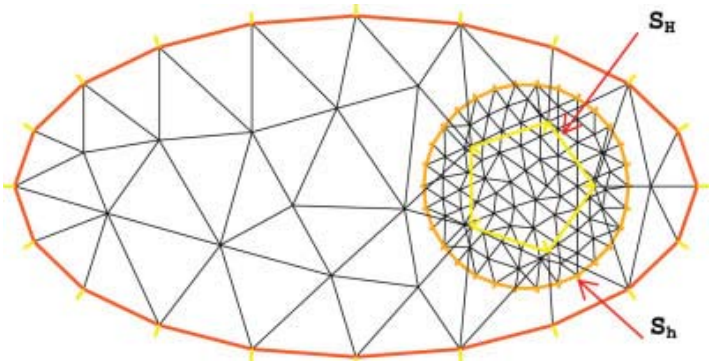


FIG. 1. The drawing shows the fine mesh of the zoom region Ω_h overlapping the coarse mesh in of Ω_H and its hole of boundary S_H. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Hypothesis H1. Assume that all the off-diagonal elements in the matrices obtained on the left hand side of the two problems in (2) are negative or zero so that the maximum principle holds for these problems. Moreover, assume that $v_H \in V_H$ solution of

$$a_H(v_H, w_H) = 0,\ \forall w_H \in V_{0H},\ v_H|_{S_H} = 1,\ v_H|_{\Gamma_H} = 0 \tag{3}$$

satisfies $\lambda := |v_H|_{\infty,S_h} < 1$.

Notice that the maximum principle is known to be true when all the angles of the triangulation are acute [12]. Error estimates in maximum norm of order $h^2 \log \frac{1}{h}$ with respect to the mesh edge size h for linear elements have been obtained in [13].

Theorem 1. *Under Hypothesis H1 the Schwarz-Zoom algorithm converges towards u_H^*, u_h^*, which satisfy*

$$\begin{aligned} &a_H\big(u_H^*, w_H\big) = (f, w_H),\ \forall w_H \in V_{0H},\ u_H^*|_{S_H} = \gamma_H u_h^*,\ u_H^*|_{\Gamma_H} = g_H \\ &a_h\big(u_h^*, w_h\big) = (f, w_h),\ \forall w_h \in V_{0h},\ u_h^*|_{S_h} = \gamma_h u_H^*. \end{aligned} \tag{4}$$

Provided the exact solution u of the Dirichlet problem for the Poisson equation is sufficiently smooth and the distance between Γ and Γ_H is of order H^2, one has

$$\|u - u_H\|_{\infty,\Omega_H} + \|u - u_h\|_{\infty,\Omega_h} \le C\left(H^2 \log\frac{1}{H}\|u\|_{2,\infty,\Omega_H} + h^2 \log\frac{1}{h}\|u\|_{2,\infty,\Omega_h}\right) \tag{5}$$

The proof can be found in [11]. It shows also that the Schwarz algorithm converges linearly at speed proportional to λ.

Remark 1. This results gives a rule to choose h/H because it is logical to have both pieces in (5) of the same size, giving

$$\frac{h^2 \log\frac{1}{h}}{H^2 \log\frac{1}{H}} \sim \frac{\|u\|_{2,\infty,\Omega_H}}{\|u\|_{2,\infty,\Omega_h}} \tag{6}$$

and since D has been chosen so that $\|u\|_{2,\infty,\Omega_h}$ is large compared with $\|u\|_{2,\infty,\Omega_H}$, it gives an h smaller than H.

III. CONVERGENCE OF THE SCHWARZ-ZOOM ALGORITHM FOR CONVECTION-DIFFUSION ON NON-MATCHING MESHES

To solve

$$\begin{aligned} &\frac{\partial u}{\partial t} + b \cdot \nabla u - \Delta u = f \text{ in } \Omega \text{ with } u = g \text{ on } \Gamma = \partial\Omega, \\ &u|_{t=0} = u_0, \end{aligned} \tag{7}$$

we proceed to analyze first the case $b = 0$.

A. Schwarz Method for the Heat Equation ($b = 0$)

To avoid further technical difficulties, we suppose that Ω is a polygonal domain. As before we choose two subdomains of Ω, Ω_H, and Ω_h, and two triangulations $\mathcal{T}_H$ of Ω_H, $\mathcal{T}_h$ of Ω_h, such that

$$\Omega_H \cup \Omega_h = \Omega, \quad \partial\Omega_h \subset \Omega_H, \quad \partial\Omega_H \backslash \Gamma \subset \Omega_h,$$

and work with the P^1 finite element spaces V_H and V_h on $\mathcal{T}_H$ and $\mathcal{T}_h$. As in Fig. 1 we denote by S_h the boundary of Ω_h and by S_H the part of the boundary of Ω_H different from Γ.

Euler-Schwarz algorithm starts from $u_H^0 = \gamma_H u_0$, $u_h^0 = \gamma_h u_0$ and finds $u_H^n \in V_H$ with $u_H^n|_\Gamma = g_H^n$ and $u_h^n \in V_h$, $n = 1, \ldots, N$ such that $\forall w_H \in V_{0H}$, $\forall w_h \in V_{0h}$:

$$\begin{aligned} &\left(\frac{u_H^n - u_H^{n-1}}{\Delta t}, w_H\right)_H + a_H\big(u_H^n, w_H\big) = \big(f^n, w_H\big),\ u_H^n|_{S_H} = \gamma_H u_h^{n-1}, \\ &\left(\frac{u_h^n - u_h^{n-1}}{\Delta t}, w_h\right)_h + a_h\big(u_h^n, w_h\big) = \big(f^n, w_h\big),\ u_h^n|_{S_h} = \gamma_h u_H^{n-1}, \end{aligned} \tag{8}$$

where $f^n = f(t^n, \cdot)$, $g_H^n = g_H(t^n, \cdot)$, $g_H = \gamma_H g$, $a_{H,h}(u, v) = \int_{\Omega_{H,h}} \nabla u \cdot \nabla v$ and γ_H (resp γ_h) is the interpolation operator on V_H (resp V_h). The parentheses $(\cdot, \cdot)$ stand for the L^2 inner product and $(\cdot, \cdot)_H$ (resp $(\cdot, \cdot)_h$) denotes mass lumping on $\mathcal{T}_H$ (resp $\mathcal{T}_h$).

Theorem 2. *Under Hypothesis H1 and provided that u, solution to* (7) *with* $b = 0$, *is sufficiently smooth,* (u_h^n, u_H^n) *in* (8) *solves approximately* (7) *with the following* L^∞ *error estimate*

$$\begin{aligned} &\max_{1 \le n \le N} \left(\left\|u_H^n - u(t^n, \cdot)\right\|_{\infty,\Omega_H}, \left\|u_h^n - u(t^n, \cdot)\right\|_{\infty,\Omega_h}\right) \\ &\quad \le CT \left(\left(\Delta t + H^2 \log \frac{1}{H}\right) \|u\|_{C^2([0,T],C(\Omega_H)) \cap C^1([0,T],H^{2,\infty}(\Omega_H))} \right. \\ &\qquad \left. + \left(\Delta t + h^2 \log \frac{1}{h}\right) \|u\|_{C^2([0,T],C(\Omega_h)) \cap C^1([0,T],H^{2,\infty}(\Omega_h))}\right). \end{aligned} \tag{9}$$

with $T = N\Delta t$ *and a constant C depending only on the domains* Ω_H *and* Ω_h.

Proof. Let $u_H^* = u_H^*(t, \cdot) \in V_H$ and $u_h^* = u_h^*(t, \cdot) \in V_h$ for any $t \in [0, T]$ be the solutions of (4) with $u = u(t, \cdot)$ and $g_H = g_H(t, \cdot)$. By Theorem 1, the solution of (4) exists for all t and the following estimate holds

$$\left\|u_H^* - u\right\|_{\infty,\Omega_H\times[0,T]} + \left\|u_h^* - u\right\|_{\infty,\Omega_h\times[0,T]}$$
$$\leq C\left(H^2 \log\frac{1}{H}\|u\|_{C([0,T],H^{2,\infty}(\Omega_H))} + h^2\log\frac{1}{h}\|u\|_{C([0,T],H^{2,\infty}(\Omega_h))}\right), \tag{10}$$

and similarly for $\|u_h^* - u\|_{\infty,\Omega_h\times[0,T]}$. Moreover, we can differentiate (4) with respect to time to obtain the estimates for the derivatives $\partial u_{H,h}^*/\partial t$:

$$\left\|\frac{\partial u_H^*}{\partial t} - \frac{\partial u}{\partial t}\right\|_{\infty,\Omega_H\times[0,T]} + \left\|\frac{\partial u_h^*}{\partial t} - \frac{\partial u}{\partial t}\right\|_{\infty,\Omega_h\times[0,T]}$$
$$\leq C\left(H^2 \log\frac{1}{H}\|u\|_{C^1([0,T],H^{2,\infty}(\Omega_H))} + h^2\log\frac{1}{h}\|u\|_{C^1([0,T],H^{2,\infty}(\Omega_h))}\right). \tag{11}$$

It is now easy to obtain an *a priori* estimates for a fully implicit version of the Euler-Schwarz algorithm (8): find $\bar{u}_H^n \in V_H$ with $\bar{u}_H^n|_\Gamma = g_H^n$ and $\bar{u}_h^n \in V_h$ such that $\forall w_H \in V_{0H}$, $\forall w_h \in V_{0h}$:

$$\left(\frac{\bar{u}_H^n - \bar{u}_H^{n-1}}{\Delta t}, w_H\right)_H + a_H(\bar{u}_H^n, w_H) = (f^n, w_H),\ \bar{u}_H^n|_{S_H} = \gamma_H \bar{u}_h^n,$$
$$\left(\frac{\bar{u}_h^n - \bar{u}_h^{n-1}}{\Delta t}, w_h\right)_h + a_h(\bar{u}_h^n, w_h) = (f^n, w_h),\ \bar{u}_h^n|_{S_h} = \gamma_h \bar{u}_H^n. \tag{12}$$

Introduce the errors $\varepsilon_H^n \in V_H$, $\varepsilon_h^n \in V_h$ as

$$\varepsilon_H^n = \bar{u}_H^n - u_H^{*n}, \quad \varepsilon_h^n = \bar{u}_h^n - u_h^{*n},$$

where $u_H^{*n} = u_H^*(t^n,\cdot)$, $u_h^{*n} = u_h^*(t^n,\cdot)$. After some simplification, the finite element problem for $\varepsilon_{H,h}^n$ reads: $\forall w_H \in V_{0H}$, $\forall w_h \in V_{0h}$

$$\left(\frac{\varepsilon_H^n - \varepsilon_H^{n-1}}{\Delta t}, w_H\right)_H + a_H\left(\varepsilon_H^n, w_H\right) = \left(r_H^n, w_H\right),\ \varepsilon_H^n|_{S_H} = \gamma_H \varepsilon_h^n,\ \varepsilon_H^n|_{\Gamma_H} = 0,$$
$$\left(\frac{\varepsilon_h^n - \varepsilon_h^{n-1}}{\Delta t}, w_h\right)_h + a_h\left(\varepsilon_h^n, w_h\right) = \left(r_h^n, w_h\right),\ \varepsilon_h^n|_{S_h} = \gamma_h \varepsilon_H^n, \tag{13}$$

with

$$r_H^n = \frac{\partial u}{\partial t}(t^n,\cdot) - \frac{u_H^{*n} - u_H^{*(n-1)}}{\Delta t}, \quad r_h^n = \frac{\partial u}{\partial t}(t^n,\cdot) - \frac{u_h^{*n} - u_h^{*(n-1)}}{\Delta t}.$$

Estimating r_H^n with the help of the tringle inequality and (11) gives

$$\left\|r_H^n\right\|_{\infty,\Omega_H} \leq \left\|\frac{\partial u}{\partial t}(t^n,\cdot) - \frac{u(t^n,\cdot) - u(t^{n-1},\cdot)}{\Delta t}\right\|_{\infty,\Omega_H} + \frac{1}{\Delta t}\left\|\int_{t^{n-1}}^{t^n}\frac{\partial}{\partial t}(u(s,\cdot) - u^*(s,\cdot))ds\right\|_{\infty,\Omega_H}$$
$$\leq C\Delta t\left\|\frac{\partial^2 u}{\partial t^2}\right\|_{\infty,\Omega_H} + \left\|\frac{\partial}{\partial t}(u - u^*)\right\|_{\infty,\Omega_H}$$
$$\leq C\left(\Delta t\|u\|_{C^2([0,T],C(\Omega_H))} + H^2\log\frac{1}{H}\|u\|_{C^1([0,T],H^{2,\infty}(\Omega_H))}\right.$$
$$\left. + h^2\log\frac{1}{h}\|u\|_{C^1([0,T],H^{2,\infty}(\Omega_h))}\right). \tag{14}$$

A similar bound holds for r_h^n. We observe also that the joint maximum of $\|\varepsilon_H^n\|_{\infty,\Omega_H}$ and $\|\varepsilon_h^n\|_{\infty,\Omega_h}$ is always attained in an iternal node of $\mathcal{T}_H$ or $\mathcal{T}_h$. Indeed, the property $\varepsilon_h^n|_{S_h} = \gamma_h \varepsilon_H^n$ entails $\|\varepsilon_h^n\|_{\infty,S_h} \leq \|\varepsilon_H^n\|_{\infty,S_h}$. Hence there is always an internal node of $\mathcal{T}_H$ such that the absolute value of u_H^n at this node is greater than $\|\varepsilon_h^n\|_{S_h}$, provided H is sufficiently small in comparison to the distance between S_H and S_h. Without loss of generality, we can now assume that the joint maximum of $|\varepsilon_H^n|$ and $|\varepsilon_h^n|$ is attained at an internal node x_n of $\mathcal{T}_h$ and $\varepsilon_h^n(x_n) \geq 0$. Taking the hat function ϕ_n associated to the node x_n in the standard basis of V_h as the test function w_h in (13) and observing that $a_h(u_n^n, \phi_n) \geq 0$ by hypothesis *H1*, we see that

$$\frac{u_h^n(x_n) - u_h^{n-1}(x_n)}{\Delta t}(1, \phi_n) \leq (r_h^n, \phi_n)$$

so that

$$\begin{aligned}\max\left(\left\|\varepsilon_H^n\right\|_{\infty,\Omega_H}, \left\|\varepsilon_h^n\right\|_{\infty,\Omega_h}\right) = u_h^n(x_n) &\leq u_h^{n-1}(x_n) + \Delta t \left\|r_H^n\right\|_{\infty,\Omega_H} \\ &\leq \max\left(\left\|\varepsilon_H^{n-1}\right\|_{\infty,\Omega_H}, \left\|\varepsilon_h^{n-1}\right\|_{\infty,\Omega_h}\right) + \Delta t \max\left(\left\|r_H^n\right\|_{\infty,\Omega_H}, \left\|r_h^n\right\|_{\infty,\Omega_h}\right). \quad (15)\end{aligned}$$

Summing up these inequalities on $n = 1, \ldots, N$ and combining with (10) and (14), we obtain the error estimates for the implicit Schwarz algorithm

$$\begin{aligned}&\max_{1\leq n\leq N}\left(\left\|\bar{u}_H^n - u(t^n,\cdot)\right\|_{\infty,\Omega_H}, \left\|\bar{u}_h^n - u(t^n,\cdot)\right\|_{\infty,\Omega_h}\right) \\ &\quad\leq CT\left(\left(\Delta t + H^2 \log\frac{1}{H}\right)\|u\|_{C^2([0,T],C(\Omega_H))\cap C^1([0,T],H^{2,\infty}(\Omega_H))}\right. \\ &\quad\quad\left.+\left(\Delta t + h^2 \log\frac{1}{h}\right)\|u\|_{C^2([0,T],C(\Omega_h))\cap C^1([0,T],H^{2,\infty}(\Omega_h))}\right). \quad (16)\end{aligned}$$

Let us now return to our original Schwarz algorithm (8). Introduce the finite differences $d_H^n = (u_H^n - u_H^{n-1})/\Delta t$, $d_h^n = (u_h^n - u_h^{n-1})/\Delta t$. The equations for them are obtained by subtracting (8) at time t^{n-1} from the same equation at time t^n: $d_H^n|_{\Gamma_H} = g_H^n - g_H^{n-1}$, $\forall w_H \in V_{0H}$, $\forall w_h \in V_{0h}$

$$\begin{aligned}&\left(\frac{d_H^n - d_H^{n-1}}{\Delta t}, w_H\right)_H + a_H(d_H^n, w_H) = \left(\frac{f^n - f^{n-1}}{\Delta t}, w_H\right),\ d_H^n|_{S_H} = \gamma_H d_h^{n-1}, \\ &\left(\frac{d_h^n - d_h^{n-1}}{\Delta t}, w_h\right)_h + a_h(d_h^n, w_h) = \left(\frac{f^n - f^{n-1}}{\Delta t}, w_h\right),\ d_h^n|_{S_h} = \gamma_h d_H^{n-1}, \quad (17)\end{aligned}$$

for $n \geq 2$. By an application of the maximum principle, similar as above, we observe that

$$\begin{aligned}\max\left(\left\|d_H^n\right\|_{\infty,\Omega_H}, \left\|d_h^n\right\|_{\infty,\Omega_h}\right) \leq \max\Bigg(&\left\|g_H^n - g_H^{n-1}\right\|_{\infty,\Gamma}, \\ &\max\left(\left\|d_H^{n-1}\right\|_{\infty,\Omega_H}, \left\|d_h^{n-1}\right\|_{\infty,\Omega_h}\right) + \Delta t\left\|\frac{f^n - f^{n-1}}{\Delta t}\right\|_{\infty,\Omega}\Bigg). \quad (18)\end{aligned}$$

The norm of the finite difference in time for f in the last line can be bounded by the maximum of $\frac{\partial f}{\partial t}$ which, in turn, is bounded by $\|u\|_{C^2([0,T],C(\Omega))\cap C^1([0,T],H^{2,\infty}(\Omega))}$ since $\frac{\partial f}{\partial t} = \frac{\partial^2 u}{\partial t^2} - \Delta\frac{\partial u}{\partial t}$. It is also

easy to obtain the appropriate estimates for $\|g_H^n - g_H^{n-1}\|_{\infty,\Gamma}$, $\|d_H^1\|_{\infty,\Omega_H}$ and $\|d_h^1\|_{\infty,\Omega_h}$, so that we have finally

$$\max_{1\le n\le N}\left(\left\|d_H^n\right\|_{\infty,\Omega_H},\left\|d_h^n\right\|_{\infty,\Omega_h}\right)\le C_d := CT\|u\|_{C^2([0,T],C(\Omega))\cap C^1([0,T],H^{2,\infty}(\Omega))}. \tag{19}$$

The last step in the proof is to estimate the difference between the original Schwarz algorithm (8) and the implicit one (12). We thus introduce $e_H^n = \bar{u}_H^n - u_H^n$, $e_h^n = \bar{u}_h^n - u_h^n$ and observe that they are solutions to the finite element problems: $\forall w_H \in V_{0H}$, $\forall w_h \in V_{0h}$

$$\begin{aligned}
&\left(\frac{e_H^n - e_H^{n-1}}{\Delta t}, w_H\right)_H + a_H\left(e_H^n, w_H\right) = 0,\ e_H^n|_{S_H} = \gamma_H e_h^n + \Delta t\gamma_H d_h^n,\ e_H^n|_{\Gamma_H} = 0,\\
&\left(\frac{e_h^n - e_h^{n-1}}{\Delta t}, w_h\right)_h + a_h\left(e_h^n, w_h\right) = 0,\ e_h^n|_{S_h} = \gamma_h e_H^n + \Delta t\gamma_h d_H^n,
\end{aligned} \tag{20}$$

with initial conditions $u_H^0 = 0, u_h^0 = 0$. We remind that all d_H^n and d_h^n are bounded in the maximum norm by C_d, take any number $A > C_d\Delta t$ and consider $s_H^n = e_H^n - A\alpha_H$, $s_h^n = e_h^n - A\alpha_h$ where the auxiliary functions $\alpha_H \in V_H$, $\alpha_h \in V_h$ satisfy

$$\begin{aligned}
&a_H(\alpha_H, w_H) = 0 \quad \forall w_H \in V_H, \quad \alpha_H|_{S_H} = \gamma_H\alpha_h + 1, \quad \alpha_H|_\Gamma = 0,\\
&a_h(\alpha_h, w_h) = 0 \quad \forall w_h \in V_h, \quad \alpha_h|_{S_h} = \gamma_h\alpha_H + 1.
\end{aligned} \tag{21}$$

We now want to prove that s_H^n and s_h^n are nonpositive everywhere and for all n. If this is not the case, we can suppose without loss of generality that the global maximum of $\{s_H^n\}$, $\{s_h^n\}$ is attained at the time step number $k \ge 1$ on $s_h^k(x_k)$ where x_k is a node of $\mathcal{T}_h$. We have

$$\begin{aligned}
&\left(\frac{s_h^k - s_h^{k-1}}{\Delta t}, w_h\right)_h + a_h\left(s_h^k, w_h\right) = 0 \quad \forall w_h \in V_h,\\
&s_h^k|_{S_h} = \gamma_h s_H^k + \Delta t\gamma_h d_H^k - A.
\end{aligned} \tag{22}$$

The first line in (22) tells us that x_k cannot be an internal node of $\mathcal{T}_h$, while the second line implies that it cannot be a boundary node either since $\Delta t d_H^k - A < 0$ on S_h. We obtain thus $\forall n \ge 1$

$$\sup_{x\in\Omega}\max\left(e_H^n(x), e_h^n(x)\right) \le AC_\alpha := A\max(\|\alpha_H\|_{\infty,\Omega_H}, \|\alpha_h\|_{\infty,\Omega_h}).$$

Note that the constant C_α can be bounded independently of H and h. Repeating the same derivation with e_H^n, e_h^n replaced by $-e_H^n$, $-e_h^n$ and tending A to $C_d\Delta t$ from above, we see that

$$\max\left(\left\|e_H^n\right\|_{\infty,\Omega_H},\left\|e_h^n\right\|_{\infty,\Omega_h}\right) \le C_\alpha C_d\Delta t,\ \forall n \ge 1. \tag{23}$$

The desired result (9) is obtained from (16) and (23) by the triangle inequality. ∎

Remark 2. The fully implicit Schwarz algorithm (12) can provide an interesting alternative to our basic algorithm (8) if several iterations of a standard Schwarz method like (2) are used on every time step to find an approximation for the couple (u_H^n, u_h^n). The convergence proof above can be easily adapted for such an algorithm. It can be preferable to (8) in terms of the cost to accuracy ratio.

Remark 3. One can associate two different time steps for the two subdomains, say Δt for the problems with u_H^n and δt for the problems with u_h^n. A variant of the Schwarz algorithm can be then constructed provided $\Delta t = m\Delta t$ with an integer m. In the zoom the problem is advanced m times and then the coarse problem is advanced once; the convergence proof is extensible to this case without difficulties.

Remark 4. Extension of our results to the operator $\nabla \cdot (K\nabla)$ in place of Δ is straightforward.

B. Extension to Convection Diffusion ($b \neq 0$)

The convection term in (7) can be discretized by the characteristic-Galerkin scheme. Combination with the Schwarz method gives the following algorithm: find $u_H^n \in V_H$, $u_h^n \in V_h$, such that $\forall w_H \in V_{0H}$, $\forall w_h \in V_{0h}$

$$\begin{aligned}
&\frac{1}{\Delta t}\left(u_H^n - u_H^{n-1} o X^{n-1}, w_H\right)_H + a_H\left(u_H^n, w_H\right) = (f, w_H), \quad u_H^n|_{S_H} = \gamma_H u_h^{n-1},\ u_H^n|_\Gamma = g_H^n,\\
&\frac{1}{\Delta t}\left(u_h^n - u_h^{n-1} o X^{n-1}, w_h\right)_h + a_h\left(u_h^n, w_h\right) = (f, w_h), \quad u_h^n|_{S_h} = \gamma_h u_H^{n-1},
\end{aligned} \tag{24}$$

where $X^{n-1}(x) = X(t^{n-1})$ and X is the solution of $\dot{X} = b(X(t), t)$, $X(t^n) = x$. We consider the case where X^{n-1} can be computed exactly (b piecewise constant for instance).

As shown in [14] and [15], the accuracy of the characteristic-Galerkin method deteriorates when the time step Δt becomes too small before the mesh size. One cannot expect therefore to extend Theorem 2 to the general case $b \neq 0$. We have, however, a weaker result:

Theorem 3. *Under Hypothesis H1 and provided that u, solution to* (7)*, is sufficiently smooth,* (u_h^n, u_H^n) *in* (24) *solves approximately* (7) *with the following* L^∞ *error estimate*

$$\begin{aligned}
\max_{1\le n\le N}\left(\left\|u_H^n - u(t^n)\right\|_{\infty,\Omega_H}, \left\|u_h^n - u(t^n)\right\|_{\infty,\Omega_h}\right) \le CT\left(\Delta t + \frac{H^2 \log\frac{1}{H}}{\Delta t}\right)\|u\|_{H^{2,\infty}(\Omega_H\times[t^{n-1},t^n])}\\
+\left(\Delta t + \frac{h^2 \log\frac{1}{h}}{\Delta t}\right)\|u\|_{H^{2,\infty}(\Omega_h\times[t^{n-1},t^n])},
\end{aligned} \tag{25}$$

with $T = N\Delta t$ *and a constant* C *depending only on the domains* Ω_H *and* Ω_h.

Proof. Denote $u^n = u(t^n, \cdot)$ and consider $\tilde{u}_H^n \in V_H$ such that $\forall w_H \in V_{0H}$

$$a_H\left(\tilde{u}_H^n, w_H\right) = a_H(u^n, w_H),\ \tilde{u}_H^n|_{S_H} = \gamma_H u(t^n),\ \tilde{u}_H^n|_\Gamma = g_H(t^n, \cdot).$$

By the results in [13], we have

$$\left\|\tilde{u}_H^n - u^n\right\|_{\infty,\Omega_H} \le CH^2 \log\frac{1}{H}\|u^n\|_{2,\infty,\Omega_H}. \tag{26}$$

We also introduce $\tilde{u}_h^n \in V_h$ in a similar way and observe that the same estimate as (26) holds for $\tilde{u}_h^n$ replacing H by h.

Denote

$$\varepsilon_H^n = u_H^n - \tilde{u}_H^n, \;\; \varepsilon_h^n = u_h^n - \tilde{u}_h^n.$$

Then $\varepsilon_H^n \in V_H$ solves $\forall w_H \in V_{0H}$

$$\left(\frac{\varepsilon_H^n - \varepsilon_H^{n-1} o X^{n-1}}{\Delta t}, w_H\right)_H + a_H\left(\varepsilon_H^n, w_H\right) = \left(r_H^n, w_H\right),$$
$$\varepsilon_H^n|_{S_H} = \gamma_H \varepsilon_h^{n-1}, \; u_H^n|_{\Gamma_H} = 0 \tag{27}$$

with

$$r_H^n = \frac{Du}{Dt}(t^n, \cdot) - \frac{\tilde{u}_H^n - \tilde{u}_H^{n-1} o X^{n-1}}{\Delta t},$$

where $\frac{Du}{Dt} = \frac{\partial u}{\partial t} + b\nabla u$ is the material derivative. By the triangle inequality and (26) we have

$$\begin{aligned}
\left\|r_H^n\right\|_{\infty,\Omega_H} &\le \left\| \frac{Du}{Dt}(t^n, \cdot) - \frac{u^n - u^{n-1} o X^{n-1}}{\Delta t} \right\|_{\infty,\Omega_H} + \left\| \frac{u^n - u^{n-1} o X^{n-1}}{\Delta t} - \frac{\tilde{u}_H^n - \tilde{u}_H^{n-1} o X^{n-1}}{\Delta t} \right\|_{\infty,\Omega_H} \\
&\le C\Delta t \|u\|_{H^{2,\infty}(\Omega_H \times [t^{n-1}, t^n])} + \frac{1}{\Delta t}\left\|u^n - \tilde{u}_H^n\right\|_{\infty,\Omega_H} + \frac{1}{\Delta t}\left\|\left(u^{n-1} - \tilde{u}_H^{n-1}\right) o X^{n-1}\right\|_{\infty,\Omega_H} \\
&\le C\left[\left(\Delta t + \frac{H^2 \log \frac{1}{H}}{\Delta t}\right)\|u\|_{H^{2,\infty}(\Omega_H \times [0,T])} + \frac{h^2 \log \frac{1}{h}}{\Delta t}\|u\|_{H^{2,\infty}(\Omega_h \times [0,T])}\right].
\end{aligned}$$

Note here the use of the inequality

$$\left\|\left(u^{n-1} - \tilde{u}_H^{n-1}\right) o X^{n-1}\right\|_{\infty,\Omega_H} \le \left\|u^{n-1} - \tilde{u}_H^{n-1}\right\|_{\infty,\Omega_H} + \left\|u^{n-1} - \tilde{u}_h^{n-1}\right\|_{\infty,\Omega_h}.$$

One cannot leave only the norm in Ω_H in the right hand side of this inequality since a characteristic starting in Ω_H at time t^n can go back to Ω_h at time t^{n-1}.

Writing down an estimate for r_h^n similar to what is done above for r_H^n and noting that

$$\max\left(\left\|\varepsilon_H^n\right\|_{\infty,\Omega_H}, \left\|\varepsilon_h^n\right\|_{\infty,\Omega_h}\right) \le \max\left(\left\|\varepsilon_H^{n-1}\right\|_{\infty,\Omega_H}, \left\|\varepsilon_h^{n-1}\right\|_{\infty,\Omega_h}\right) + \Delta t \max\left(\left\|r_H^n\right\|_{\infty,\Omega_H}, \left\|r_h^n\right\|_{\infty,\Omega_h}\right),$$

we conclude the proof in the same way as that of the estimate (16) for the implicit Schwarz method in the proof of Theorem 2. ∎

IV. NUMERICAL TESTS

We shall apply the method to solve one of Andra's COUPLEX test cases (see [16]). but before that let us verify the error estimate on a simpler example. A closed form solution $u = tr^{-2}$, $r^2 = 0.01 + x^2 + y^2$ is computed in a square $(-1, 1)^2$ by solving the heat equation with an appropriate source term $f = (r^{-2} + 4tr^{-4} - 8t(x^2 + y^2)r^{-6}$. The subdomains are the square minus the disk of radius 0.25 and a disk of radius 0.3. The left (resp right) part of Table I displays the L^∞ errors on the coarse (resp fine) subdomain for several values of H and Δt when $h = H/8$. The

TABLE I. The left (resp right) part displays the L^∞ errors on the coarse (resp fine) subdomain for several values of H and Δt when $h = H/8$.

$\Delta t \backslash H$	0.2	0.1	0.05	0.2	0.1	0.05
0.25	0.124378	0.124378	0.124378	9.31509	9.70629	9.8834
0.125	0.0621891	0.0621891	0.0621891	7.26138	8.28583	8.75552
0.0625	0.0310945	0.0310945	0.0310945	3.83163	5.8366	6.72392
0.03125	0.0155473	0.0155473	0.0155473	0.355462	2.9373	4.32695
0.015625	3.02283	0.00777363	0.00777363	4.10578	0.442487	2.29146

If Δt is too small the quadrature errors become too large; this is why the errors for $H = 0.25$ and $\Delta t = 0.015625$ are large (the same happens to the problem without domain decomposition).

COUPLEX test case is made first of an elliptic problem $\nabla \cdot (K \nabla H) = 0$ to obtain the hydrostatic pressure H from the porosity K and convection velocity as $-K \nabla H$ and then of a convection diffusion equation for the concentration $c(x, y, t)$ of the radionucleide in the underground. The difficulties come from the large differences of K in each geological layer and from the different length scales in the problem. The details of the calculation are not very important but Fig. 2 shows that the Schwarz-zoom algorithm works for this problem.

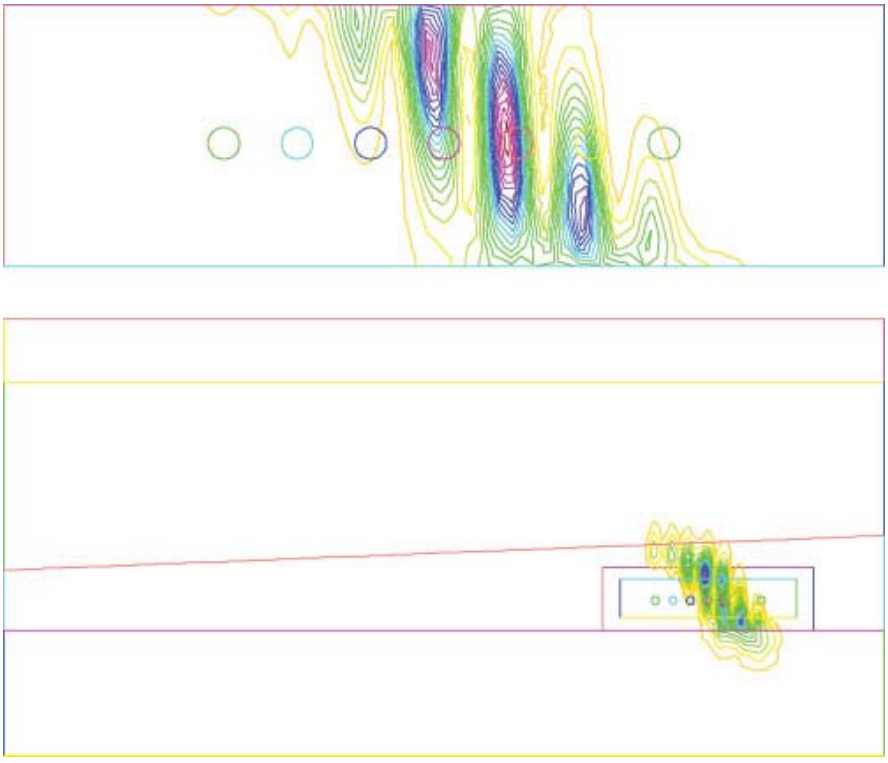


FIG. 2. The top pictures shows detailed level lines of the concentration of radionucleide nearby the canisters (the seven circles) after 600 time steps. A concentration is given at initial time in the canister which are all supposed to leak due to rust and spread their content by convection and diffusion in the clay. The calculation is obtained from a Schwarz-zoom refinement. The bottom picture shows the full domain, its decomposition into two subdomains and a superposition of both fields of level lines of c_H and c_h; both fields match fairly well. The domain is physically a rectangle (0,10 km) × (0,500 m) but it is rescaled for the simulation; the long horizontal lines are the separations between (from top to bottom), marl, limestone, clay, and dogger layers. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

V. CONCLUSION

The article has explained why the standard Schwarz algorithm can be used for a numerical zoom procedure, with standard interpolations at the boundaries. The method works for time dependent problems as well and it is not necessary to iterate the Schwarz algorithm because the time stepping procedure has the same effect. A compatibility condition between the time step and the mesh size may be necessary in the case of convection diffusion problems. However, the same feature is present at the standard finite element characteristic-Galerkin simulations on a single mesh so that it cannot be attributed to the Schwarz method.

Application to flows through porous media is covered by the two theorems of this article; it would be nice to extend the results to other fluid flows such as Navier-Stokes flow but the proofs here rely heavily on the maximum principle which is not valid for the generalized Stokes operator.

References

1. R. Glowinski, J. He, J. Rappaz, and J. Wagner, Approximation of multi-scale elliptic problems using patches of finite elements, C R Math Acad Sci Paris 337 (2003), 679–684.
2. J. He, A. Lozinski, and J. Rappaz, Accelerating the method of finite element patches using approximately harmonic functions, C R Math Acad Sci Paris 345 (2007), 107–112.
3. J. L. Steger and J. A. Benek, On the use of composite grid schemes in computational aerodynamics, Comput Methods Appl Mech Eng 64 (1987), 301–320.
4. C. D. Mote, Global-local finite element method, Int J Num Methods Eng 3 (1971), 565–574.
5. J. Fish and T. Belytschko, Elements with embedded localization zones for large deformation problems, Comput Struct 30 (1988), 247–256.
6. H. B. Dhia and G. Rateau, Mathematical analysis of the mixed Arlequin method, C R Math Acad Sci Paris 332 (2001), 649–654.
7. P. L. Lions, On the Schwarz alternating method, I,II,III. International Symposium on Domain Decomposition Methods for Partial Differential Equations, SIAM, Philadelphia, 1988, 89–90.
8. J.-L. Lions and O. Pironneau, Domain decomposition methods for CAD, C R Acad Sci Paris Sér I Math 328 (1999), 73–80.
9. Y. Achdou and Y. Maday, The mortar element method with overlapping subdomains, SIAM J Numer Anal 40 (2002), 601–628.
10. F. Brezzi, J. L. Lions, and O. Pironneau, Analysis of a chimera method, C R Math Acad Sci Paris 332 (2001), 655–660.
11. F. Hecht, A. Lozinski, and O. Pironneau, Numerical zoom and the Schwarz algorithm, M. Bercovier, M. Ganders, R. Kornhuber, and O. Widlund, editors, In Proceeding Domain Decomposition Conference dd18, Jerusalem 2008, Lecture Notes in Computational Science and Engineering, Vol. 70, Springer Verlag, Berlin, 2009, pp. 63–73.
12. P. G. Ciarlet and P.-A. Raviart, Maximum principle and uniform convergence for the finite element method, Comput Methods Appl Mech Engrg 2 (1973), 17–31.
13. A. H. Schatz and L. B. Wahlbin, On the quasi-optimality in L_∞ of the $\dot{H}_1$-projection into finite element spaces, Math Comp 38 (1982), 1–22.
14. O. Pironneau, On the transport-diffusion algorithm and its applications to the Navier-Stokes equations, Numerische Mathematik 38 (1982), 309–312.
15. O. Pironneau and M. Tabata, Galerkin-characteristics finite element method, Int J Numer Methods Fluids, to appear.
16. ANDRA, Couplex test cases (2001). Available at: http://www.andra.fr/couplex.

On discretization in time in simulations of particulate

par

Matthieu Hillairet, Alexei Lozinski and Marcela Szopos

à paraître dans

Discrete and Continuous Dynamical Systems - Series B (DCDS-B)

Manuscript submitted to
AIMS' Journals
Volume **X**, Number **0X**, XX **200X**

Website: http://AIMsciences.org

pp. **X–XX**

ON DISCRETIZATION IN TIME IN SIMULATIONS OF PARTICULATE FLOWS

Matthieu Hillairet

Ceremade, UMR CNRS 7534, Université Dauphine
Place du Maréchal De Lattre De Tassigny
F-75016 Paris, France

Alexei Lozinski

Institut de Mathématiques de Toulouse
UMR 5219, Université de Toulouse et CNRS
118 route de Narbonne
F-31062 Toulouse Cedex 9, France

Marcela Szopos

Institut de Recherche Mathématique Avancée
UMR 7501, Université de Strasbourg et CNRS
7 rue René Descartes
F-67084 Strasbourg Cedex, France

(Communicated by the associate editor name)

Abstract. We propose a time discretization scheme for a class of ordinary differential equations arising in simulations of fluid/particle flows. The scheme is intended to work robustly in the lubrication regime when the distance between two particles immersed in the fluid or between a particle and the wall tends to zero. The idea consists in introducing a small threshold for the particle-wall distance below which the real trajectory of the particle is replaced by an approximated one where the distance is kept equal to the threshold value. The error of this approximation is estimated both theoretically and by numerical experiments. Our time marching scheme can be easily incorporated into a full simulation method where the velocity of the fluid is obtained by a numerical solution to Stokes or Navier-Stokes equations. We also provide a derivation of the asymptotic expansion for the lubrication force (used in our numerical experiments) acting on a disk immersed in a Newtonian fluid and approaching the wall. The method of this derivation is new and can be easily adapted to other cases.

1. **Introduction.** One of the challenges for fluid/particle flow simulations is to provide an accurate resolution of the lubrication regime when the distance between two particles immersed in the fluid or between a particle and the wall becomes very small. The presence of extremely high gradients of the velocity in the narrow gap between the particle and the wall (or between the two particles) results in very strong drag forces (which can be referred to as the lubrication forces in this case)

2000 *Mathematics Subject Classification.* Primary: 65Z05, 76T99; Secondary: 65L04.

Key words and phrases. Particulate flow, near contact, lubrication force.

The work of the first author was supported by ANR project ANR-08-JCJC-0104-01. The work of the last two authors was partially supported by the project "BQR AO1" of Université Paul Sabatier.

and thus requires special attention when one attempts to construct an efficient and accurate discretization in space. The corrections specific to the lubrication forces were proposed in the context of the boundary integral methods [22, 12, 17], in that of the Stokesian Dynamics [1, 23], and in that of the force-coupling method [4]. Short-range repulsion forces mimicking the lubrication ones were also used in [8] in the context of fictitious domain finite element methods. Taking aside the formidable problems related to the discretization in space, we focus our attention in this article on the discretization in time. Indeed, the presence of lubrication force of very high magnitude in the equations governing the motion of a particle necessitates the use of very small time steps in the lubrication regime in order to obtain a physically acceptable solution. Our idea is to prohibit the particle from approaching too closely the wall during a simulation. We shall thus choose a threshold q_s for the distance q between the particle and the wall and replace the true trajectory of the particle by an approximated one, in which the distance q is kept equal to q_s until a possible rebound of the particle from the wall. The moment of rebound is predicted using an auxiliary quantity (a crude approximation of the velocity) that is computed all along the period of time when the particle is stuck at the distance q_s. This approach reminds the gluey particle model of [18, 15] where q_s is set to zero and the limit of vanishing viscosity is considered. However, our motivations are quite different from that behind the gluey particle model. This model is intended as a simple alternative to the standard governing equations of Navier-Stokes type, possibly corrected by taking into account the roughness of the particle surface. On the other hand, our approach is to take the standard fluid equations for granted (assuming the particle surface to be smooth) and to provide a tool for a robust time discretization of them. It means, in particular, that we would need an accurate enough method valid for any given value of the viscosity, not necessarily small. Note also, that our threshold q_s will typically depend on the time step size, so that it is indeed a numerical device and it has no physical meaning.

The plan of the article is as follows: we start by reminding the governing equations in a general setting and by explaining in more detail the difficulties related to the simulations in the lubrication regime in the next section. Section 3 is the core of the paper. The idea of the threshold is rigorously introduced and studied there for the model ordinary differential equation representing the essence of the general setting in the simplest case of a circular (or spherical) particle approaching the wall. The discussion is held on the continuous level in Section 3. The discretization in time is introduced in Section 4 where several implementations of our idea are proposed on the discrete level followed by numerical experiments. We also include an appendix detailing a derivation of the asymptotic expansion for the lubrication force acting on a disk approaching the wall. The method of this derivation is new and can be easily adapted to other cases.

2. Motivations: governing equations for the fluid/particle flows and some difficulties arising in their simulations.

The general setting of this work is a study of the motion of a rigid particle immersed in a viscous, incompressible fluid with a particular emphasis on situations when the particle approaches the plane. To set the notations, we assume in general that the fluid (with the particle inside) fills a fixed domain $\Omega \subset \mathbb{R}^d$ with $d = 2$ or 3, while the region occupied by the particle $\mathcal{B}_t \subset \Omega$ varies with time t. We denote the time-dependent fluid domain $\mathcal{F}_t$ so that $\Omega = \overline{\mathcal{B}_t} \cup \mathcal{F}_t$ at any time t. Supposing that the inertial effects are negligible in the

fluid and the no-slip conditions are valid on the boundaries of Ω and $\mathcal{B}_t$, the fluid motion is governed by the Stokes equations

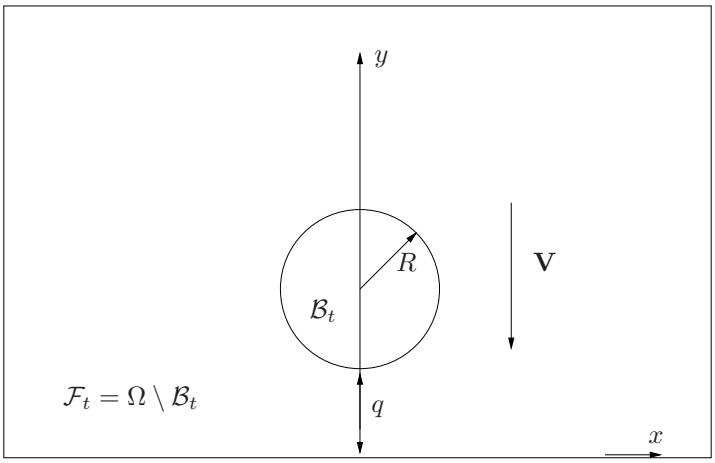


FIGURE 1. Notations

$$\begin{cases} -\nu\Delta\mathbf{u} + \nabla p = \rho_f \mathbf{g}, & \text{in } \mathcal{F}_t \\ \nabla \cdot \mathbf{u} = 0, & \text{in } \mathcal{F}_t \\ \mathbf{u} = \mathbf{0}, & \text{on } \partial\Omega \\ \mathbf{u} = \mathbf{V} + \omega \times \mathbf{r} & \text{on } \partial\mathcal{B}_t \end{cases} \tag{1}$$

where $\mathbf{u}$ and p are the velocity and the pressure in the fluid, ν and ρ_f are the viscosity and the density of the fluid, $\mathbf{g}$ is the external force, $\mathbf{V} = \mathbf{V}(t)$ and $\omega = \omega(t)$ are the translational and angular velocities of the rigid body $\mathcal{B}_t$, $\mathbf{r} = \mathbf{x} - \mathbf{G}$ is the vector pointing from the center of mass of the particle $\mathbf{G}$ to a point $\mathbf{x}$ on its boundary. The more realistic Navier-Stokes equations may be accommodated into the framework (1) by including the convective term $\mathbf{u} \cdot \nabla\mathbf{u}$ into $\mathbf{g}$.

The fluid exerts a net force $\mathbf{F}$ and a torque $\mathbf{T}$ on the particle given by

$$\begin{aligned} \mathbf{F} = \mathbf{F}(\mathcal{B}_t, \mathbf{V}, \omega) &= \int_{\partial\mathcal{B}_t} (2\nu D(\mathbf{u}) - pI)\mathbf{n}\mathrm{d}\sigma, \\ \mathbf{T} = \mathbf{T}(\mathcal{B}_t, \mathbf{V}, \omega) &= \int_{\partial\mathcal{B}_t} \mathbf{r} \times (2\nu D(\mathbf{u}) - pI)\mathbf{n}\mathrm{d}\sigma, \end{aligned} \tag{2}$$

where $D(\mathbf{u})$ stands for the symmetric gradient of $\mathbf{u}$ and $\mathbf{n}$ is the unit normal vector on $\partial\mathcal{B}_t$ directed towards the fluid domain. Note that $\mathbf{F}$ and $\mathbf{T}$ are indeed functions of only the placement of the particle $\mathcal{B}_t$ and its translational and angular velocities since the velocity $\mathbf{u}$ and the pressure p are uniquely determined in the fluid by the parameters $\mathcal{B}_t$, $\mathbf{V}$ and ω as the solution to the Stokes equations (1). Moreover, the dependence of $\mathbf{F}$ and $\mathbf{T}$ on $\mathbf{V}$ and ω is linear. Using these notations we write out the equations of motion of the particle as follows

$$\begin{cases} m\frac{d\mathbf{V}}{dt} = \mathbf{F}(\mathcal{B}_t, \mathbf{V}, \omega) + m\mathbf{g}, \\ \mathbb{I}_t\frac{d\omega}{dt} + \omega \times \mathbb{I}_t\omega = \mathbf{T}(\mathcal{B}_t, \mathbf{V}, \omega), \end{cases} \tag{3}$$

where m is the mass of the particle and $\mathbb{I}_t$ is its inertia tensor, expressed in the fixed Cartesian frame and thus dependent on time. Equations (3) are coupled with the

equations describing the propagation of the particle, *i.e.* $\dot{\mathbf{G}} = \mathbf{V}$ for the center of mass and $\dot{\mathbf{r}}_i = \omega \times \mathbf{r}_i$, $i = 1, \dots, d$, for the vectors $\mathbf{r}_i$ fixed in the particle. The force $\mathbf{F}$ is a sum of the Archimedes force due to gravity and of the drag force which is purely hydrodynamic, *i.e.* obtained from (1) by setting $\mathbf{g} = 0$. The particularity of the drag is that it tends very rapidly to ∞ when the particle approaches the wall, thus preventing collisions between them. Indeed, it has been proved in [10] (2D case), [11] (3D case), that a smooth rigid body embedded in a viscous fluid cannot touch the wall in finite time. In the regime of very small distance between the particle and the wall, the drag force is also known as the lubrication force and it is notoriously difficult to take into account in a numerical simulation.

It is noteworthy that these considerations are valid only for smooth surfaces; however, modelling surface roughness of the wall or particle is a much more delicate issue. Several experimental ([24], [14], [13], [25]) and theoretical ([13], [21]) works propose in this case to modify the expression for the lubrication drag force by introducing a shift to the distance to the wall, of magnitude strictly lower than the roughness size. This physically means that roughness decreases the dissipation in the system, and that the interaction is similar to that between equivalent smooth surfaces, located at some intermediate position, between the peaks and the valleys of asperities.

To simulate the motion of the particle governed by equation described above, one should be able to solve numerically the Stokes system (1) for any given particle position $\mathcal{B}_t$ and for any given velocities $\mathbf{V}$ and ω. This is a formidable task in itself especially because the position of the particle is not known a priori but changes with time. We do not make precise the choice of the numerical method for the Stokes system. We just assume for the moment that the force $\mathbf{F}$ and the torque $\mathbf{T}$ can be computed for any $\mathcal{B}_t$, $\mathbf{V}$, ω but this computation is in general very expensive. Our primary goal in this article is to devise an efficient discretization in time of (3) using as few as possible solutions of the Stokes system. The simplest idea is to use the following scheme

$$m\frac{\mathbf{V}^k - \mathbf{V}^{k-1}}{\Delta t} = \mathbf{F}(\mathcal{B}_{t_{k-1}}, \mathbf{V^k}, \omega^k) + m\mathbf{g}(t_k), \tag{4}$$

$$\mathbb{I}_{t_{k-1}}\frac{\omega^k - \omega^{k-1}}{\Delta t} + \omega^{k-1} \times \mathbb{I}_{t_{k-1}}\omega^k = \mathbf{T}(\mathcal{B}_{t_{k-1}}, \mathbf{V^k}, \omega^k), \tag{5}$$

$$\frac{\mathbf{G}^k - \mathbf{G}^{k-1}}{\Delta t} = \mathbf{V}^k, \tag{6}$$

$$\frac{\mathbf{r}_i{}^k - \mathbf{r}_i{}^{k-1}}{\Delta t} = \mathbf{r}_i{}^k \times \omega^k, \quad i = 1, \dots, d. \tag{7}$$

We have introduced here the uniform grid in time $t_k = k\Delta t$ and have denoted the quantities computed at the time step t_k by the superscript k. The idea behind the scheme (4)–(7) is to compute first the velocities by (4)–(5) and then to propagate the particle using the last available values of the velocities. The equations (4)–(5) are thus coupled together and also coupled with the solution of the Stokes system on a fixed geometry given by position of the particle $\mathcal{B}_{t_{k-1}}$ on the previous time step t_{k-1}. The cost of such a computation is normally essentially the same as that of the Stokes system (1) with prescribed $\mathbf{V}$ and ω. We need thus one solution of the Stokes system per time step. This approach was successfully used, for example, in [16] in conjunction with a fictitious domain discretization of the Stokes system as in [8]. However, independently from the discretization in space,

one would encounter problems when the particle approaches the wall. Indeed, in this lubrication regime the force $\mathbf{F}$ explodes and thus the scheme (4)–(5) is no longer valid unless an extremely low value for the time step is used that makes the simulation prohibitively expensive. A commonly used cure for this problem is to introduce short-range repulsion forces between the particle and the wall, as in [8], for example. However, the influence of these (not necessarily realistic) forces on the accuracy of a simulation is not well understood. Another simple idea is just to stop the particle when it tries to penetrate the wall during a numerical simulation. However, it is then not necessarily clear what criterion should be chosen to decide if the particle should eventually bounce off the wall and when should it happen. These questions have a partial answer in the articles [18, 15] on the gluey particle model. It is shown there that the particle trajectory satisfies an integro-differential equation in the limit of vanishing viscosity, which is easy to discretize in time using moderate time steps and which predicts the moment of an eventual rebound from the wall. We pursue a similar idea in this article but our aim is to construct an approximated trajectory of the particle in the lubrication regime that would be accurate enough for any given value of the viscosity, not necessarily small.

3. A model ordinary differential equation with lubrication forces.

3.1. The model.

Let us consider the simplest setting of the problem described in the previous section: assume $\Omega \subset \mathbb{R}^2$ is the half-plane $\{(x, y),\ y > 0\}$ and the particle is a disk of radius R. Let moreover $\mathbf{g}(t) = g(t)\mathbf{e}_2$ and assume the particle is at rest at the initial time. The x-component of the particle velocity and its angular velocity will then vanish at all time. The position of the particle is fully determined by its distance q from the bottom, as in Figure 1. The net force $\mathbf{F}$ is the sum of the drag, which is a function of q and $\mathbf{V}$, linear in $\mathbf{V}$, and of the Archimedes force:

$$\mathbf{F} = -n(q)\mathbf{V} + m_a g(t)\mathbf{e}_2, \text{ with } m_a = m - \rho_f|\mathcal{B}_t|,$$

where $n(q)$ is the drag coefficient computed by the Stokes equations (1). Denoting the y-component of the velocity by v, we are thus led to the following differential equations

$$\begin{cases} m\dot{v} = -n(q)v + m_a g, \\ \dot{q} = v. \end{cases} \tag{8}$$

We are especially interested in the lubrication regime of small q. The asymptotic of $n(q)$ when $q \to 0$ is given by $3\sqrt{2}\pi\nu\left(\frac{R}{q}\right)^{\frac{3}{2}}$ (see Appendix A) in the 2D case. After eliminating v from the system (8) and going to non-dimensional variables (see Appendix B for the details) we obtain the following equation for $q(t)$

$$\ddot{q} = -\varepsilon\frac{\dot{q}}{q^{\frac{3}{2}}} + g, \tag{9}$$

with $\varepsilon = \frac{3\nu}{\rho_s R^{\frac{3}{2}}}\sqrt{\frac{2\rho_s}{g_{char}(\rho_s - \rho_f)}}$ where ρ_s is the density of the solid disk and g_{char} is the characteristic value of g.

In the same way, we can consider the analogous three dimensional problem setting Ω to be the half-space $\{(x, y, z),\ z > 0\}$ and the particle to be a ball of radius R. The asymptotic expression is well known in this case (cf. Remark 5) and

is given by $6\pi V\nu\frac{R^2}{q}$. Performing the same non-dimensionalizations as in the 2D case, we arrive at the equation for $q(t)$ (the distance from the particle to the wall):

$$\ddot{q} = -\varepsilon\frac{\dot{q}}{q} + g, \tag{10}$$

with $\varepsilon = \frac{9\nu}{2\rho_s R^{\frac{3}{2}}}\sqrt{\frac{\rho_s}{g_{char}(\rho_s-\rho_f)}}$.

liquid	ρ_s/ρ_f	R (mm)	ε in (9)	ε in (10)
water	1.1	1	0.13	0.14
	1.5	0.1	1.55	1.64
glycerin	1.1	1	152	161
	1.5	0.1	1843	1954

TABLE 1. Typical values of the parameter ε in (9) or in (10) taking the viscosity and density of either water or glycerin and different value of the particle radius R and density ρ_s.

We remind that equations of the type (9)–(10) are at the basis of the gluey particle model of [18, 15]. The model consists in fact in considering the limit $\varepsilon \to 0$, which is physically the limit of vanishing viscosity. We see, however, that ε is not necessarily small.

3.2. **Estimates and an approximated solution for the model ODE.** Consider the ordinary differential equation

$$\ddot{q} = -n(q)\dot{q} + g \tag{11}$$

or, equivalently, the system of first-order equations

$$\begin{cases} \dot{v} &= -n(q)v + g \\ \dot{q} &= v \end{cases} \tag{12}$$

where $n(q)$ is a given differentiable decreasing positive function on $(0,\infty)$ with $n(q) = N'(q)$ such that $N(q) \to -\infty$ as $q \to 0$. Note that this is valid for $n(q) = \varepsilon/q^{\frac{3}{2}}$ or $n(q) = \varepsilon/q$, which give the asymptotic of the lubrication force in 2D and 3D respectively. We assume $g \in L^1_{loc}(\mathbb{R}^+)$ from now on. Under this hypothesis, Problem (11), completed with appropriate initial conditions, is well-posed, as proved for instance in [18, Prop. 1.1] for the case $n(q) = \varepsilon/q$. The proof is generalized without problem to any $n(q)$ satisfying the above hypotheses, as mentioned in [18, Section 5.1].

Proposition 1. *Given $(q_0, v_0) \in \mathbb{R}^2$ with $q_0 > 0$ there exists a unique positive global solution $q \in W^{1,\infty}_{loc}(\mathbb{R}^+)$ to (11) with initial conditions:*

$$q(0) = q_0 \quad \dot{q}(0) = v_0. \tag{13}$$

In the following, we are especially interested in the lubrication regime of small $q \to 0$, when the drag coefficient $n(q)$ tends to infinity, which can create difficulties when computing numerical solutions of (11). In order to overcome this problem, one should either use a very small time step when discretizing the equation, but this can become computationally heavy, or propose a strategy to prevent q from becoming too small. The approach we propose here, inspired by the latter idea, consists

in introducing a threshold value q_s for the distance q, below which, instead of solving (11), we introduce an alternative model and construct a sufficiently accurate "approximate" solution $\bar{q}$ instead of the "true" solution q. We will first pursue this strategy at the continuous level and then implement it at the discrete level in the next section.

Our construction is the following: we first choose some threshold value q_s and next suppose that q goes below this value after some time t_1 when $q(t)$ hits the threshold q_s for the first time. Our aim is to find a suitable approximation of q after t_1 which enables to predict the time t_2 when q goes above the threshold q_s without solving (11). To this end, we introduce

$$\bar{v}(t) = \dot{q}(t_1) + \int_{t_1}^{t} g(s)ds \tag{14}$$

for $t \geq t_1$ and note that $\bar{v}(t_2) = \dot{q}(t_2)$. Indeed, integrating (11) from t_1 to $t > t_1$ shows that

$$\dot{q}(t) = \dot{q}(t_1) + N(q_s) - N(q(t)) + \int_{t_1}^{t} g(s)ds = N(q_s) - N(q(t)) + \bar{v}(t). \tag{15}$$

Setting here $t = t_2$ and noting that $q(t_2) = q_s$ gives the desired result. Our approximation of the trajectory $q(t)$, denoted by $\bar{q}(t)$, stems from the assumption (verified afterwards in Proposition 3) that the velocity $\dot{q}(t_2)$ at the return point t_2 is small provided the threshold q_s is small. Consequently, a good approximation for the time t_2 should be provided by the time $\bar{t}_2$ defined as the first time larger than t_1 when $\bar{v}(t) = 0$. The construction of the approximated trajectory is hence the following: we first assume that $\bar{q}(t)$ is the same as $q(t)$ until the latter hits q_s for the first time at $t = t_1$. Next, the trajectory $\bar{q}(t)$ is frozen until the time $t = \bar{t}_2$ and resumes then again as a solution to (11) starting from q_s with zero velocity:

$$\bar{q}(t) = \begin{cases} q(t), \text{ for } 0 < t < t_1 \\ q_s, \text{ for } t_1 \leq t < \bar{t}_2 \\ \text{solution to (11) with } q(\bar{t}_2) = q_s,\ \dot{q}(\bar{t}_2) = 0, \text{ for } t \geq \bar{t}_2 \end{cases} \tag{16}$$

(see Figure 2).

In order to study the error committed by introducing the approximated trajectory $\bar{q}(t)$, we start by infering the following estimate:

Proposition 2. *The time $\bar{t}_2$ provides a lower bound for t_2, i.e. $\bar{t}_2 \leq t_2$.*

Proof. There are three cases. If t_2 does not exist, there is nothing to prove. If $t_2 = t_1$ this means t_1 is a local minimum for q so that $\dot{q}(t_1) = 0$ and $t_1 = \bar{t}_2 = t_2$. Finally, if t_2 exists and satisfies $t_1 < t_2$, since $q(t_1) = q(t_2) = q_s$ the function $q(t)$ has a local minimum $t_{\min} \in (t_1, t_2)$ where $\dot{q}(t_{\min}) = 0$. Evaluating (15) at $t = t_{\min}$ shows then $\bar{v}(t_{\min}) = N(q(t_{\min})) - N(q_s) \leq 0$. On the other hand, $\bar{v}(t_2) = \dot{q}(t_2) \geq 0$ since the function $q(t)$ is increasing at t_2 by the definition of t_2. The intermediate value theorem tells now that there exists $\bar{t} \in [t_{\min}, t_2]$ such that $\bar{v}(\bar{t}) = 0$. Thus the first time $t = \bar{t}_2$ when $\bar{v}(t) = 0$ is certainly in $(t_1, \bar{t}] \subset (t_1, t_2]$.

□

We have proved, in particular, that if the true trajectory $q(t)$ ever returns to the values above the threshold q_s, *i.e.* $t_2 < \infty$, the indicator $\bar{t}_2$ will show it, *i.e.* $\bar{t}_2 < \infty$. In Proposition 5, we compute an upper bound $\tilde{t}_2$ for t_2. In particular, under suitable assumptions on g the time t_2 exists.

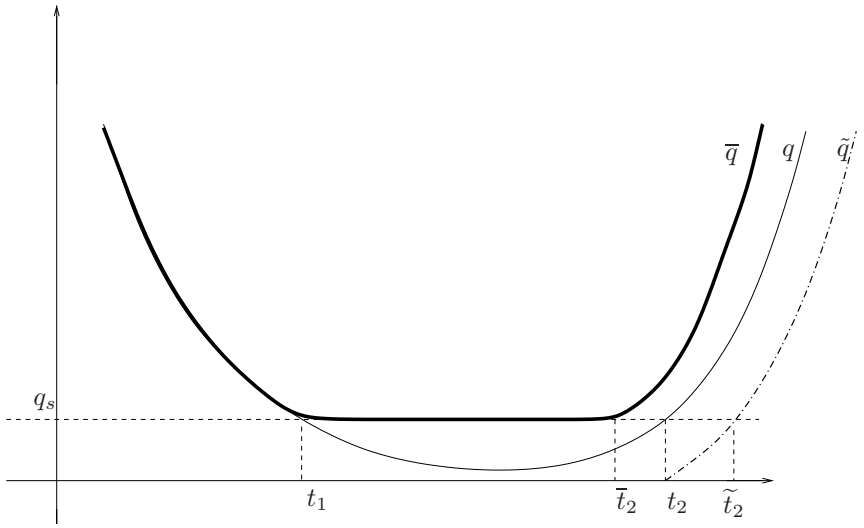


FIGURE 2. Construction of the approximated trajectories $\bar{q}(t)$ and $\tilde{q}(t)$.

To see moreover that $\bar{t}_2$ may be a good approximation to t_2 we need, as mentioned before, to control $\dot{q}(t_2)$. The following proposition proves that $\dot{q}(t_2)$ is not too large with respect to q_s.

Proposition 3. *Let $\dot{q}^+$ (resp. g^+) be the positive part of $\dot{q}$ (resp. g): $\dot{q}^+ = \max(\dot{q}, 0)$ (resp. $g^+ = \max(g, 0)$). Then*

$$\|\dot{q}^+\| \leq \frac{\|g^+\|}{n(q_s)}, \tag{17}$$

where $\|\cdot\|$ denotes the supremum on $[t_1, t_2]$. In particular, $\dot{q}(t_2) \geq 0$, so that $\dot{q}(t_2) \leq \frac{\|g^+\|}{n(q_s)}$.

Proof. Let $\dot{q}$ be non-negative on an interval $[t_{\min}, t_{\max}] \subset [t_1, t_2]$. Without loss of generality, we can thus assume that $t_{\min}$ is a local minimum of $q(t)$ and $t_{\max}$ is either a local maximum or $t_{\max} = t_2$. Then, we introduce

$$f(t) := \dot{q}(t) - \frac{\|g^+\|}{n(q(t))}, \quad \forall\, t \in [t_1, t_2].$$

If $f(t_0) > 0$ at some $t_0 \in (t_{\min}, t_{\max})$ then $f(t)$ remains positive on a small interval around t_0 so that $\ddot{q} < 0$ by invoking (11) and thus $\dot{q}(t)$ is decreasing on this interval. Since $q(t)$ is increasing on $[t_{\min}, t_{\max}]$ and $n(q)$ is decreasing, we obtain that $f(t)$ is decreasing locally around t_0 so that $f(t) \geq f(t_0) > 0$ for all $t \in [t_0 - \delta, t_0]$ with some $\delta > 0$. Repeating this argument for times down to $t_{\min}$ we see that $f(t)$ should be positive on $[t_{\min}, t_0]$ but this is impossible since $\dot{q}(t_{\min}) = 0$ and $f(t_{\min}) < 0$. This proves that $f(t) \leq 0$ on $[t_{\min}, t_{\max}]$. □

The results above allow us to fully characterize the error of approximating $q(t)$ by $\bar{q}(t)$ up to time $\bar{t}_2$ and to bound it by certain quantities depending on $\bar{q}$ alone for time after $\bar{t}_2$, at least in the special case when $g(t)$ is given by

$$g(t) = \begin{cases} g_-(t) < 0, & \text{for } t \leq t_0, \\ g_+ > 0, & \text{for } t > t_0, \end{cases} \tag{18}$$

with some positive constants t_0, g_+ and a negative function $g_-(t)$. Indeed, we prove in the following proposition that the error $\tau = t_2 - \bar{t}_2$ is bounded by $1/n(q_s)$ and we remind that $n(q_s) \to \infty$ as $q_s \to 0$.

Proposition 4. *If $g(t)$ is given by (18), then*

$$\begin{aligned} &0 \le t_2 - \bar{t}_2 \le \frac{1}{n(q_s)}, \\ &\sup_{t\in[0,\bar{t}_2]} |q(t) - \bar{q}(t)| \le q_s, \\ &|\dot{q}(\bar{t}_2) - \dot{\bar{q}}(\bar{t}_2)| = \dot{q}(\bar{t}_2) \le \frac{g_+}{n(q_s)}. \end{aligned} \tag{19}$$

Moreover, both $q(t)$ and $\bar{q}(t)$ are non-decreasing for $t \ge \bar{t}_2$ and

$$|q(t) - \bar{q}(t)| \le \begin{cases} \bar{q}(t), & \text{if } \bar{t}_2 \le t < t_2, \\ \bar{q}(t) - \bar{q}(t-\tau), & \text{if } t \ge t_2, \end{cases} \tag{20}$$

with $\tau = t_2 - \bar{t}_2$.

Proof. We first note that $t_2 \ge \bar{t}_2 \ge t_0$. The first inequality here is already proved in Proposition 2. The second inequality follows from $\bar{v}(\bar{t}_2) = 0$, which can be rewritten as

$$-\dot{q}(t_1) = \int_{t_1}^{\bar{t}_2} g(s)ds.$$

Since $\dot{q}(t_1) \le 0$, the last equality cannot hold if g is negative everywhere on $[t_1, t_2]$. Since $\bar{t}_2 \ge t_0$, we have $g(t) = g_+ > 0$ for $t \ge \bar{t}_2$. Evaluating (15) at $t = t_2$ gives

$$\dot{q}(t_2) = \int_{\bar{t}_2}^{t_2} g(s)ds = g_+(t_2 - \bar{t}_2). \tag{21}$$

Recalling Proposition 3, we see that

$$t_2 - \bar{t}_2 = \frac{\dot{q}(t_2)}{g_+} \le \frac{1}{n(q_s)},$$

which is the first inequality in (19). The second inequality in (19) is obvious since $0 < q(\bar{t}_2) \le q_s$ and $\bar{q}(\bar{t}_2) = q_s$. The third inequality follows from Proposition 3 since $\dot{q}(\bar{t}_2) \ge 0$. Indeed, (15) evaluated at $t = \bar{t}_2$ gives $\dot{q}(\bar{t}_2) = N(q_s) - N(q(\bar{t}_2))$.

We now turn to the study of $q(t)$ and $\bar{q}(t)$ for time $t \ge \bar{t}_2$, in order to prove that $q(t)$ and $\bar{q}(t)$ are non-decreasing on this interval. We first observe that $q(t)$ cannot have local maximums at $t > t_0$. Indeed, at any local extremum $t_e > t_0$, equation (11) would imply $\ddot{q}(t_e) = g_+ > 0$. Since $q(t)$ is non-decreasing at $\bar{t}_2 > t_0$, it should be non-decreasing also everywhere on $[\bar{t}_2, \infty)$. The same reasoning applies to $\bar{q}(t)$. The estimate (20) is now evident in the case $\bar{t}_2 \le t \le t_2$ since $\bar{q}(t) \ge q_s$ and $q(t) \le q_s$ for such t.

In order to prove (20) in the case $t \ge t_2$ we will show that $q(t)$ is squeezed between $\bar{q}_\tau(t)$ and $\bar{q}(t)$ for $t \ge t_2$ where $\bar{q}_\tau(t) = \bar{q}(t-\tau)$. Indeed both $\bar{q}_\tau(t)$ and $q(t)$ satisfy the same equation (11) with the same initial condition at $t = t_2$ ($q(t_2) = \bar{q}_\tau(t_2) = q_s$) but possibly different initial velocities $\dot{q}(t_2) \ge 0$, $\dot{\bar{q}}_\tau(t_2) = 0$. Writing (11) for $\bar{q}_\tau(t)$ and $q(t)$, taking the difference thereof and integrating from t_2 to t yields

$$\dot{q}(t) - \dot{\bar{q}}_\tau(t) = \dot{q}(t_2) - N(q(t)) + N(\bar{q}_\tau(t)).$$

This shows that if the trajectories of $q(t)$ and $\bar{q}_\tau(t)$ intersect at some time t' then $\dot{q}(t') \geq \dot{\bar{q}}_\tau(t')$ so that $q(t) \geq \bar{q}_\tau(t)$ at least for some time after t'. Since $q(t) \geq \bar{q}_\tau(t)$ holds also for some time after t_2 it should hold for all $t \in [t_2, \infty)$.

It remains to prove that $q(t) \leq \bar{q}(t)$ for all $t \in [t_2, \infty)$. To this end, we integrate (11) for $q(t)$ from t_2 to t and that for $\bar{q}(t)$ from $\bar{t}_2$ to t. This yields with the aid of (21)

$$\dot{q}(t) = \dot{q}(t_2) + N(q_s) - N(q(t)) + \int_{t_2}^{t} g(s)ds = N(q_s) - N(q(t)) + \int_{\bar{t}_2}^{t} g(s)ds, \quad (22)$$

$$\dot{\bar{q}}(t) = N(q_s) - N(\bar{q}(t)) + \int_{\bar{t}_2}^{t} g(s)ds. \quad (23)$$

We see now that if $q(t') = \bar{q}(t')$ at some time $t' \geq \bar{t}_2$ then also $\dot{q}(t') = \dot{\bar{q}}(t')$. It means by the uniqueness of solutions to (11) that the trajectories of $q(t)$ and $\bar{q}(t)$ either coincide or do not intersect on $[\bar{t}_2, \infty)$. Since $q(\bar{t}_2) \leq \bar{q}(\bar{t}_2)$ this implies $q(t) \leq \bar{q}(t)$ on $[\bar{t}_2, \infty)$.

□

It is possible to relax the hypotheses of the last Proposition on the particular form of the function g in several ways, if we still remain in the case when q goes below the threshold value q_s a finite number of times (in particular, g should be allowed to change sign only a finite number of times). Although we do not have an analogue of these results for a general $g(t)$, we can always provide an easily computable upper bound for t_2 alongside the lower bound $\bar{t}_2$.

Proposition 5. *Let $\tilde{t}_2$ be the first time after $\bar{t}_2$ such that*

$$\int_{\bar{t}_2}^{\tilde{t}_2} (\tilde{t}_2 - s)g(s)ds = q_s. \quad (24)$$

Then $\tilde{t}_2 \geq t_2$.

Proof. Consider $\tilde{q}(t)$ defined for $t \geq \bar{t}_2$ as the solution to

$$\dot{\tilde{q}}(t) = \int_{\bar{t}_2}^{t} g(s)ds, \quad \tilde{q}(\bar{t}_2) = 0$$

We recall now the property (22) of the original solution and note $N(q(t)) < N(q_s)$ for $t \in (\bar{t}_2, t_2)$ so that

$$\dot{q}(t) \geq \int_{\bar{t}_2}^{t} g(s)ds \geq \dot{\tilde{q}}(t).$$

Since $q(\bar{t}_2) > \tilde{q}(\bar{t}_2)$, the trajectory of $q(t)$ lies above that of $\tilde{q}(t)$ for $t \in (\bar{t}_2, t_2)$ so that $q(t)$ hits the threshold q_s before $\tilde{q}(t)$. In other words, $t_2 < \tilde{t}_2$, where $\tilde{t}_2$ is the first time after $\bar{t}_2$ such that $\tilde{q}(t) = q_s$. This definition of $\tilde{t}_2$ is equivalent to that in (24). □

4. **Discretization in time of the model ODE.** We are now going to give a discretized in time version of some of the ideas elaborated for the model ODE (11) in the previous section. We shall start by a straightforward Euler discretization in time of (11) rewritten as the system (12) to remind that this scheme (referred to as Algorithm 1) does not always work in the near-contact regime. We then turn to a discrete version of the method constructing the approximate trajectory $\bar{q}(t)$ defined in (16) and suitable in the near-contact case. This is done in Algorithm 2, where an

a priori threshold value $q_s > 0$ is used to compute an approximated solution to the problem. We pursue by discussing the strategy to choose an appropriate value for q_s which will typically depend on the time step Δt. Finally, we combine an adaptative strategy for the choice of the time step with the threshold approximation given by Algorithm 2. This gives Algorithm 3.

Although all these developments are done in this article only for the model equation (11), we have in mind a possible incorporation of our time marching schemes into a full simulation of flow with solid particles. The coefficient $n(q)$ will be then given by a solution to the Stokes system at each time step where q represents the surface-wall distance in the case of a particle approaching to the wall or the inter-particle distance in the case of two approaching particles. The "external" force $g(t)$ in (11) will then include the components of the physical external forces, like gravitation, in the direction connecting the closest points on the surfaces of nearly touching particles, plus other corrections. Most importantly, these include the part of hydrodynamic drag due to the convective terms in Navier-Stokes, if the convective terms in the fluid motion equation are not neglected. The time marching scheme should thus work robustly for any right hand side $g(t)$, in particular that changing the sign. Indeed, a possible scenario even in the simplest test case of a particle falling under its weight on a horizontal wall (so that $g(t)$ is negative in the beginning) could be that the convective terms in Navier-Stokes equations become so important at some point in time that they prevail over the gravitation force and $g(t)$ becomes positive, which may lead to the rebound of the particle.

4.1. **Three schemes for ODE (11).** Introducing the time step Δt and the discrete times $t_k = k\Delta t$, $k = 1, 2, \ldots$ we first describe in Algorithm 1 a straightforward Euler discretization in time of (11) rewritten as the system (12).

Algorithm 1 (straightforward Euler)

Step 0. Initialize $(v^0,\, q^0)$.

Step 1. For $k = 1, 2, \ldots$ update $(v^k,\, q^k)$ as follows

$$\begin{cases} \frac{v^k - v^{k-1}}{\Delta t} = -n(q^{k-1})v^k + g(t_k) \\ \frac{q^k - q^{k-1}}{\Delta t} = v^k. \end{cases} \tag{25}$$

Algorithm 1 is inspired by the real fluid-particle simulations in which one first finds the new velocity at each time step and then moves the particle with this velocity. An immediately evident drawback of the scheme in Algorithm 1 is that it does not necessarily provide a positive approximation q^k. We remind that negative values of q are unphysical. Moreover, even if approximations q^k remain positive but become tiny at some time steps, one would require very small stepsize to obtain an accurate solution.

Remark 1. One may argue that this can be cured by resorting to a fully implicit scheme that couples the evaluation of the velocity with the displacement of the particle:

$$\begin{cases} \frac{v^k - v^{k-1}}{\Delta t} = -n(q^k)v^k + g(t_k) \\ \frac{q^k - q^{k-1}}{\Delta t} = v^k. \end{cases} \tag{26}$$

Indeed, this scheme provides a positive solution for q, for example, in the important case $n(q) = \varepsilon/q$. To see this we eliminate v^k from (26), which gives an equation for

q^k only

$$\frac{q^k}{\Delta t} - \frac{\varepsilon q^{k-1}}{q^k} = \frac{q^{k-1}}{\Delta t} - \varepsilon + v^{k-1} + \Delta t g(t_k)$$

The function of q^k in the left-hand side is increasing form $-\infty$ to ∞ when q^k goes from 0 to ∞. This means that there is a unique positive solution q^k for any given v^{k-1} and q^{k-1}. However, scheme (26) would be too expensive in a real fluid/particle simulation where $n(q^k)$ should be recomputed after any change in q^k through a numerical approximation of the Stokes equations.

Remark 2. Difficulties related to the discretization of these differential equations have already been pointed out in [18] (see Remark 3.1): it is shown that very small values of q can be reached (below the smallest real number which can be stored by standard numerical softwares and also below intermolecular distances). Different approaches have been proposed in previous related works to deal with similar problems: in [18, Section 5.2], the author suggests the introduction of a cut-off function for the microscopic distance $\gamma = \varepsilon \ln(q)$, defined in terms of the roughness of the surface (see also [15, Section 2.5]); in [15, Section 2.4.3], for the case of fluid/particle simulation, a constraint for the distance q is defined in terms of the mesh size of the fluid domain.

In order to overcome these difficulties, in this work we propose to introduce a small threshold value $q_s > 0$, to replace the exact solution $q(t)$ by the approximated one $\bar{q}(t)$, as summarized in (16) on the continuous level. Discretization in time of this idea introducing also an approximation of $\bar{v}(t)$ defined in (14) is detailed in Algorithm 2. Note that the exact threshold is replaced in Step 2 of this algorithm by the last available value of q^k before passing below q_s, *i.e.* q^{k_1-1}, so that the passage time t_1 is legitimately approximated by t_{k_1-1}.

Algorithm 2 (with an *a priori* fixed threshold)

Step 0. Initialize (v^0, q^0) and choose $q_s > 0$.

Step 1. For $k = 1, 2, \ldots$ update (v^k, q^k) by (25) until $q^k \le q_s$ at some $k = k_1$. We then redefine $v^{k_1} = 0$, $q^{k_1} = q^{k_1-1}$, initialize $\bar{v}$ by $\bar{v}^{k_1} = v^{k_1-1} + g(t_{k_1})\Delta t$ and switch to Step 2.

Step 2. For $k = k_1 + 1, k_1 + 2, \ldots$ update $\bar{v}^k$ as follows

$$\bar{v}^k = \bar{v}^{k-1} + g(t_k)\Delta t \tag{27}$$

keeping $q^k = q^{k_1-1}$, $v^k = 0$ until $\bar{v}^k \ge 0$ at some $k = k_2$. We then abandon the calculation of $\bar{v}$ and switch to Step 3.

Step 3. For $k = k_2 + 1, k_2 + 2, \ldots$ update (v^k, q^k) as in Step 1. If q^k goes again under q_s at some time step, then reintroduce $\bar{v}$ and keep switching between Steps 1 and 2 back and forth.

As suggested by Proposition 4 a good choice for q_s would be such that $1/n(q_s) = C\Delta t$ with some constant C so that the criterion for switching from Step 1 to Step 2 can be rewritten as $n(q^k) \ge 1/(C\Delta t)$. Indeed, extrapolating the result of Proposition 4 to a general $g(t)$, we see then that the error of approximation of the exact return time t_2 by $\bar{t}_2 \approx k_2\Delta t$ is of order Δt. Moreover, the error $|\bar{q}(t) - q(t)|$ is dominated by $|\bar{q}(t) - \bar{q}(t-\tau)|$ with τ of order Δt. Thus the error caused by the introduction of the threshold should be of of the same order as that of Euler scheme

itself. However, an optimal choice of the constant C above would require some *a posteriori* error indicators to control the error of the Euler discretization.

Algorithm 3 (with adapted variable stepsize and automatically chosen threshold)

Step 0. Initialize (v^0, q^0) for $t_0 = 0$, choose the first time step $\Delta t_1 > 0$, the minimal tolerated time step $\Delta t_{\min} > 0$ and the error tolerance per time step $tol > 0$.

Step 1. For $k = 1, 2, \ldots$ set $t_k = t_{k-1} + \Delta t_k$, update (v^k, q^k) by

$$\begin{cases} \frac{v^k - v^{k-1}}{\Delta t_k} = -n(q^{k-1})v^k + g(t_k) \\ \frac{q^k - q^{k-1}}{\Delta t_k} = v^k. \end{cases} \tag{28}$$

and calculate the approximation for the error committed on this time step by $e_k = \max(|\ddot{q}(t_k)|, |\ddot{v}(t_k)|)\Delta t_k^2/2$ where $\ddot{q}(t_k)$, $\ddot{v}(t_k)$ are computed using (29)–(30). If $e_k \leq tol$ and $q_k > 0$ then we accept the just computed values (v^k, q^k) and proceed to the next time step increasing the time step to $\Delta t_{k+1} = \sqrt{\frac{tol}{e_k}}\Delta t_k$. Otherwise, if $e_k > tol$ or $q_k \leq 0$, we reject the approximation (v^k, q^k) and try to recompute it by (28) with a smaller time step $\Delta t_k := \min(\sqrt{\frac{tol}{e_k}}\Delta t_k, \Delta t_k/2)$. If the new approximation (v^k, q^k) is still not sufficiently accurate, we try to diminish the time step again and again until an acceptable approximation (v^k, q^k) is computed and proceed only then to the next time step taking Δt_{k+1} equal to the smallest value of Δt_k used on step k. However, if in the process of reducing Δt_k we come to a time step smaller than $\Delta t_{\min}$ at some $k = k_1$, we abandon the calculation of (v^{k_1}, q^{k_1}) and switch to Step 2 taking the initial values $\bar{v}^{k_1-1} = v^{k_1-1}$, $q_s = q^{k_1-1}$ and setting Δt to the current value of Δt_k.

Step 2. For $k = k_1, k_1 + 1, \ldots$ update $\bar{v}^k$ by (27) keeping $q^k = q_s$, $v^k = 0$ until $\bar{v}^k \geq 0$ at some $k = k_2$. We then abandon the calculation of $\bar{v}$ and switch to Step 3.

Step 3. For $k = k_2 + 1, k_2 + 2, \ldots$ update (v^k, q^k) as in Step 1. If q^k goes again under q_s at some time step, then reintroduce $\bar{v}$ and keep switching between Steps 1 and 2 back and forth.

Generally speaking, a discretization of an ODE using the constant time step is often not optimal concerning the CPU time needed to achieve a desired accuracy. This is especially true for our ODE (11) since the coefficient $n(q(t))$ can change enormously during a simulation. One could try therefore to modify our algorithms by introducing a non constant stepsize chosen at each step using an error indicator. A general recipe for an optimal adaptation of the stepsize, according to [2] is to choose the stepsizes so that the discretization error per time step remains approximately constant during the whole simulation. As is well known, the error per time step in a Euler scheme like that of Algorithm 1, is given to the leading order by $\max(|\ddot{q}(t_k)|, |\ddot{v}(t_k)|)\Delta t_k^2/2$ where $\Delta t_k = t_k - t_{k-1}$ is the stepsize on step k which can be varying. Assuming that (q^{k-1}, v^{k-1}) and (q^k, v^k) are Euler approximations at times t_{k-1} and t_k respectively, the second derivatives of q and v can be computed approximately by deriving the equations in (12) with respect to time and replacing

the missing derivatives by finite differences. This gives

$$\ddot{q}(t_k) \approx -n(q^k)v^k + g(t_k) \tag{29}$$

$$\ddot{v}(t_k) \approx n^2(q^k)v^k - n(q^k)g(t_k) - \frac{n(q^k) - n(q^{k-1})}{\Delta t_k}v^k + \frac{g(t_k) - g(t_{k-1})}{\Delta t_k} \tag{30}$$

Thus, denoting $E_k = \max(|\ddot{q}(t_k)|, |\ddot{v}(t_k)|)/2$ where the second derivatives are replaced by the formulas above, the strategy to choose the time step would be to require $E_k \Delta t_k^2 = tol = const$ where tol is prescribed tolerance. As this recipe can lead to extremely small Δt_k we combine it with the threshold approximation by switching to $\bar{q}(t)$ only when Δt_k becomes smaller than some minimal admissible time step size. This idea is implemented in Algorithm 3.

4.2. **Some numerical tests.** As an illustration, we consider the ODE (11) with $n(q) = \varepsilon/q^{3/2}$ corresponding to the lubrication force in 2D. We report several numerical results setting

$$g(t) = \begin{cases} -2, & \text{for } t \le 2, \\ 2, & \text{for } t > 2, \end{cases} \tag{31}$$

and the initial conditions $q(0) = 1$, $\dot{q}(0) = 0$. In all the cases, we have at our disposal a very accurate reference solution, which we call "exact" in what follows.

Remark 3. The source term $g(t)$ represents an important and relevant test case in view of the possible incorporation of our time marching schemes into the full fluid/particle simulations, as explained in the beginning of Section 4. Note that (31) was also used in [15] to test the gluey particle model.

We first compare Algorithm 1 (without threshold) vs. Algorithm 2 with the threshold such that $n(q_s) = 1/(20\Delta t)$. The choice of the coefficient 20 in the last formula is somewhat arbitrary, but it works fine in our test cases. The first series of numerical experiments is performed taking $\varepsilon = 0.1$. The results are reported in Figure 3. Note that, although the solution obtained with Algorithm 1 is positive at $\Delta t = 0.01$ and smaller, the introduction of the threshold in Algorithm 2 seems to enhance the quality of the solution at large time.

Let us now turn to another series of numerical experiments taking $\varepsilon = 10^{-3}$. The results are reported in Figure 4. Algorithm 1 is now unable to produce a physically acceptable solution even with very small $\Delta t = 0.0001$. Algorithm 2, on the other hand, works fine. In order to observe better the quasi-contact region and the effect of introducing the threshold, we zoom on small distances by passing to a log-scale in Figure 5 (left). We report also some numerical experiments aiming at the determination of an optimal value of the threshold q_s. As said already, our strategy is to choose it so that $n(q_s) = 1/(C\Delta t)$ taking $C = 20$ in all the preceding experiments. Figure 5 (left) illustrates the results obtained fixing $\Delta t = 0.001$ and taking three different values for q_s by varying C. These results confirm two general observations: taking q_s too small may deteriorate the accuracy of the solution (indeed, the original goal of the introduction of q_s was to avoid the extremely small values of q). On the other hand, taking q_s too large may unnecessarily perturb the solution in the regions where a straight discretization could give more precise results. Thus, there should be an optimal choice of q_s which seems to be not too far from the formula mentioned above with $C = 20$.

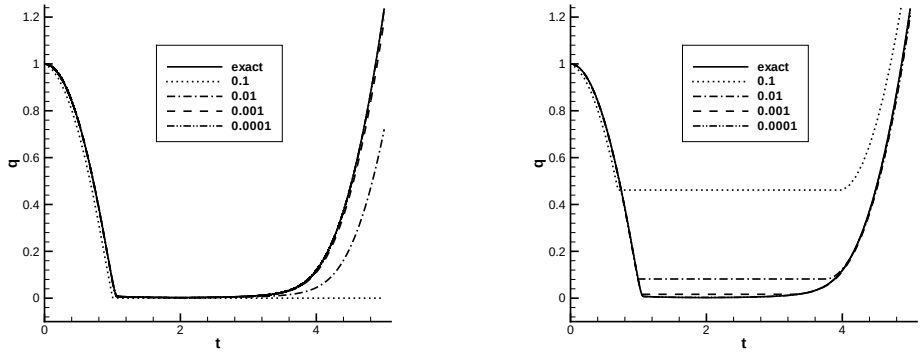


FIGURE 3. Solution to the ODE (11) with $n(q) = \varepsilon/q^{3/2}$, $\varepsilon = 0.1$. Left: the exact solution and solutions obtained by Algorithm 1 with Δt ranging from 0.1 to 0.0001. Right: solutions obtained by Algorithm 2 with Δt in the same range. The results obtained with $\Delta t = 0.0001$ are visually indistinguishable from the exact solution.

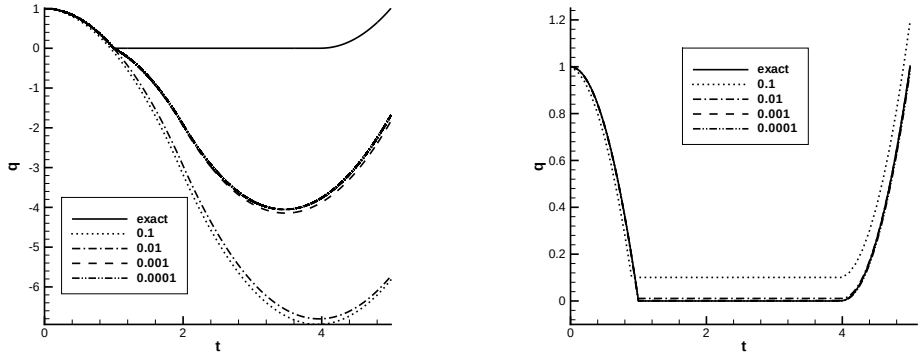


FIGURE 4. Solution to the ODE (11) with $n(q) = \varepsilon/q^{3/2}$, $\varepsilon = 10^{-3}$. Left: the exact solution and solutions obtained by Algorithm 1 with Δt ranging from 0.1 to 0.0001. Right: solutions obtained by Algorithm 2 with Δt in the same range. The results obtained with $\Delta t \leq 0.01$ on the right are visually indistinguishable from the exact solution.

We conclude by a series of experiments using Algorithm 3 with adapted stepsize. We take again the same governing equation with $\varepsilon = 10^{-3}$ and run Algorithm 3 setting the tolerance to $tol = 10^{-5}$ and taking a number of values for $\Delta t_{\min}$. The results are reported in Figure 6 with the evolution of $q(t^k)$ on the left and that of Δt_k on the right. We give in particular the results with $\Delta t_{\min} = 0$ so that Algorithm 3 is reduced to a standard time marching scheme with automatically adapted step sizes. It is interesting to note that step sizes Δt of the order 10^{-4} are actually sufficient to satisfy our error tolerance criterion almost everywhere apart

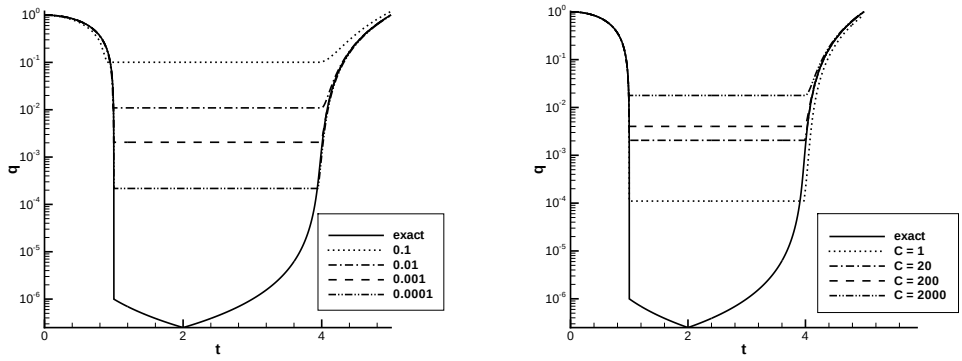


FIGURE 5. Left: same results as on Figure 4, *i.e.* varying Δt and choosing q_s so that $n(q_s) = 1/(20\Delta t)$. Right: same test with $\Delta t = 0.001$ fixed and different values of q_s calculated so that $n(q_s) = 1/(C\Delta t)$ and C varying in the range $[1, 2000]$.

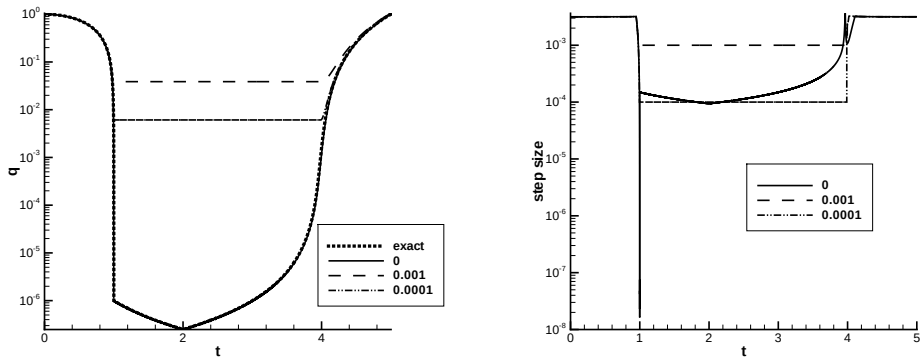


FIGURE 6. Results obtained by Algorithm 3 with $tol = 10^{-5}$ and different values of $\Delta t_{\min}$ specified in the legend. Left: evolution of $q(t^k)$ as compared to the exact solution. Right: evolution of Δt_k.

from a very small range of time around 1 where the Algorithm chooses Δt of the order 10^{-8}. Setting $\Delta t_{\min}$ to 10^{-3} or 10^{-4} avoid such small time steps and gives fairly good results.

Acknowledgements. We would like to thank the organizers of the conference ECM'09 (International Conference of Microfluidics and Complex flow) Lassaad El Asmi and Mourad Ismail for giving us the opportunity to present this work. We had a number of useful discussions during this conference, especially with Aline Lefebvre-Lepot and François Feuillebois, to whom we are very grateful.

Appendix A. **An estimate for the drag (lubrication) force.** This part is devoted to computing the first order expansion of the viscous drag exerted by a fluid on a disk approaching a plane wall. Let $\Omega \subset \mathbb{R}^2$ be the half plane $\Omega =$

$\{(x,y)$ with $y>0\}$, the boundary of Ω is $\partial\Omega=\{(x,0),\ x\in\mathbb{R}\}$ (see Figure 1). We assume that the fluid fills the domain $\mathcal{F}_q:=\Omega\setminus\overline{\mathcal{B}_q}$ where $\mathcal{B}_q:=B((0,R+q),R)$ is the domain occupied by the solid disk of radius $R>0$ and placed at the distance $q>0$ from the boundary $\partial\Omega$. Our notations are lightly different here from that of Section 2. Indeed, domains $\mathcal{F}_q$ and $\mathcal{B}_q$ are now indexed by q rather than t since the eventual dependence of q on time does not interest us in this Appendix. Assume moreover that the disk has a prescribed velocity $\mathbf{V}=V\mathbf{e}_2$ and zero angular velocity. The motion of the fluid is governed by Stokes equations (1) with $\mathbf{g}=0$ and zero boundary conditions at the infinity for $\mathbf{u}$. It is classical that (1) is well-posed in a suitable framework and has a unique solution $(\mathbf{u},p)$ (the pressure p is defined up to a constant). In particular, the viscous drag force $\mathbf{F}$ exerted by a fluid flow $(\mathbf{u},p)$ on $\mathcal{B}_q$ and computed by (2) is well-defined as a function of V and q. Our aim is to prove the following result:

Theorem A.1. *The viscous drag force satisfies* $\mathbf{F}=-n(q)\mathbf{V}$ *where*

$$n(q)=3\sqrt{2}\pi\nu\left(\frac{R}{q}\right)^{\frac{3}{2}}\left[1+\varepsilon(q)\right],$$

with $\varepsilon(q)\to 0$ *when* $q\to 0$.

Remark 4. The equivalent result in the three-dimensional setting is well-known but it is new in two dimensions. Whatever the novelty of this result, the main contribution in this section is that we provide a new way for computing the first order expansion of the viscous drag. This new method is more robust than the known computations [3, 5, 19, 20]. In particular, we claim it extends to the three-dimensional setting with small changes and can be adapted to other boundary conditions. As an illustration, our method is a well-designed tool for estimating the distance between solid bodies in solutions to the fluid-structure interaction system with the full newtonian Navier Stokes equation on the fluid domain [7, 10, 11].

The proof of Theorem A.1 is divided into three steps. First, we recall the variational formulation for the solution $(\mathbf{u},p)$ to (1) and apply it to compute the associated drag coefficient $n(q)$ in the expression for viscous drag $\mathbf{F}=-n(q)\mathbf{V}$. Secondly, we deduce from the variational formulation the lower and upper bounds for $n(q)$. We conclude the proof by computing an asymptotic expansion of the bounds of this range.

1. The variational formulation for the Stokes system (1). As $(\mathbf{u},p)$ in (1) and the formula (2) for the drag depend linearly on $\mathbf{V}$, we set $\mathbf{V}=\mathbf{e}_2$ in what follows. It is classical to extend the velocity $\mathbf{u}$ to the whole domain Ω by setting it to $\mathbf{e}_2$ inside $\mathcal{B}_q$ and to look for $\mathbf{u}$ in the following space

$$D_0^{1,2}(\Omega):=\Big\{\mathbf{w}\in L^1_{loc}(\Omega)\text{ with }\nabla\mathbf{w}\in L^2(\Omega),\ \operatorname{div}\mathbf{w}=0\text{ in }\Omega\text{ and }\mathbf{w}=0\text{ on }\partial\Omega\Big\}.$$

The solution of (1) is associated with the minimization of Dirichlet integral $\mathbb{D}(\mathbf{w})=\int_\Omega|\nabla(\mathbf{w})(x,y)|^2\mathrm{d}x\mathrm{d}y$ over the subset Y_q of $D_0^{1,2}(\Omega)$:

$$Y_q=\Big\{\mathbf{w}\in D_0^{1,2}(\Omega)\text{ such that }\mathbf{w}_{|\mathcal{B}_q}=\mathbf{e}_2\Big\}.$$

Let us denote this minimum by $E_D(q)$, *i.e.*

$$E_D(q):=\min\left\{\mathbb{D}(\mathbf{w})\ ;\ \mathbf{w}\in Y_q\right\},\tag{32}$$

As the Dirichlet integral $\mathbb{D}$ is a strictly convex functional on Y_q it has a unique minimizer $\mathbf{u}_q$ for any $q > 0$. It is easy to see that $\mathbf{u}_q$ gives the solution of (1). Indeed, we have for any $\mathbf{w} \in \mathcal{C}_c^\infty(\mathcal{F}_q)$ such that div $\mathbf{w} = 0$,

$$\int_\Omega \nabla \mathbf{u}(x,y) : \nabla \mathbf{w}(x,y) \mathrm{d}x\mathrm{d}y = 0,$$

which is the variational formulation of (1). The pressure $p_q \in L^2_{loc}(\Omega)$ can be then recovered as the Lagrange multiplier corresponding to the constraint div $\mathbf{u} = 0$. We notice that p_q is unique up to a constant and the ellipticity of the Stokes problem implies that $(\mathbf{u}_q, p_q)$ is smooth on $\overline{\mathcal{F}}_q$ so that it furnishes a classical solution to (1). We refer to [6] for more details and also for a proof of the converse implication.

Given $\mathbf{w} \in \mathcal{C}_c^\infty(\Omega)$ such that div $\mathbf{w} = 0$ and $\mathbf{w} = \mathbf{e}_2$ on $\mathcal{B}_q$, a straightforward integration by parts yields (recall that $\mathbf{n}$ is the unit normal looking into $\mathcal{F}_q$):

$$\begin{aligned}\nu \int_\Omega \nabla \mathbf{u}_q(x,y) : \nabla \mathbf{w}(x,y) \mathrm{d}x\mathrm{d}y &= 2\nu \int_{\mathcal{F}_q} D(\mathbf{u}_q)(x,y) : \nabla \mathbf{w}(x,y) \mathrm{d}x\mathrm{d}y \\ &= -\int_{\partial \mathcal{B}_q} [2\nu D(\mathbf{u}_q) - p_q I_2]\mathbf{n} \mathrm{d}\sigma \cdot \mathbf{e}_2.\end{aligned}$$

Taking a suitable family of approximation of $\mathbf{u}_q$ (see [6] for details) we obtain in the limit:

$$\nu E_D(q) = -\int_{\partial \mathcal{B}_q} [2\nu D(\mathbf{u}_q) - p_q I_2]\mathbf{n} \mathrm{d}\sigma \cdot \mathbf{e}_2.$$

The drag $\mathbf{F}$ is parallel to the velocity $\mathbf{V}$ for symmetry reasons, *i.e.* $\mathbf{F} = (\mathbf{F} \cdot \mathbf{e}_2)\mathbf{e}_2$. We have thus the following result.

Lemma A.2. *For any given $q > 0$ and $\mathbf{V}$ parallel to $\mathbf{e}_2$, the drag force is given by $\mathbf{F} = -n(q)\mathbf{V}$ with $n(q) = \nu E_D(q)$.*

2. Upper and lower bounds for $E_D(q)$. We apply the above variational formulation to bound $E_D(q)$. To this end, given $q > 0$, we first prove that the Dirichlet integral of any $\mathbf{u} \in Y_q$ is greater than some $m_D(q)$ depending only on q. This gives us a lower bound for $E_D(q)$. Then, we construct a suitable $\tilde{\mathbf{u}}_q \in Y_q$ and compute its Dirichlet integral $M_D(q)$. This gives an upper bound for $E_D(q)$. We finally compare m_D and M_D for small values of q to prove Theorem A.1.

As the scale invariance of our problem implies that $n(\cdot)$ is actually a function of q/R, it is sufficient to consider the case $R = 1$, which we admit in what follows. For any $x \in (-1,1)$, $y < 1+q$ such that $(x,y) \in \partial\mathcal{B}_q$ there holds $y = 1+q-\sqrt{1-x^2} := \delta_q(x)$ and we denote

$$\Omega_q := \{(x,y) \in \mathbb{R}^2 \text{ such that } x \in (-1/2, 1/2) \text{ and } y \in (0, \delta_q(x))\}.$$

Given a smooth $\mathbf{u} \in Y_q$, we introduce

$$\psi(x,y) = -\int_0^y u_1(x,z)\mathrm{d}z.$$

Then, straightforward computations imply that $\psi \in \mathcal{C}^\infty(\overline{\mathcal{F}}_q)$ and satisfy $\mathbf{u} = \nabla^\perp \psi = (-\partial_y\psi, \partial_x\psi)$, so that we have the boundary conditions:

$$\left\{ \begin{array}{rcll} \psi(x,y) & = & 0, \quad \partial_y\psi(x,y) = 0, & \text{on } \partial\Omega, \\ \partial_x\psi(x,y) & = & 1, \quad \partial_y\psi(x,y) = 0, & \text{on } \partial\mathcal{B}_q. \end{array} \right. \tag{33}$$

On the other hand, we compute the Dirichlet integral $\mathbb{D}(\mathbf{u})$ with respect to ψ. This yields:

$$\mathbb{D}(\mathbf{u}) \geq \int_{\Omega_q} |\partial_{yy}\psi(x,y)|^2 \mathrm{d}x\mathrm{d}y.$$

We denote $I(\psi)$ the integral on the right-hand side of the above inequality. Thus, $\mathbb{D}(\mathbf{u})$ is greater than the minimum of $I(\psi)$ over smooth ψ satisfying the boundary conditions (33). This minimum is computed in the following lemma:

Lemma A.3. *Given $q > 0$ and $\psi \in \mathcal{C}^\infty(\overline{\Omega}_q)$, satisfying boundary conditions (33) there holds $I(\psi) \geq I(\psi_q)$ where :*

$$\psi_q(x,y) = x\left[\frac{y}{\delta_q(x)}\right]^2\left(3 - 2\left[\frac{y}{\delta_q(x)}\right]\right), \quad \forall\,(x,y) \in \overline{\Omega}_q.$$

Proof. Given any $\psi(x,y) \in \mathcal{C}^\infty(\overline{\Omega}_q)$, satisfying (33), boundary conditions on $\partial\mathcal{B}_q$ imply $\psi(x,\delta_q(x)) = C + x$, for all $x \in (-1/2, 1/2)$, with some $C \in \mathbb{R}$. Thus ψ satisfies for all $x \in (-1/2, 1/2)$

$$\psi(x,0) = 0, \quad \psi(x,\delta_q(x)) = C + x, \quad \frac{\partial\psi}{\partial y}(x,0) = 0, \quad \frac{\partial\psi}{\partial y}(x,\delta_q(x)) = 0.$$

Then, for arbitrary $x \in (-1/2, 1/2)$, with the possible exception of $x = -C$, we introduce $\chi_x(t) = \psi(x, t\delta_q(x))/(x + C)$, $t \in (0,1)$, which satisfies

$$\chi_x(0) = 0, \quad \chi_x(1) = 1, \quad \chi_x'(0) = 0, \quad \chi_x'(1) = 0. \tag{34}$$

By a standard optimization argument we see that the minimum of $\int_0^1 |\chi_x''(t)|^2 dt$ over all smooth $\chi_x(t)$ satisfying (34) is attained on the function $\eta(t)$ such that $\eta^{(4)}(t) = 0$ on $(0,1)$, which is given by $\eta(t) = t^2(3 - 2t)$. This yields

$$\begin{aligned}\int_{\Omega_q} |\partial_{yy}\psi(x,y)|^2 \mathrm{d}x\mathrm{d}y &= \int_{-1/2}^{1/2}\int_0^1 \frac{(x+C)^2}{|\delta_q(x)|^3}|\chi_x''(t)|^2 \mathrm{d}t\mathrm{d}x \\ &\geq \int_{-1/2}^{1/2}\int_0^1 \frac{(x+C)^2}{|\delta_q(x)|^3}|\eta''(t)|^2 \mathrm{d}t\mathrm{d}x. \end{aligned} \tag{35}$$

Minimizing the last integral with respect to C, we obtain $C = 0$. The last inequality becomes equality if we take $\psi(x,y) = \psi_q(x,y) = x\eta(y/\delta_q(x))$. □

Finally, a lower bound is given by

$$m_D(q) = \int_{\Omega_q} |\partial_{yy}\psi_q(x,y)|^2 \mathrm{d}x\mathrm{d}y.$$

On the other hand, it is always possible to extend ψ_q on the whole Ω_q to some $\tilde{\psi}_q \in \mathcal{C}^\infty(\mathcal{F}_q) \cap \mathcal{C}^\infty(\mathcal{B}_q) \cap \mathcal{C}(\Omega)$ and such that $\tilde{\mathbf{u}}_q := \nabla\tilde{\psi}_q \in Y_q$. For instance one might interpolate (in the x-variable) ψ_q with a suitable truncation of $\psi_0(x,y) = x$. As, outside Ω_q, the solid disk remains at a positive distance of $\partial\Omega$, the truncation and interpolation function can be made independent of q. In particular there holds:

$$\begin{aligned}&\int_\Omega |\nabla\tilde{\mathbf{u}}_q(x,y)|^2 \mathrm{d}x\mathrm{d}y \\ &\qquad = m_D(q) + r(q) + \int_{\Omega_q}\left[|\partial_{xx}\psi_q(x,y)|^2 + |\partial_{xy}\psi_q(x,y)|^2\right]\mathrm{d}x\mathrm{d}y := M_D(q),\end{aligned}$$

with r a bounded function of q.

3. Asymptotic expansion of $m_D(q)$ and $M_D(q)$.

Lemma A.4. *There holds, for small values of q :*

$$m_D(q) = 3\sqrt{2}\pi \frac{1+\varepsilon(q)}{q^{3/2}}, \tag{36}$$

with $\varepsilon(q) \to 0$ when $q \to 0$.

Proof. As already seen in the proof of the preceding Lemma, $m_D(q)$ is given by the integral (35) with $C = 0$ and $\eta(t) = t^2(3-2t)$. Integrating with respect to t and performing a change of variables $x \to \sqrt{q}x$ yields

$$m_D(q) = 12\int_{-1/2}^{1/2} \frac{x^2}{|\delta_q(x)|^3}\mathrm{d}x = \frac{12}{q^{3/2}}\int_{-1/(2\sqrt{q})}^{1/(2\sqrt{q})} \frac{x^2\mathrm{d}x}{(\delta_q(\sqrt{q}x)/q)^3}$$

Expanding to the first order, we have:

$$\lim_{q\to 0} \frac{x^2}{(\delta_q(\sqrt{q}x)/q)^3} = \lim_{q\to 0} \frac{x^2}{[(1+q-\sqrt{1-qx^2})/q]^3} = \frac{x^2}{(1+x^2/2)^3}.$$

We remark moreover that standard geometric arguments imply that $\delta_q(x) \geq q + x^2/2$ or $\delta_q(\sqrt{q}x)/q \geq 1 + x^2/2$ so that we can apply the Lebesgue theorem to obtain:

$$\lim_{q\to 0} \int_{-1/(2\sqrt{q})}^{1/(2\sqrt{q})} \frac{x^2\mathrm{d}x}{(\delta_q(\sqrt{q}x)/q)^3} = \int_{-\infty}^{\infty} \frac{x^2\mathrm{d}x}{(1+x^2/2)^3} = \frac{\pi}{2\sqrt{2}}.$$

□

One can also check that $\partial_{yy}\psi_q$ is the second derivative of ψ_q which diverges the fastest when q goes to 0. Thus, $M_D(q) - m_D(q) = o(m_D(q))$ as $q \to 0$ so that (36) holds also for $M_D(q)$, which proves Theorem A.1.

Remark 5. The same method can be easily adapted to the three dimensional case. More precisely, the result of Lemma A.2 is still valid : the viscous drag exerted by a fluid on a sphere can be computed as the result of a minimization problem of the same type as the one given by equation (32). Next, since the problem is axisymmetric, one can use cylindrical coordinates and perform the same asymptotical analysis as in the 2D case, in order to obtain that

$$n(q) = 6\pi V\nu \frac{R^2}{q}\left[1+\varepsilon(q)\right],$$

with $\varepsilon(q) \to 0$ when $q \to 0$.

Numerical evaluation of the drag force. In order to compare the theoretical estimate given by Proposition A.1 to a numerical evaluation of the force $\mathbf{F}$, we carry out the following simulations: we begin with solving the Stokes problem in a fluid with viscosity $\nu = 1$ and then compute the corresponding force exerted by the fluid on a particle of radius $R = 0.1$ situated above a plane at a distance q. The computations are performed by using the Finite-Element solver FreeFem++ [9] and the results are plotted in Figure 7 (dashed line with squares). They agree with the asymptotic expansion $F \sim 3\sqrt{2}\pi V\nu \left(\frac{R}{q}\right)^{\frac{3}{2}}$ (solid line).

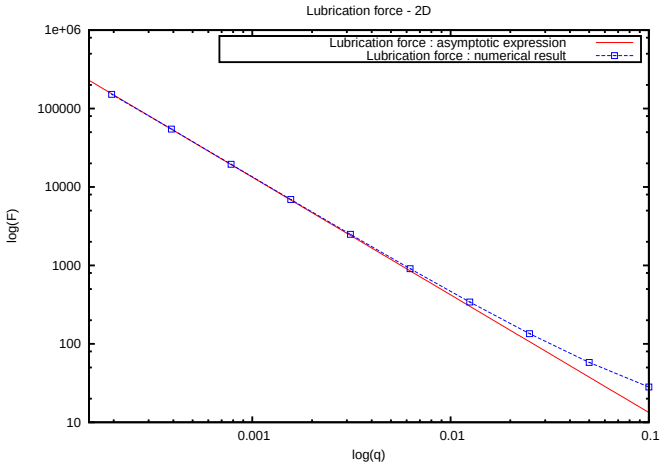


FIGURE 7. Approximation of the lubrication force

Appendix B. **Non-dimensional form of the governing equations (8) involving the lubrication force.** Let us first consider the 2D case. Eliminating the velocity v from (8) and specifying $m = \rho_s \pi R^2$ with ρ_s the density of the solid, we arrive at

$$\rho_s \pi R^2 \ddot{q} = -3\sqrt{2}\pi\nu \left(\frac{R}{q}\right)^{\frac{3}{2}} \dot{q} + (\rho_s - \rho_f)\pi R^2 g.$$

In order to understand the typical values of parameters in this ODE, we should pass to non-dimensional variables, which can be introduced as follows:

$$q = Rq', \ t = Tt', \ g = g_{char} g'$$

where the radius R is used as the length scale, the time scale is denoted by T and g_{char} is a typical value of the external force density, so that $g_{char} \sim 10 m/s^2$. We choose then the time scale T so that the non-dimensional external force becomes of order 1, *i.e.* $\frac{T^2}{R}\frac{\rho_s - \rho_f}{\rho_s} g_{char} = 1$ and obtain the non-dimensional ODE (dropping the primes) (9).

The asymptotic in the 3D case is is given by $n(q) = 6\pi\nu\frac{R^2}{q}$ (cf. Remark 5) so that starting from the dimensional ODE

$$\rho_s \frac{4}{3}\pi R^3 \ddot{q} = -6\pi\nu R^2 \frac{\dot{q}}{q} + (\rho_s - \rho_f)\frac{4}{3}\pi R^3 g \qquad (3D).$$

and using the similar non-dimensionalizations as before we obtain (10).

REFERENCES

[1] J. F. Brady and G. Bossis, *Stokesian dynamics*, Annual Review of Fluid Mechanics, **20** (1988), 111–157.

[2] J. Butcher, "Numerical methods for ordinary differential equations," John Wiley & Sons Ltd., Chichester, 2003.

[3] M. Cooley and M. O'Neill, *On the slow motion generated in a viscous fluid by the approach of a sphere to a plane wall or stationary sphere.*, Mathematika, **16** (1969), 37–49.

[4] S. L. Dance and M. R. Maxey, *Incorporation of lubrication effects into the force-coupling method for particulate two-phase flow*, Journal of Computational Physics, **189** (2003), 212–238.

[5] W. R. Dean and M. E. O'Neill, *A slow motion of viscous liquid caused by the rotation of a solid sphere*, Mathematika, **10** (1963), 13–24.
[6] G. P. Galdi, "An introduction to the mathematical theory of the Navier-Stokes equations. Vol. I, Linearized steady problems," vol. 38 of Springer Tracts in Natural Philosophy, Springer-Verlag, New York, 1994.
[7] D. Gérard-Varet and M. Hillairet, *Regularity issues in the problem of fluid structure interaction*, Arch. Ration. Mech. Anal., **195** (2010), 375–407.
[8] R. Glowinski, T.-W. Pan, T. I. Hesla, and D. D. Joseph, *A distributed lagrange multiplier/fictitious domain method for particulate flows*, Int. J. Multiphase Flow, **24** (1999), 755–794.
[9] F. Hecht, O. Pironneau, A. L. Hyaric, and K. Ohtsuka, *Freefem++, ver. 3.7*, http://www.freefem.org/ff++, (2009).
[10] M. Hillairet, *Lack of collision between solid bodies in a 2D incompressible viscous flow*, Comm. Partial Differential Equations, **32** (2007), 1345–1371.
[11] M. Hillairet and T. Takahashi, *Collisions in three-dimensional fluid structure interaction problems*, SIAM J. Math. Anal., **40** (2009), 2451–2477.
[12] M. S. Ingber, A. Mammoli, P. Vorobieff, T. McCollam, and A. Graham, *Experimental and numerical analysis of irreversibilities among particles suspended in a couette device*, Journal of Rheology, **50** (2006), 99–114.
[13] N. Lecocq, R. Anthore, B. Cichocki, P. Szymczak, and F. Feuillebois, *Drag force on a sphere moving towards a corrugated wall*, J. Fluid Mech., **513** (2004), 247–264.
[14] N. Lecocq, F. Feuillebois, N. Anthore, R. Anthore, F. Bostel, and C. Petipas, *Precise measurement of particle-wall hydrodynamic interactions at low reynolds number using laser interferometry*, Phys. Fluids A, **5** (1993), 3–12.
[15] A. Lefebvre, *Numerical simulation of gluey particles*, M2AN Math. Model. Numer. Anal., **43** (2009), 53–80.
[16] A. Lozinski and M. Romerio, *Motion of gas bubbles, considered as massless bodies, affording deformations within a prescribed family of shapes, in an incompressible fluid under the action of gravitation and surface tension*, M3AS Math. Mod. Meth. Appl. Sci., **17** (2007), 1445–1478.
[17] A. A. Mammoli, *The treatment of lubrication forces in boundary integral equations*, Royal Society of London Proceedings Series A, **462** (2006), 855–881.
[18] B. Maury, *A gluey particle model*, ESAIM: Proceedings, **18** (2007), 133–142.
[19] M. O'Neill, *A slow motion of viscous liquid caused by a slowly moving solid sphere*, Mathematika, **11** (1964), 67–74.
[20] M. O'Neill and K. Stewartson, *On the slow motion of a sphere parallel to a nearby plane wall*, J. Fluid Mech., **27** (1967), 705–724.
[21] L. Pasol, M. Chaoui, S. Yahiaoui, and F. Feuillebois, *Analytical solutions for a spherical particle near a wall in axisymmetrical polynomial creeping flows*, Phys. Fluids, **17** (2005), 1–13.
[22] F. Qi, N. Phan-Tien, and X. J. Fan, *Effective moduli of particulate solids: The completed double layer boundary element method with lubrication approximation*, Zeitschrift Angewandte Mathematik und Physik, **51** (2000), 92–113.
[23] A. Sierou and J. F. Brady, *Accelerated Stokesian Dynamics simulations*, Journal of Fluid Mechanics, **448** (2001), 115–146.
[24] J. Smart and D. Leighton, *Measurements of the hydrodynamic roughness of non colloidal spheres*, Phys. Fluids A, **1** (1989), 52–60.
[25] O. Vinogradova and G. Yakubov, *Surface roughness and hydrodynamic boundary conditions*, Phys. Rev. E, **73** (2006), p. 045302(R).

Received xxxx 20xx; revised xxxx 20xx.

E-mail address: hillairet@ceremade.dauphine.fr
E-mail address: alexei.lozinski@math.univ-toulouse.fr
E-mail address: szopos@math.unistra.fr

MOTION OF GAS BUBBLES, CONSIDERED AS MASSLESS BODIES, AFFORDING DEFORMATIONS WITHIN A PRESCRIBED FAMILY OF SHAPES, IN AN INCOMPRESSIBLE FLUID UNDER THE ACTION OF GRAVITATION AND SURFACE TENSION

par

ALEXEI LOZINSKI AND MICHEL ROMERIO

paru dans

Mathematical Models and Methods in Applied Sciences

Mathematical Models and Methods in Applied Sciences
Vol. 17, No. 9 (2007) 1445–1478
© World Scientific Publishing Company

World Scientific
www.worldscientific.com

MOTION OF GAS BUBBLES, CONSIDERED AS MASSLESS BODIES, AFFORDING DEFORMATIONS WITHIN A PRESCRIBED FAMILY OF SHAPES, IN AN INCOMPRESSIBLE FLUID UNDER THE ACTION OF GRAVITATION AND SURFACE TENSION

ALEXEI LOZINSKI

Department of Mathematics, University of Houston,
4800 Calhoun Rd., Houston TX 77204-3476, USA

MICHEL V. ROMERIO

Institut d'Analyse et de Calcul Scientifique,
École Polytechnique Fédérale de Lausanne,
1015 Lausanne, Switzerland

Received 26 May 2006
Revised 19 April 2007
Communicated by N. Bellomo

A model allowing to describe motion and coalescence of gas bubbles in a liquid under the action of gravitation and surface tension is proposed. The shape of the bubbles is described by a pre-defined family of mappings, for example ellipsoids with a fixed volume and the effects of the gas motions inside the bubbles are neglected. The motion of a bubble is obtained in a Lagrangian form using the D'Alembert principle of virtual works. The set of equations is numerically solved with the help of the fictitious domain technique in which the Navier–Stokes equations in the domain formed by both fluid and gas are considered. The equations governing the bubbles motion are imposed by introducing Lagrange multipliers on the bubbles boundaries. Numerical results in 2D and 3D are presented.

Keywords: Gas bubbles; Lagrangian description; fictitious domain.

1. Introduction

The problem we are dealing with in this paper originates from the production of aluminum in a reduction cell. This reduction process gives birth to a large amount of carbon oxide which is released in the so-called channels between the anodes (Fig. 1). These anodes, which are formed of a porous agglomerate mainly constituted in carbon, are burnt during the process; the gas produced is stocked into the alveolate structure of this material. The bubbles are generated in small cavities located at the bottom of the anodes. Once they reach some critical size they can

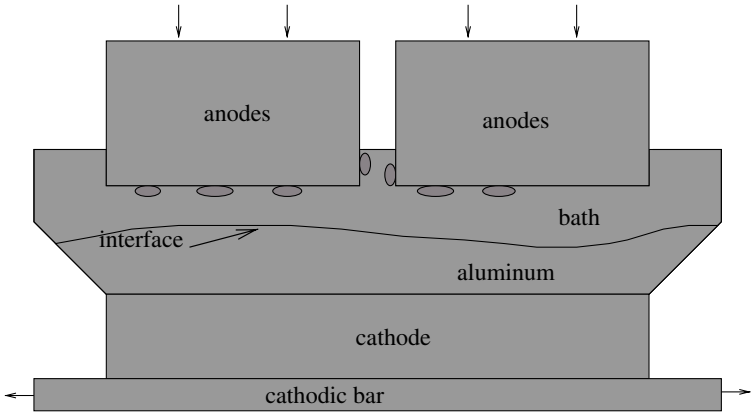


Fig. 1. The aluminum reduction cell.

detach from the surface of the anode and then move freely on a small layer of fluid. This peculiar behavior of bubbles called detachment phenomena has been studied both experimentally and theoretically, see for example Refs. 10 and 12.

As already pointed out in Ref. 6, when, in such a situation, a large set of bubbles, initially uniformly distributed is growing through coalescence only the uniform character is rapidly lost in the course of time. This implies that, in this case, the use of a homogenization method for the description of the gas flux is certainly inappropriate.

It is the purpose of this work to describe the motion of such a gas bubble flux in an incompressible fluid under the action of both gravitation and surface tension. Here the action of the Lorentz forces are disregarded for their effects are small with respect to those resulting from the other forces.

It should be noted that in order to get a description of this nonhomogeneous gas flux an accurate description of each bubble and of its generation is certainly not necessary. Having this in mind we make the assumption that at some initial time all the bubbles which constitute the gas flux, have the same size and are located under a portion of a plane, representing the bottom of the anode. We suppose that this size corresponds to the critical one given by the detachment phenomena, so that the bubbles can move freely on a thin layer of fluid (around 100 microns thick). It is also assumed that they are no longer growing during their motions but by a coalescence procedure.

As already pointed out since our goal is to describe a non-homogeneous gas flux constituted of a large number of bubbles, we make use of a modeling in which the bubble shapes are described by a few parameters only. This means that we renounce an exact description of the bubble shapes but keep what we consider as the main features of the gas flux corresponding to the situation under consideration. This implies that we also disregard methods like level set (see Refs. 11 and 14) or the volume of fluid (see Refs. 3 and 9) which could hardly deal with the large number of bubbles which are present in the problem mentioned above.

Taking these remarks into account we derive our modeling under the following assumptions.

- In the absence of coalescence the gas amount contained in a bubble at some initial time is preserved at later time.
- Bubbles are massless deformable bodies the shapes of which belong to some predefined family of surfaces. In particular, during the deformation processes, all the successive shapes are elements of this family.

It will be shown below that under these assumptions the motion of a bubble is described by a restricted number of equations, namely

- barycenter equations,
- rotation equations (rotation around the barycenter),
- equations describing the body deformations, compatible with the constraints imposed by the belonging to the predefined family of surfaces.

These equations are naturally expressed in terms of the barycenter coordinates, the rotation coordinates, and the parameters describing the predefined family of surfaces subject to the volume conservation constraint. We stick to the incompressibility assumption on the gas bubbles which is justified by the weak variation of the hydraulic pressure in our situation.[10] Another possible assumption would be to consider that pressure P and volume V satisfy the following relation for an adiabatic transformation, $PV^\gamma = \text{const}$, where γ is the adiabatic constant for ideal gas.

The fluid, in which the bubbles are moving, is assumed to be governed by Navier–Stokes equations. Its motion results from the bubbles displacements, which are produced by the Archimedes drag, from the deformations resulting from the surface tension and from some velocity conditions on the boundary of the domain containing the fluid and the bubbles. The mathematical well-possedness of a similar problem is well studied in the case of rigid bodies moving in a fluid with two space variables. Existance of gobal weak solutions is obtained in Ref. 13, however the solution is non-unique after the collision between the bodies.[15] To the best of our knowledge, no similar study exists in the case of gas bubbles immersed in a fluid.

Since the bubble shapes in our modelling belong to a predefined family of surfaces the coalescence phenomena, which takes place when two bubbles are getting into contact, requires a particular treatment. In order to preserve the constraint on the surfaces we assume that the coalescence process starts as soon as two bubbles are sufficiently close. The process then takes places along a few time steps during which the smallest bubble empty itself into the largest one. The intermediate shapes of the two bubbles remain evidently elements of the predefined family of surfaces. Our simple model of coalescence does not take into account the fact that bubbles may stay in contact without immediately coalescing. From the observation performed on industrial cells one knows that large bubbles are generated under the anodes before the release of the gas in the channels. Since we are mainly interested

in the effect of this gas motion on the fluid velocity we assume that this motion is not strongly affected by the fact that some of the bubbles did not coalesce but stayed in contact with a large bubble. In this last situation the volume of gas which is released remains the same whereas the shape of the domain occupied by the two bubbles in contact is only slightly modified.

In order to avoid numerical problems related to the time dependence of the fluid domain we make use of a fictitious domain technique already introduced in Ref. 7 in a similar context. The fluid is thus considered to occupy the fixed domain Ω, namely that of both fluid and gas. A Lagrange multiplier is moreover introduced to take care of the velocity requirement at the boundaries between fluid and gas.

This paper is organized as follows: original governing equations are presented in Sec. 2. Section 3 is devoted to a description of the kinematics and dynamics of a deformable body in the frame of analytical mechanics. The general motion equations corresponding to such a situation are obtained with the help of D'Alembert's principle. Our numerical approach (based on the method of the fictitious domain) is considered in Sec. 4 and our implementation of the coalescence process is explained in Sec. 5. Finally, numerical results in two- and three-dimensional situations are reported in Sec. 6.

2. Original Governing Equations

We consider a system consisting of fluid and gas bubbles. We denote by Ω the domain (not changing in time) occupied by the whole system (fluid and gas) and by $P(t)$ the time-dependent sub-domain of Ω occupied by the gas bubbles. We assume that the fluid is Newtonian with density ρ and viscosity η. It is described by Navier–Stokes equations in the domain $\Omega_l(t) = \Omega \setminus \overline{P(t)}$. Together with the conditions on the boundary $\Gamma = \partial\Omega$ they read:

$$\rho \frac{D\mathbf{u}}{Dt} - \eta\Delta\mathbf{u} + \nabla p = \rho\mathbf{g}, \tag{2.1}$$

$$\nabla \cdot \mathbf{u} = 0, \tag{2.2}$$

$$\mathbf{u}|_\Gamma = \mathbf{0}, \tag{2.3}$$

where $\mathbf{u}$ and p are respectively the velocity and the pressure in the fluid and $\mathbf{g}$ is some prescribed body force, such as gravity. D/Dt stands for the material derivative, i.e. $D/Dt = \partial/\partial t + \mathbf{u} \cdot \nabla$.

The gas in the bubbles is supposed to be inviscid, isothermal and incompressible with density ρ_g. Its dynamics is thus governed by the Euler equations

$$\rho_g \frac{D\mathbf{v}}{Dt} + \nabla P = \rho_g\mathbf{g}, \tag{2.4}$$

$$\nabla \cdot \mathbf{v} = 0, \tag{2.5}$$

where $\mathbf{v}$ and P are respectively the velocity and the pressure in the gas.

In order to close the system we should add some relations on the liquid–gas interface $\partial P(t)$. The choice of these relations depends on the chemical properties of both the gas and the liquid. One can consider two extreme cases:

- "Clean" bubbles. We impose the immiscibility

$$\mathbf{u}\cdot\mathbf{n}|_{\partial P(t)} = \mathbf{v}\cdot\mathbf{n}|_{\partial P(t)} \tag{2.6}$$

and the force balance

$$\sigma_f\mathbf{n}|_{\partial P(t)} = (-P\mathbf{n} + \alpha H\mathbf{n})|_{\partial P(t)}, \tag{2.7}$$

where $\mathbf{n}$ is the unit normal to $\partial P(t)$ pointing towards the fluid domain, α is the surface tension coefficient, H is the mean curvature of $\partial P(t)$ and σ_f denotes the Cauchy stress tensor in the fluid

$$\sigma_f = -p\mathbf{I} + \eta(\nabla\mathbf{u} + \nabla\mathbf{u}^T). \tag{2.8}$$

Equation (2.7) thus expresses the equilibrium between the hydrodynamic force in the liquid, the force produced by the pressure inside the bubble and the surface tension.

- "Dirty" bubbles, i.e. those containing a surface active film which makes them behave very much like a rigid body. We impose therefore the no-slip condition in this case:

$$\mathbf{u}|_{\partial P(t)} = \mathbf{v}|_{\partial P(t)}. \tag{2.9}$$

Requiring in addition the force balance Eq. (2.7) would lead to an over-determined system. We should choose instead some "relaxed" interface conditions. To this end, we revisit the derivation of the motion equation of an ideal gas immersed into a fluid.

Take any volume ω inside $\overline{P(t)}$ and write the momentum equation for the gas contained in it, namely,

$$\frac{d}{dt}\int_\omega \rho_g\mathbf{v}d\mathbf{x} = \int_\omega \rho_g\mathbf{g}d\mathbf{x} - \int_{\partial\omega\cap Int(P(t))} P\mathbf{n}ds + \int_{\partial\omega\cap\partial P(t)} (\sigma_f\mathbf{n} - \alpha H\mathbf{n})ds$$

or

$$\int_\omega \left(\rho_g\frac{D\mathbf{v}}{Dt} + \nabla P\right)\cdot\mathbf{w}d\mathbf{x} = \int_\omega \rho_g\mathbf{g}\cdot\mathbf{w}d\mathbf{x} + \int_{\partial\omega\cap\partial P(t)} (P\mathbf{n} + \sigma_f\mathbf{n} - \alpha H\mathbf{n})\cdot\mathbf{w}ds$$

for any vector $\mathbf{w}$.

Let ω_i, $i = 1,\ldots,N$ be a partition of $P(t)$. Taking, in each ω_i, a vector $\mathbf{w}_i$ in the above formula, we sum up the contributions of all the subdomains. At the limit where the number of subdomains tends to infinity and where their respective diameters tend to zero, we get

$$\int_{P(t)} \left(\rho_g\frac{D\mathbf{v}}{Dt} + \nabla P\right)\cdot\mathbf{w}d\mathbf{x} = \int_{P(t)} \rho_g\mathbf{g}\cdot\mathbf{w}d\mathbf{x} + \int_{\partial P(t)} (P\mathbf{n} + \sigma_f\mathbf{n} - \alpha H\mathbf{n})\cdot\mathbf{w}ds,$$

which holds for any vector function $\mathbf{w}$ in $H^1(P(t))$. If we require furthermore that $\nabla \cdot \mathbf{w}$ vanishes in $P(t)$ and integrate by parts, we obtain

$$\int_{P(t)} \rho_g \frac{D\mathbf{v}}{Dt} \cdot \mathbf{w} d\mathbf{x} = \int_{P(t)} \rho_g \mathbf{g} \cdot \mathbf{w} d\mathbf{x} + \int_{\partial P(t)} (\sigma_f \mathbf{n} - \alpha H \mathbf{n}) \cdot \mathbf{w} ds, \tag{2.10}$$

which is the weak form of the Euler equations with the boundary conditions suited for the dirty bubbles.

In what follows we will deal only with the second option (2.9)–(2.10).

In view of constructing an approximation for the bubble dynamics, we adopt from now on a Lagrangian viewpoint on the equations of gas motion (2.10) and (2.5). We restrict ourselves to the case of a single bubble, the generalization to the case of multiple bubbles being straightforward if one disregards the coalescence which will be considered later. Let us index the gas particles by the parameter $\boldsymbol{\chi}$ running over a fixed connected domain $X \subset R^3$, i.e. consider the set of mappings

$$\begin{aligned} \mathcal{M} = \{\mathbf{x} : X \to P \,|\, & P \text{ is a connected domain in } R^3, \\ & \mathbf{x} \text{ is one-to-one and onto as well as volume preserving}\}, \end{aligned} \tag{2.11}$$

and let $\mathbf{x}(\cdot, t) : X \to P(t)$ be the mappings such that:

$$\begin{aligned} \mathbf{x}(\boldsymbol{\chi}; 0) &= \mathbf{x}_0(\boldsymbol{\chi}), \\ \partial_t \mathbf{x}(\boldsymbol{\chi}, t) &= \mathbf{v}(\mathbf{x}(\boldsymbol{\chi}, t), t), \end{aligned} \tag{2.12}$$

where $\mathbf{x}_0 \in \mathcal{M} : X \to P(0)$ is some arrangement of the initial configuration of gas particles. We see by (2.5) that $\mathbf{x}(\cdot, t) \in \mathcal{M}$, the latter set can therefore be called the configurational space of the gas bubble.

Note that $D\mathbf{v}/Dt = \ddot{\mathbf{x}}$ where the dots stand for the derivation in time, hence (2.10) can be rewritten as

$$\int_{P(t)} \rho_g \ddot{\mathbf{x}} \cdot \mathbf{w} d\mathbf{x} = \int_{P(t)} \rho_g \mathbf{g} \cdot \mathbf{w} d\mathbf{x} + \int_{\partial P(t)} (\sigma_f \mathbf{n} - \alpha H \mathbf{n}) \cdot \mathbf{w} ds. \tag{2.13}$$

Recall that (2.13) holds for any $\mathbf{w} \in H^1(P(t))$ such that $\nabla \cdot \mathbf{w} = 0$ in $P(t)$.

3. Approximate Representation of Bubble Dynamics by Deformable Bodies

In this section our purpose is to derive a set of equations which describes the motion of a bubble approximated by a deformable body whose possible deformations are described by a finite number of parameters. In order to motivate such an approximation we start from a rather non-rigorous remark. Following the spirit of Ref. 1, the configurational space $\mathcal{M}$ can be formally regarded as an "infinite dimensional manifold" with the tangent space $T\mathcal{M}_{\mathbf{x}}$ at any $\mathbf{x} \in \mathcal{M} : X \to P$ consisting of the

divergence free vector fields over P. Hence, Eq. (2.13) can be formally rewritten as the relation

$$\int_{P(t)} \rho_g \ddot{\mathbf{x}} \cdot \delta\mathbf{x} d\mathbf{x} = \int_{P(t)} \rho_g \mathbf{g} \cdot \delta\mathbf{x} d\mathbf{x} + \int_{\partial P(t)} (\sigma_f \mathbf{n} - \alpha H \mathbf{n}) \cdot \delta\mathbf{x} ds, \tag{3.1}$$

which should hold for any element $\delta\mathbf{x}$ of the tangent space $T\mathcal{M}_{\mathbf{x}(t)}$. Our idea is then to choose a subset $\mathcal{M}^n$ of $\mathcal{M}$, which can be considered as a (finite-dimensional) submanifold of $\mathcal{M}$, and look for the approximate solution $\mathbf{x}(\cdot, t) \in \mathcal{M}^n$ such that (3.1) holds for any $\delta\mathbf{x} \in T\mathcal{M}^n_{\mathbf{x}(t)}$. The latter manifold can be thought of as a configurational manifold of a deformable body.

The first step in this direction consists of getting an expression for the position vector of an arbitrary point of the deformable body. In order to get such an expresion we are inspired by the classical description of a rigid body in which a frame of reference is bound to the body. The positions of the different points are then fixed, with respect to this moving frame, and the position of this last one is described, from a fixed frame of reference, by a translation and a rotation around the origin, usually the barycenter, of the moving frame.

In the general case the deformation is described in a frame of reference the origin of which is the barycenter of the deformable body. A general displacement is then obtained by a translation and a rotation as in the case of the solid body.

Let us introduce some notations.

3.1. *Notations*

- $\mathbf{G}$: barycenter of the deformable body; $\mathbf{G}$ is independent of the mass density which is supposed to be homogeneous.
- $\{O, \mathbf{e}_i,\ i = 1, 2, 3\}$, an orthonormal frame of reference, with origin O, of the three-dimensional Euclidean space.
- $\{\mathbf{G}, \boldsymbol{\epsilon}_j,\ j = 1, 2, 3\}$, a moving orthonormal frame of reference with origin $\mathbf{G}$.
- $\mathbf{f}(\boldsymbol{\chi}; q_1, \ldots, q_n)$, where $\boldsymbol{\chi}$ runs over the domain X of R^3 : the family of mappings, depending on the parameters $q_1, \ldots, q_n$. We suppose that the mappings are volume preserving, i.e.

$$det \left| \frac{D\mathbf{f}(\boldsymbol{\chi}; q_1, \ldots, q_n)}{D\boldsymbol{\chi}} \right| = 1.$$

  Moreover

$$\int_X \mathbf{f}(\boldsymbol{\chi}; q_1, \ldots, q_n) d\boldsymbol{\chi} = 0. \tag{3.2}$$

- $P(t)$: domain occupied by the deformable body. $P(t)$ is supposed to belong to the family $\mathbf{f}$ taken in the moving frame $\{\mathbf{G}, \boldsymbol{\epsilon}_j,\ j = 1, 2, 3\}$ at any time t.
- $\Omega_l(t) = \Omega \setminus \overline{P(t)}$: domain occupied by the fluid.

3.2. *The configuration manifold*

Let $f^j(\boldsymbol{\chi}; q_1, \ldots, q_n)$ be the components of $\mathbf{f}(\boldsymbol{\chi}; q_1, \ldots, q_n)$. With the above notations, any point $\mathbf{x} \in P(t)$ can be expressed in the form:

$$\mathbf{x} = \mathbf{G} + \sum_{j=1}^{3} f^j(\boldsymbol{\chi}; q_1, \ldots, q_n)\boldsymbol{\epsilon}_j \tag{3.3}$$

for some values of $\boldsymbol{\chi} \in X$ and $q_1, \ldots, q_n$.

We introduce some notations.

- $G^i = q_{n+i}$ for $i = 1, 2, 3$,
- $\alpha = q_{n+4}, \beta = q_{n+5}, \gamma = q_{n+6}$, being the Euler's angles describing the rotation from the fixed frame $\{\mathbf{e}_i,\ i = 1, 2, 3\}$ to the moving one $\{\boldsymbol{\epsilon}_i,\ i = 1, 2, 3\}$,
- $\tilde{\mathbf{q}} = \{q_1, \ldots, q_{n+6}\}$,
- $\mathbf{q} = \{q_1, \ldots, q_n\}$.

We can now consider the manifold $\mathcal{M}^n$ of dimension $n+6$ consisting of mappings $\boldsymbol{\chi} \to \mathbf{x}$ of the form (3.3) and depending on $n + 6$ parameters $\tilde{\mathbf{q}}$. By construction $\mathcal{M}^n$ is suited for the description of a body which can afford some prescribed types of deformations, defined by $\mathbf{f}(\cdot; \mathbf{q})$, and which is submitted to some external forces. The last manifold will be called the configuration manifold of the deformable body. The motion of this body is fully determined by a knowledge of the values the parameters $\tilde{\mathbf{q}}$ take in the course of time.

The elements of the tangent space $TM^n_{\tilde{\mathbf{q}}}$ at the point $\tilde{\mathbf{q}}$ are usually called virtual displacements. With (3.3) the elements of $T\mathcal{M}_{\tilde{\mathbf{q}}}$ can be written in the form

$$\sum_{k=1}^{n+6} \frac{\partial \mathbf{x}(\boldsymbol{\chi}; \tilde{\mathbf{q}})}{\partial q_k} \delta q_k, \tag{3.4}$$

where the δq_k, for $k = 1, \ldots, n + 6$, have arbitrary real values. Since the bubbles are assumed to be incompressible the divergence of (3.4) is vanishing

$$\nabla \cdot \frac{\partial \mathbf{x}(\boldsymbol{\chi}; \tilde{\mathbf{q}})}{\partial q_k} = 0, \quad \text{for } k = 1, \ldots, n + 6. \tag{3.5}$$

3.3. *The velocity field*

We now allow $\mathbf{x} \in M^n$ to evolve in time, i.e. $q_1, \ldots, q_{n+6}$ are now regarded as functions of t. In order to perform the time derivation of (3.3) we introduce the angular velocity $\boldsymbol{\omega}(t)$ corresponding to the rotation of the frame $\{\mathbf{G}, \boldsymbol{\epsilon}_j, j = 1, 2, 3\}$. In order to do this, we note that

$$\frac{d}{dt}(\boldsymbol{\epsilon}_i \cdot \boldsymbol{\epsilon}_j) = \dot{\boldsymbol{\epsilon}}_i \cdot \boldsymbol{\epsilon}_j + \boldsymbol{\epsilon}_i \cdot \dot{\boldsymbol{\epsilon}}_j = 0, \quad i, j = 1, 2, 3,$$

hence the matrix $\Omega : \Omega_{i,j} = \boldsymbol{\epsilon}_i \cdot \dot{\boldsymbol{\epsilon}}_j,\ i, j = 1, 2, 3$ is antisymmetric and one can consider the action of the corresponding operator as given by the vector product

with an axial vector $\boldsymbol{\omega}(t)$. In particular,

$$\dot{\boldsymbol{\epsilon}}_i(t) = \Omega\boldsymbol{\epsilon}_i(t) = \boldsymbol{\omega}(t) \wedge \boldsymbol{\epsilon}_i(t). \tag{3.6}$$

We are now ready to derive the expression for the velocity field. Differentiating (3.3) with respect to the time variable and taking (3.6) into account we get

$$\dot{\mathbf{x}} = \mathbf{V} + \boldsymbol{\omega} \wedge \sum_{j=1}^{3} f^j(\boldsymbol{\chi}; \mathbf{q}(t))\boldsymbol{\epsilon}_j + \sum_{j=1}^{3}\sum_{k=1}^{n} \frac{\partial f^j(\boldsymbol{\chi}; \mathbf{q}(t))}{\partial q_k}\dot{q}_k\boldsymbol{\epsilon}_j, \tag{3.7}$$

where $\mathbf{V} = \dot{\mathbf{G}}$.

3.4. *Lagrange equations*

Our next task is thus to derive the equations satisfied by the $q_k,\ k = 1, \ldots, n+6$, considered as functions of t. We recall that the dynamics of our deformable body is governed by the variational principle (3.1) restricted to $\mathcal{M}^n$. More precisely we require that

$$\int_{P(t)} \rho_g \ddot{\mathbf{x}} \cdot \delta\mathbf{x} d\mathbf{x} = \int_{P(t)} \rho_g \mathbf{g} \cdot \delta\mathbf{x} d\mathbf{x} + \int_{\partial P(t)} \mathbf{F} \cdot \delta\mathbf{x} ds, \tag{3.8}$$

holds for any $\delta\mathbf{x} \in \mathcal{M}^n_{\mathbf{x}(t)}$. In (3.8)

$$\mathbf{F} = \sigma_f \mathbf{n} - \alpha H \mathbf{n} \tag{3.9}$$

denotes the resultant of the hydrodynamic forces and of the surface tension.

Remark 3.1. The relation (3.8) can be identified with d'Alembert's principle in classical mechanics.[8] We consider here $\mathbf{F}$ as the exterior force applied to the deformable body $P(t)$ and postulate the existence of some interior constraint forces such that their work on virtual displacements is zero.

Let us now recast (3.8) in the more convenient form of the Lagrange equations. As a first step in this direction we compute the expression of the kinetic energy.

3.4.1. *The kinetic energy*

By definition the kinetic energy is given by the expression

$$E_{kin} = \int_{P(t)} \rho_g \dot{\mathbf{x}}^2 d\mathbf{x} = \int_X \rho_g \dot{\mathbf{x}}^2 d\boldsymbol{\chi}.$$

In order to express E_{kin} in a more usual way we need the following results.

Lemma 3.1. *For any* $\mathbf{V}$, $\boldsymbol{\omega}$, q_k *and* $\dot{q}_k$, $k = 1, \ldots, n$ *we have*

$$\int_X \mathbf{V} \cdot \boldsymbol{\omega} \wedge \mathbf{f}(\boldsymbol{\chi}; \mathbf{q}) d\boldsymbol{\chi} = 0,$$

$$\int_X \mathbf{V} \cdot \sum_{k=1}^{n} \frac{\partial \mathbf{f}(\boldsymbol{\chi}; \mathbf{q})}{\partial q_k}\dot{q}_k d\mathbf{x} = 0.$$

Proof. Making use of (3.2) the first relation is trivial for $\boldsymbol{\omega}$ and $\mathbf{V}$ are both independent of $\boldsymbol{\chi}$. We write the second one in the form

$$\sum_{k=1}^{n} \dot{q}_k \mathbf{V} \cdot \frac{\partial}{\partial q_k} \int_X \mathbf{f}(\boldsymbol{\chi}; \mathbf{q}) d\boldsymbol{\chi},$$

which is also vanishing by (3.2). □

Let

$$\mathbf{L} = \sum_{i,j=1}^{3} \int_X \rho_g f_i(\boldsymbol{\chi}; \mathbf{q}(t)) \boldsymbol{\epsilon}_i \wedge \dot{f}^j(\boldsymbol{\chi}, t) \boldsymbol{\epsilon}_j d\boldsymbol{\chi},$$

where

$$\dot{f}^j(\boldsymbol{\chi}, t) = \sum_{k=1}^{n} \frac{\partial f^j(\boldsymbol{\chi}; \mathbf{q})}{\partial q_k} \dot{q}_k.$$

$\mathbf{L}$ is formally the angular momentum of the deformable body one would get when the basis vectors $\boldsymbol{\epsilon}_j,\ j = 1, 2, 3$, are time independent. Let

$$E_{\text{def}} = \frac{1}{2} \int_X \rho_g |\dot{\mathbf{f}}(\boldsymbol{\chi})|^2 d\boldsymbol{\chi}.$$

E_{def} be the kinetic energy due to the deformation inside the moving frame.

With these definitions we can state the following result.

Proposition 3.1.

$$E_{\text{kin}}(t) = \frac{1}{2} M \mathbf{V}(t)^2 + \frac{1}{2} (\boldsymbol{\omega}, I\boldsymbol{\omega}) + \boldsymbol{\omega} \cdot \mathbf{L} + E_{\text{def}}, \tag{3.10}$$

where M is the mass of the deformable body and I is the tensor of inertia the coefficients of which are given by

$$I_{i,j} = \int_X \rho_g (\mathbf{f}^2(\boldsymbol{\chi}; \mathbf{q}) \delta^{i,j} - f^i(\boldsymbol{\chi}; \mathbf{q}) f^j(\boldsymbol{\chi}; \mathbf{q})) dx.$$

Proof. From Lemma 3.1 when the expression of $\dot{\mathbf{x}}$, given by (3.7), is introduced into E_{kin}, four terms only are remaining, namely:

(1) $\frac{1}{2} \int_X \rho_g \mathbf{V}^{\mathbf{2}}(t) d\boldsymbol{\chi} = \frac{1}{2} \mathbf{V}^{\mathbf{2}}(t) \int_X \rho_g d\boldsymbol{\chi} = \frac{1}{2} M \mathbf{V}^{\mathbf{2}}(t).$

(2) $\frac{1}{2} \int_X \rho_g \Big(\boldsymbol{\omega}(t) \wedge \sum_{i,j=1}^{3} f^i(\boldsymbol{\chi}; \mathbf{q}) \boldsymbol{\epsilon}_i \,.\, \boldsymbol{\omega}(t) \wedge \sum_{i,j=1}^{3} f^j(\boldsymbol{\chi}; \mathbf{q}) \boldsymbol{\epsilon}_i \Big) d\boldsymbol{\chi}$

$= \frac{1}{2} \int_X \rho_g (\boldsymbol{\omega}(t) \wedge \mathbf{f}(\boldsymbol{\chi}; \mathbf{q}) \,.\, \boldsymbol{\omega}(t) \wedge \mathbf{f}(\boldsymbol{\chi}; \mathbf{q})) d\boldsymbol{\chi}$

$= \frac{1}{2} \omega_i(t)\, \omega_j(t) \int_X \rho_g (\mathbf{f}^2(\boldsymbol{\chi}; \mathbf{q}) \delta^{i,j} - f^i(\boldsymbol{\chi}; \mathbf{q}) f^j(\boldsymbol{\chi}; \mathbf{q})) d\boldsymbol{\chi} = \frac{1}{2} \boldsymbol{\omega}(t) \cdot I\boldsymbol{\omega}(t).$

(3) $\frac{1}{2} \int_X \rho_g \Big(\boldsymbol{\omega}(t) \wedge \sum_{i,j=1}^{3} f^i(\boldsymbol{\chi}; \mathbf{q}) \boldsymbol{\epsilon}_i \Big) \cdot \dot{f}^j(\boldsymbol{\chi}; \mathbf{q}) \boldsymbol{\epsilon}_j d\boldsymbol{\chi} = \boldsymbol{\omega}(t) \cdot \mathbf{L}.$

(4) $\frac{1}{2} \int_X \rho_g \sum_{i=1}^{3} \dot{f}^i(\boldsymbol{\chi}; \mathbf{q}) \dot{f}^i(\boldsymbol{\chi}; \mathbf{q}) d\boldsymbol{\chi}.$

We therefore obtain expression (3.10). □

3.4.2. *Generalized forces*

In order to bring the derivation of Lagrange equations to an end we substitute (3.4) into (3.8). This yields

$$\sum_{1}^{n+6}(P_j - Q_j)\delta q_j = 0, \tag{3.11}$$

where, for $j = 1, \ldots, n+6$,

$$Q_j = \int_{P(t)} \rho_g \mathbf{g} \cdot \frac{\partial \mathbf{x}(\boldsymbol{\chi}; \tilde{\mathbf{q}})}{\partial q_j} d\mathbf{x} + \int_{\partial P(t)} \mathbf{F} \cdot \frac{\partial \mathbf{x}(\boldsymbol{\chi}; \tilde{\mathbf{q}})}{\partial q_j} ds,$$

are the generalized forces and

$$P_j = \int_{P(t)} \rho \ddot{\mathbf{x}} \cdot \frac{\partial \mathbf{x}(\boldsymbol{\chi}; \tilde{\mathbf{q}})}{\partial q_j} d\mathbf{x} = \int_X \rho \ddot{\mathbf{x}}(\boldsymbol{\chi}; \tilde{\mathbf{q}}) \cdot \frac{\partial \mathbf{x}(\boldsymbol{\chi}; \tilde{\mathbf{q}})}{\partial q_j} d\boldsymbol{\chi}$$

are the generalized inertial forces.

Since the δq_i, $i = 1, \ldots, n+6$, are independent (3.11) leads to

$$(P_j - Q_j) = 0, \quad j = 1, \ldots, n+6. \tag{3.12}$$

The usual form of Lagrange equations then results from the following lemma.

Lemma 3.2. *Considering the kinetic energy* E_{kin} *as a function of* q_j *and* $\dot{q}_j$, $j = 1, \ldots, n+6$, *we have*

$$P_j = \frac{d}{dt}\frac{\partial E_{\text{kin}}}{\partial \dot{q}_j} - \frac{\partial E_{\text{kin}}}{\partial q_j}, \quad j = 1, \ldots, n+6.$$

Proof.

$$\int_X \rho \ddot{\mathbf{x}}(\boldsymbol{\chi}; \tilde{\mathbf{q}}) \cdot \frac{\partial \mathbf{x}(\boldsymbol{\chi}; \tilde{\mathbf{q}})}{\partial q_j} d\boldsymbol{\chi}$$
$$= \int_X \rho \frac{d}{dt}\left(\dot{\mathbf{x}}(\boldsymbol{\chi}; \tilde{\mathbf{q}}) \cdot \frac{\partial \mathbf{x}(\boldsymbol{\chi}; \tilde{\mathbf{q}})}{\partial q_j}\right) - \int_X \rho \dot{X}(\boldsymbol{\chi}; \tilde{\mathbf{q}}) \cdot \frac{d}{dt}\frac{\partial \mathbf{x}(\boldsymbol{\chi}; \tilde{\mathbf{q}})}{\partial q_j} d\boldsymbol{\chi}.$$

It is an easy matter to see that

$$\frac{\partial \mathbf{x}(\boldsymbol{\chi}; \tilde{\mathbf{q}})}{\partial q_\alpha} = \frac{\partial \dot{\mathbf{x}}(\boldsymbol{\chi}; \tilde{\mathbf{q}})}{\partial \dot{q}_\alpha}$$

and that

$$\frac{d}{dt}\frac{\partial \mathbf{x}(\boldsymbol{\chi}; \tilde{\mathbf{q}})}{\partial q_\alpha} = \frac{\partial \dot{\mathbf{x}}(\boldsymbol{\chi}; \tilde{\mathbf{q}})}{\partial q_\alpha}.$$

The proof then follows by noticing that the integration domain X is time independent. □

From this lemma and from (3.11) the Lagrange equations read

$$\frac{d}{dt}\frac{\partial E_{\text{kin}}}{\partial \dot{q}_i} - \frac{\partial E_{\text{kin}}}{\partial q_i} = Q_i, \quad \text{for } i = 1, 2, \ldots, n+6. \tag{3.13}$$

3.5. *The case of a massless deformable body in a viscous and incompressible fluid*

From now on we make the assumption that the deformable bodies which represent the bubbles have negligible masses. With this assumption the kinetic energy is vanishing so that, from (3.13), the motion equations reduce to

$$Q_j = \int_{\partial P(t)} \mathbf{F} \cdot \frac{\partial \mathbf{x}(\boldsymbol{\chi}; \tilde{\mathbf{q}})}{\partial q_j} ds = 0, \quad \text{for } j = 1, \ldots, n+6.$$

Taking (3.3) and (3.9) into account, these equations can be rewritten as

$$\int_{\partial P(t)} (\sigma_f \mathbf{n} - \alpha H \mathbf{n}) ds = 0, \tag{3.14}$$

$$\int_{\partial P(t)} \mathbf{r} \wedge (\sigma_f \mathbf{n} - \alpha H \mathbf{n}) ds = 0, \tag{3.15}$$

$$\int_{\partial P(t)} \mathbf{f}_k \cdot (\sigma_f \mathbf{n} - \alpha H \mathbf{n}) ds = 0, \quad k = 1, \ldots, n, \tag{3.16}$$

where we make use of the notations

$$\mathbf{r} = \mathbf{x} - \mathbf{G} = \sum_{j=1}^{3} f^j(\boldsymbol{\chi}; \mathbf{q}(t)) \boldsymbol{\epsilon}_j \tag{3.17}$$

and

$$\mathbf{f}_k = \frac{\partial \mathbf{x}(\boldsymbol{\chi}; \widetilde{\mathbf{q}}(t))}{\partial q_k} = \sum_{j=1}^{3} \frac{\partial f^j(\boldsymbol{\chi}; \mathbf{q}(t))}{\partial q_k} \boldsymbol{\epsilon}_j. \tag{3.18}$$

4. The Numerical Approach

Before presenting the numerical scheme let us recapitulate the governing equations and rewrite them in a weak form. We recall that we are looking for the parameters defining the bubble propagation, i.e. $\mathbf{V}(t)$, $\boldsymbol{\omega}(t)$ and $\dot{q}_k(t)$, $k = 1, \ldots, n$, as well as fluid velocity $\mathbf{u}$ and pressure p in $\Omega \setminus \overline{P(t)}$.

The governing equations include the Navier–Stokes equations

$$\rho \int_{\Omega \setminus \overline{P(t)}} \frac{D\mathbf{u}}{Dt} \cdot \mathbf{v} d\mathbf{x} - \int_{\Omega \setminus \overline{P(t)}} p \nabla \cdot \mathbf{v} d\mathbf{x} + \eta \int_{\Omega \setminus \overline{P(t)}} \nabla \mathbf{u} : \nabla \mathbf{v} d\mathbf{x}$$
$$= \int_{\Omega \setminus \overline{P(t)}} \mathbf{g} \cdot \mathbf{v} d\mathbf{x}, \ \ \forall\ \mathbf{v} \in H_0^1(\Omega \setminus \overline{P(t)})$$
$$\int_{\Omega \setminus \overline{P(t)}} \mathbf{u} \cdot \nabla q = 0, \ \ \forall\ q \in L_0^2(\Omega \setminus \overline{P(t)})$$

combined with the boundary conditions

$$\mathbf{u}|_{\Gamma} = \mathbf{0}, \tag{4.1}$$

$$\mathbf{u}|_{\partial P(t)} = \mathbf{V}(t) + \boldsymbol{\omega}(t) \wedge \mathbf{r} + \sum_{k=1}^{n} \mathbf{f}_k \dot{q}_k(t), \tag{4.2}$$

where notations (3.17) and (3.18) are used, and Eqs. (3.14)–(3.16) for the deformable body.

The bubble domain $P(t)$ is given by the parameters $\mathbf{G}(t)$, $\boldsymbol{\epsilon}_j(t)$, $j = 1, 2, 3$ and $q_k(t)$, $k = 1, \ldots, n$ which are to be found from the equations

$$\frac{d\mathbf{G}}{dt} = \mathbf{V}(t), \tag{4.3}$$

$$\frac{d\boldsymbol{\epsilon}_j}{dt} = \boldsymbol{\omega}(t) \wedge \boldsymbol{\epsilon}_j, \quad j = 1, 2, 3, \tag{4.4}$$

$$\frac{dq_k}{dt} = \dot{q}_k(t), \quad k = 1, \ldots, n. \tag{4.5}$$

4.1. *Fictitious domain technique and Lagrange multiplier*

In this subsection we rewrite the Navier–Stokes equations with two technical changes which will be useful for a numerical approach of the solutions. For the numerical reasons mentioned above it is convenient to consider that the domain $P(t)$ is also filled with a fluid having the same density but another stress tensor which is chosen in such a way that the conditions imposed on $\partial P(t)$ are fitted. In other words, the motion of the fluid in Ω_l is not affected by the presence of the other fluid in $P(t)$. This approach is usually called fictitious domain technique (see for example Ref. 7 for more details).

The second change consists in introducing a Lagrange multiplier, we name $\boldsymbol{\lambda}$, which takes care of relation (4.2) relative to the matching of the velocities on $\partial P(t)$. The term containing the Lagrange parameter has the following form:

$$\int_{\partial P(t)} \boldsymbol{\lambda} \cdot \mathbf{v} ds.$$

$\boldsymbol{\lambda}$ can thus be viewed as a force field which allows to satisfy condition (4.2) on $\partial P(t)$.

In view of the above changes the equations for $\mathbf{u}(\cdot, t) \in H_0^1(\Omega)$, $p(\cdot, t) \in L_0^2(\Omega)$ and $\boldsymbol{\lambda} \in H^{-\frac{1}{2}}(\partial P(t))$ take the following form:

$$\rho \int_\Omega \frac{D\mathbf{u}}{Dt} \cdot \mathbf{v} d\mathbf{x} - \int_\Omega p \nabla \cdot \mathbf{v} d\mathbf{x} + \eta \int_\Omega \nabla \mathbf{u} : \nabla \mathbf{v} d\mathbf{x}$$
$$= \int_\Omega \mathbf{g} \cdot \mathbf{v} d\mathbf{x} + \int_{\partial P(t)} \boldsymbol{\lambda} \cdot \mathbf{v} ds, \ \forall\ \mathbf{v} \in H_0^1(\Omega) \tag{4.6}$$

$$\int_\Omega \mathbf{u} \cdot \nabla q = 0, \ \forall\ q \in L_0^2(\Omega) \tag{4.7}$$

$$\int_{\partial P(t)} \mathbf{u} \cdot \boldsymbol{\mu} ds = \int_{\partial P(t)} \left(\mathbf{V}(t) + \boldsymbol{\omega}(t) \wedge \mathbf{r} + \sum_{k=1}^{n} \mathbf{f}_k \dot{q}_k(t) \right) \cdot \boldsymbol{\mu} ds, \ \forall\ \boldsymbol{\mu} \in H^{-\frac{1}{2}}(\partial P(t)). \tag{4.8}$$

The system should be combined again with the force balance Eqs. (3.14)–(3.16) and the propagation Eqs. (4.3)–(4.5).

We now have to specify under which condition the presence of the fluid in the domain $P(t)$ does not affect the motion in Ω_l. In order to do that we split, in Eq. (4.6), the domain of integration Ω into the two sub-domains Ω_l and $P(t)$. Then performing standard integration by part we draw the following condition on the interface $\partial P(t)$

$$\boldsymbol{\lambda} = \left(-p\mathbf{n} + \eta\frac{\partial \mathbf{u}}{\partial \mathbf{n}}\right)\Big|_{\text{fluid}}^{\text{gas}} = (-p\mathbf{n} + \eta(\nabla\mathbf{u})\mathbf{n})|_{\text{fluid}}^{\text{gas}}\,. \tag{4.9}$$

Let us take any point O on the interface and consider the Cartesian coordinates (ξ_1, ξ_2, ξ_3) attached to this point such that the axis ξ_1 is in the direction of the normal vector $\mathbf{n}$ and the axes ξ_2 and ξ_3 are tangential to $\partial P(t)$. Let u_1, u_2, u_3 denote for the moment the components of $\mathbf{u}$ with respect to this frame. We have then (in the same frame):

$$(\nabla\mathbf{u}^{\mathrm{T}})\mathbf{n} = \begin{pmatrix} \dfrac{\partial u_1}{\partial \xi_1} \\ \dfrac{\partial u_1}{\partial \xi_2} \\ \dfrac{\partial u_1}{\partial \xi_3} \end{pmatrix} = \begin{pmatrix} -\dfrac{\partial u_2}{\partial \xi_2} - \dfrac{\partial u_3}{\partial \xi_3} \\ \dfrac{\partial u_1}{\partial \xi_2} \\ \dfrac{\partial u_1}{\partial \xi_3} \end{pmatrix}$$

and we see that $(\nabla\mathbf{u}^{\mathrm{T}})\mathbf{n}$ is continuous for it contains only tangential derivatives of the continuous field $\mathbf{u}$. This remark allows us to rewrite the equation for $\boldsymbol{\lambda}$ as (cf. (2.8))

$$\boldsymbol{\lambda} = (\sigma_g - \sigma_f)\mathbf{n}, \tag{4.10}$$

where σ_g is the fictitious stress tensor in $P(t)$.

4.2. *The gauge reformulation*

The numerical treatment of the Navier–Stokes equations is prone to some well-known difficulties, most notably the need to satisfy the inf-sup condition or to use some stabilization. One of the techniques to get around these difficulties and to construct an efficient time marching scheme is the so-called Gauge method.[5] This approach consists in reformulating the Navier–Stokes equations by introducing the new variables $\mathbf{a}$ and ϕ, which are related to the fluid velocity $\mathbf{u}$ through the relation

$$\mathbf{u} = \mathbf{a} + \nabla\phi \tag{4.11}$$

in such a way that the pressure is eliminated and Laplace-like equations only have to be solved for $\mathbf{a}$ and ϕ on each time step. In order to do this we substitute (4.11) into (4.6) and split the terms there, into two groups, namely

$$\rho\int_\Omega \frac{\partial \mathbf{a}}{\partial t}\cdot\mathbf{v}d\mathbf{x} + \eta\int_\Omega \nabla\mathbf{a} : \nabla\mathbf{v}d\mathbf{x} = \int_{\partial P(t)} \boldsymbol{\lambda}\cdot\mathbf{v}ds - \int_\Omega ((\mathbf{u}\cdot\nabla)\mathbf{u})\cdot\mathbf{v}d\mathbf{x} \tag{4.12}$$

and

$$\rho \int_\Omega \nabla \frac{\partial \phi}{\partial t} \cdot \mathbf{v} d\mathbf{x} - \int_\Omega p \nabla \cdot \mathbf{v} d\mathbf{x} + \eta \int_\Omega \nabla \nabla \phi : \nabla \mathbf{v} d\mathbf{x} = \rho \int_\Omega \mathbf{g} \cdot \mathbf{v} d\mathbf{x}. \tag{4.13}$$

We then require that both relations (4.12) and (4.13) be satisfied for all $\mathbf{v} \in H_0^1(\Omega)$ for some $\mathbf{a}(t, \cdot) \in H^1(\Omega)$ and $\phi(t, \cdot) \in H^2(\Omega)$. In order to see that (4.13) is solvable for ϕ we perform an integration by parts in all the terms of (4.13). In particular,

$$\begin{aligned} \int_\Omega \nabla \nabla \phi : \nabla \mathbf{v} d\mathbf{x} &= \sum_{i,j=1}^3 \int_\Omega \frac{\partial^2 \phi}{\partial x_i \partial x_j} \frac{\partial v_i}{\partial x_j} d\mathbf{x} \\ &= \sum_{i,j=1}^3 \int_\Gamma n_i \frac{\partial \phi}{\partial x_j} \frac{\partial v_i}{\partial x_j} ds - \sum_{i,j=1}^3 \int_\Omega \frac{\partial \phi}{\partial x_j} \frac{\partial^2 v_i}{\partial x_i \partial x_j} d\mathbf{x}. \end{aligned}$$

If we require furthermore that

$$\left. \frac{\partial \phi}{\partial \mathbf{n}} \right|_\Gamma = 0 \tag{4.14}$$

then the boundary term $\sum_{i,j=1}^3 \int_\Gamma n_i \frac{\partial \phi}{\partial x_j} \frac{\partial v_i}{\partial x_j} ds$ vanishes. Indeed, we have at any point on Γ

$$\sum_{i,j=1}^3 n_i \frac{\partial \phi}{\partial x_j} \frac{\partial v_i}{\partial x_j} = \sum_{i=1}^3 \sum_{j=1}^2 n_i (\boldsymbol{\tau}_j \cdot \nabla \phi)(\boldsymbol{\tau}_j \cdot \nabla v_i) = 0,$$

where $\boldsymbol{\tau}_1$ and $\boldsymbol{\tau}_2$ are two orthonormal vectors tangential to Γ. We have thus

$$\int_\Omega \nabla \nabla \phi : \nabla \mathbf{v} d\mathbf{x} = - \int_\Omega \nabla \phi \cdot \nabla (\nabla \cdot \mathbf{v}) d\mathbf{x},$$

so that, in the case where $\mathbf{g}$ has a constant value, Eq. (4.13) is equivalent to

$$\int_\Omega p \nabla \cdot \mathbf{v} d\mathbf{x} = -\rho \int_\Omega \frac{\partial \phi}{\partial t} \nabla \cdot \mathbf{v} d\mathbf{x} - \eta \int_\Omega \nabla \phi \cdot \nabla (\nabla \cdot \mathbf{v}) d\mathbf{x} + \rho \int_\Omega (\mathbf{g} \cdot \mathbf{x})(\nabla \cdot \mathbf{v}) d\mathbf{x}.$$

It is now clear that (4.13) is satisfied if we require that

$$p = -\rho \frac{\partial \phi}{\partial t} + \eta \Delta \phi + \rho \mathbf{g} \cdot \mathbf{x}.$$

Finally, the equation for ϕ is obtained by substituting (4.11) into (4.7) and taking (4.14) into account. This yields

$$\int_\Omega \nabla \phi \cdot \nabla q = - \int_\Omega \mathbf{a} \cdot \nabla q. \tag{4.15}$$

Hence, the Navier–Stokes equations (4.6)–(4.7) are replaced by the relations (4.12) and (4.15) which should hold for any $\mathbf{v} \in H_0^1(\Omega)$ and $q \in H^1(\Omega)$.

The boundary conditions for $\mathbf{a}$ are obtained from (4.1) and (4.14) in the form

$$\mathbf{a} \cdot \boldsymbol{\tau}|_\Gamma = - \left. \frac{\partial \phi}{\partial \boldsymbol{\tau}} \right|_\Gamma, \quad \mathbf{a} \cdot \mathbf{n}|_\Gamma = 0.$$

Substituting (4.11) into (4.8) yields the interface conditions

$$\int_{\partial P(t)} \mathbf{a}\cdot\boldsymbol{\mu} ds = \int_{\partial P(t)} \left(-\nabla\phi + \mathbf{V}(t) + \boldsymbol{\omega}(t) \wedge \sum_{j=1}^{3} f^j \boldsymbol{\epsilon}_j + \sum_{j=1}^{3}\sum_{k=1}^{n} \frac{\partial f^j}{\partial q_k} \dot{q}_k(t) \boldsymbol{\epsilon}_j \right) \cdot \boldsymbol{\mu} ds, \quad \forall\ \boldsymbol{\mu} \in H^{-\frac{1}{2}}(\partial P(t)).$$

4.3. *Force equations*

A direct discretization of the force equations (3.14)–(3.16) would involve some difficulties for the Cauchy tensor contains the normal derivatives of the velocity on $\partial P(t)$, i.e. precisely where these last ones are discontinuous and consequently not straightforward to compute. For this reason we reformulate them in a way where the computation of the Cauchy tensor on the bubbles boundaries can be avoided.

Lemma 4.1. *If the relations* (4.10) *and* (4.12) *hold, then Eqs.* 3.14–3.16 *are equivalent to*

$$\int_{\partial P(t)} \boldsymbol{\lambda} ds = \rho \left(\frac{d}{dt}\int_{P(t)} \mathbf{a} d\mathbf{x} + \frac{d}{dt}\int_{\partial P(t)} \phi \mathbf{n} ds - \mathbf{g}|P(t)| \right) - \alpha \int_{\partial P(t)} H\mathbf{n} ds, \tag{4.16}$$

$$\int_{\partial P(t)} \mathbf{r}\wedge\boldsymbol{\lambda} ds = \rho \left(\frac{d}{dt}\int_{P(t)} \mathbf{r}\wedge\mathbf{a} d\mathbf{x} + \frac{d}{dt}\int_{\partial P(t)} \phi\mathbf{r}\wedge\mathbf{n} ds + \mathbf{V}\wedge\int_{P(t)} \mathbf{u} d\mathbf{x} \right) - \alpha \int_{\partial P(t)} H\mathbf{r}\wedge\mathbf{n} ds \tag{4.17}$$

and

$$\begin{aligned}\int_{\partial P(t)} \mathbf{f}_k\cdot\boldsymbol{\lambda} ds = \rho &\left(\frac{d}{dt}\int_{P(t)} \mathbf{f}_k\cdot\mathbf{a} d\mathbf{x} + \frac{d}{dt}\int_{\partial P(t)} \phi\mathbf{f}_k\cdot\mathbf{n} ds \right.\\ &\left. - \int_{P(t)} \frac{\partial \mathbf{f}_k}{\partial t}\cdot\mathbf{u} d\mathbf{x} - \int_{P(t)} \mathbf{u}\cdot((\mathbf{u}\cdot\nabla)\mathbf{f}_k) d\mathbf{x} \right)\\ &+ \eta \left(\int_{\partial P(t)} \left(\mathbf{n}\cdot((\mathbf{u}\cdot\nabla)\mathbf{f}_k) + \mathbf{u}\cdot\frac{\partial}{\partial\mathbf{n}}\mathbf{f}_k \right) ds - \int_{P(t)} \mathbf{u}\cdot\Delta\mathbf{f}_k d\mathbf{x} \right)\\ &- \alpha \int_{\partial P(t)} H\mathbf{f}_k\cdot\mathbf{n} ds. \end{aligned} \tag{4.18}$$

The field $\mathbf{f}_k$ *is defined by* (3.18) *and is treated in* (4.18) *as a function of* $\mathbf{x}$ *and* t.

Proof. First, we rewrite the Eqs. (3.14)–(3.16) by using (4.10)

$$\int_{\partial P(t)} \boldsymbol{\sigma}_g \mathbf{n} ds = \int_{\partial P(t)} (\boldsymbol{\lambda} + \alpha H \mathbf{n}) ds, \tag{4.19}$$

$$\int_{\partial P(t)} \mathbf{r} \wedge \boldsymbol{\sigma}_g \mathbf{n} ds = \int_{\partial P(t)} \mathbf{r} \wedge (\boldsymbol{\lambda} + \alpha H \mathbf{n}) ds, \tag{4.20}$$

$$\int_{\partial P(t)} \mathbf{f}_k \cdot \boldsymbol{\sigma}_g \mathbf{n} ds = \int_{\partial P(t)} \mathbf{f}_k \cdot (\boldsymbol{\lambda} + \alpha H \mathbf{n}) ds. \tag{4.21}$$

Then the terms containing $\boldsymbol{\sigma}_g$ in (4.19)–(4.21) can be rewritten in a form containing the integrals of $\mathbf{a}$, ϕ and $\mathbf{u}$ only, but not their derivatives. To this end, we use the momentum equation

$$\rho \left(\frac{D\mathbf{u}}{Dt} - \mathbf{g} \right) = \nabla \cdot \boldsymbol{\sigma}_g \tag{4.22}$$

in the fictitious domain $P(t)$ and apply the divergence and transport theorems. Thus, (4.19) yields with (4.11)

$$\begin{aligned}
\int_{\partial P(t)} \boldsymbol{\sigma}_g \mathbf{n} ds &= \int_{P(t)} \nabla \cdot \boldsymbol{\sigma}_g d\mathbf{x} \\
&= \rho \int_{P(t)} \left(\frac{D\mathbf{u}}{Dt} - \mathbf{g} \right) d\mathbf{x} \\
&= \rho \left(\frac{d}{dt} \int_{P(t)} \mathbf{u} d\mathbf{x} - \mathbf{g} |P(t)| \right) \\
&= \rho \left(\frac{d}{dt} \int_{P(t)} \mathbf{a} d\mathbf{x} + \frac{d}{dt} \int_{\partial P(t)} \phi \mathbf{n} ds - \mathbf{g} |P(t)| \right)
\end{aligned}$$

so that (4.19) can be replaced by (4.16).

Analogously for the angular momentum equation (4.20) we have

$$\begin{aligned}
\int_{\partial P(t)} \mathbf{r} \wedge \boldsymbol{\sigma}_g \mathbf{n} ds &= \int_{P(t)} \mathbf{r} \wedge (\nabla \cdot \boldsymbol{\sigma}_g) d\mathbf{x} = \rho \int_{P(t)} \mathbf{r} \wedge \left(\frac{D\mathbf{u}}{Dt} - \mathbf{g} \right) d\mathbf{x} \\
&= \rho \left(\frac{d}{dt} \int_{P(t)} \mathbf{r} \wedge \mathbf{u} d\mathbf{x} - \int_{P(t)} \frac{D\mathbf{r}}{Dt} \wedge \mathbf{u} d\mathbf{x} \right)
\end{aligned}$$

since $\int_{P(t)} \mathbf{r} d\mathbf{x} = \mathbf{0}$. We now note that $\frac{D\mathbf{r}}{Dt} = \frac{D}{Dt}(\mathbf{x} - \mathbf{G}) = \mathbf{u} - \mathbf{V}$ and substitute (4.11) for $\mathbf{u}$ to conclude

$$\begin{aligned}
\int_{\partial P(t)} \mathbf{r} \wedge \boldsymbol{\sigma}_g \mathbf{n} ds &= \rho \left(\frac{d}{dt} \int_{P(t)} \mathbf{r} \wedge (\mathbf{a} + \nabla \phi) d\mathbf{x} + \int_{P(t)} \mathbf{V} \wedge \mathbf{u} d\mathbf{x} \right) \\
&= \rho \left(\frac{d}{dt} \int_{P(t)} \mathbf{r} \wedge \mathbf{a} d\mathbf{x} + \frac{d}{dt} \int_{\partial P(t)} \phi \mathbf{r} \wedge \mathbf{n} ds + \mathbf{V} \wedge \int_{P(t)} \mathbf{u} d\mathbf{x} \right)
\end{aligned}$$

so that (4.20) can be replaced by (4.17).

Finally, the left-hand side in the remaining force Eq. (4.21) is rewritten as

$$\int_{\partial P(t)} \mathbf{f}_k \cdot \boldsymbol{\sigma}_g \mathbf{n} ds = \int_{P(t)} \mathbf{f}_k \cdot (\nabla \cdot \boldsymbol{\sigma}_g) d\mathbf{x} + \int_{P(t)} \boldsymbol{\sigma}_g : \nabla \mathbf{f}_k d\mathbf{x}. \tag{4.23}$$

We now treat the two terms in the last relation separately. Note first that

$$\nabla \cdot \mathbf{f}_k = 0 \tag{4.24}$$

by comparing (3.5) and (3.18). For the first term in (4.23) we get now with (4.22), (3.2) and (4.24)

$$\begin{aligned}
\int_{P(t)} \mathbf{f}_k \cdot (\nabla \cdot \boldsymbol{\sigma}_g) d\mathbf{x} &= \rho \int_{P(t)} \mathbf{f}_k \cdot \left(\frac{D\mathbf{u}}{Dt} - \mathbf{g} \right) d\mathbf{x} \\
&= \rho \left(\frac{d}{dt} \int_{P(t)} \mathbf{f}_k \cdot \mathbf{u} d\mathbf{x} - \int_{P(t)} \frac{D\mathbf{f}_k}{Dt} \cdot \mathbf{u} d\mathbf{x} \right) \\
&= \rho \left(\frac{d}{dt} \int_{P(t)} \mathbf{f}_k \cdot \mathbf{a} d\mathbf{x} + \frac{d}{dt} \int_{\partial P(t)} \phi \mathbf{f}_k \cdot \mathbf{n} ds \right. \\
&\qquad \left. - \int_{P(t)} \frac{\partial \mathbf{f}_k}{\partial t} \cdot \mathbf{u} d\mathbf{x} - \int_{P(t)} \mathbf{u} \cdot ((\mathbf{u} \cdot \nabla) \mathbf{f}_k) d\mathbf{x} \right).
\end{aligned} \tag{4.25}$$

For the second term in (4.23) we get

$$\int_{P(t)} \boldsymbol{\sigma}_g : \nabla \mathbf{f}_k d\mathbf{x} = \eta \sum_{j=1}^{3} \int_{P(t)} (\nabla \mathbf{u} + \nabla \mathbf{u}^T) \boldsymbol{\epsilon}_j \cdot \nabla \frac{\partial f^j}{\partial q_k} d\mathbf{x} \tag{4.26}$$

since, with (4.24)

$$\int_{P(t)} p\mathbf{I} : \nabla \mathbf{f}_k d\mathbf{x} = \int_{P(t)} p \sum_{j=1}^{3} \boldsymbol{\epsilon}_j \cdot \nabla \frac{\partial f^j}{\partial q_k} d\mathbf{x} = \int_{P(t)} p \nabla \cdot \frac{\partial \mathbf{x}}{\partial q_k} d\mathbf{x} = 0.$$

We now denote the Cartesian coordinates in the frame $(G, \boldsymbol{\epsilon}_j)$ by ξ_j and the components of the vectors $\mathbf{u}$ and $\mathbf{n}$ in the same frame by $\widetilde{u}_j$ and $\widetilde{n}_j$. We then write the right-hand side of (4.26) in an explicit form and integrate it by parts using (4.24) for a simplification. This yields with (3.18)

$$\begin{aligned}
&\sum_{j=1}^{3} \int_{P(t)} (\nabla \mathbf{u} + \nabla \mathbf{u}^T) \boldsymbol{\epsilon}_j \cdot \nabla \frac{\partial f^j}{\partial q_k} d\mathbf{x} \\
&\quad = \sum_{i,j=1}^{3} \int_{P(t)} \left(\frac{\partial \widetilde{u}_i}{\partial \xi_j} + \frac{\partial \widetilde{u}_j}{\partial \xi_i} \right) \frac{\partial}{\partial \xi_i} \frac{\partial f^j}{\partial q_k} d\mathbf{x} \\
&\quad = \sum_{i,j=1}^{3} \int_{\partial P(t)} (\widetilde{u}_i \widetilde{n}_j + \widetilde{u}_j \widetilde{n}_i) \frac{\partial}{\partial \xi_i} \frac{\partial f^j}{\partial q_k} ds
\end{aligned}$$

$$
\begin{aligned}
&- \sum_{i,j=1}^{3} \int_{P(t)} \widetilde{u}_j \frac{\partial^2}{\partial \xi_i^2} \frac{\partial f^j}{\partial q_k} d\mathbf{x} \\
&= \int_{\partial P(t)} \left(\mathbf{n} \cdot ((\mathbf{u} \cdot \nabla)\mathbf{f}_k) + \mathbf{u} \cdot \frac{\partial}{\partial \mathbf{n}} \mathbf{f}_k \right) ds \\
&- \sum_{j=1}^{3} \int_{P(t)} \mathbf{u} \cdot \Delta \mathbf{f}_k d\mathbf{x}. \qquad (4.27)
\end{aligned}
$$

Putting (4.23), (4.25) and (4.27) together we see that (4.21) can be replaced by (4.18). □

4.4. *Time marching scheme*

Let us now present our time marching scheme. Assuming that all the unknowns $\mathbf{a}^{n-1}$, ϕ^{n-1}, $\mathbf{G}^{n-1}$, $\boldsymbol{\epsilon}_j^{n-1}$, q_k^{n-1}, $\mathbf{V}^{n-1}$, ω^{n-1}, $\dot{q}_k^{n-1}$ ($j = 1, 2, 3$ and $k = 1, \ldots, n$) are already computed at time $t_{n-1} = (n-1)\Delta t$ we proceed with their calculation at time $t_n = n\Delta t$ through the four following stages:

(1) Propagation of bubbles

$$
\begin{aligned}
\mathbf{G}^n &= \mathbf{G}^{n-1} + \mathbf{V}^{n-1}\Delta t \\
\boldsymbol{\epsilon}_j^n &= \boldsymbol{\epsilon}_j^{n-1} + \boldsymbol{\omega}^{n-1} \wedge \boldsymbol{\epsilon}_j^{n-1}\Delta t, \quad j = 1, 2, 3 \\
q_k^n &= q_k^{n-1} + \dot{q}_k^{n-1}\Delta t, \quad k = 1, \ldots, n.
\end{aligned}
$$

The bubble domain P^n at time t_n has now been determined.

(2) Updating the variable ϕ

$$
\int_\Omega \nabla \phi^n \cdot \nabla q = - \int_\Omega \mathbf{a}^{n-1} \cdot \nabla q, \quad \forall\, q \in L_0^2(\Omega). \qquad (4.28)
$$

(3) Updating the vector field $\mathbf{a}^n$, the Lagrange multiplier $\boldsymbol{\lambda}^n$ and the velocity parameters of the bubbles $\mathbf{V}^n$, $\boldsymbol{\omega}^n$ and $\dot{q}_k^n$:

$$
\begin{aligned}
&\frac{\rho}{\Delta t} \int_\Omega \mathbf{a}^n \cdot \mathbf{v} d\mathbf{x} + \eta \int_\Omega \nabla \mathbf{a}^n : \nabla \mathbf{v} d\mathbf{x} - \int_{\partial P^n} \boldsymbol{\lambda}^n \cdot \mathbf{v} ds \\
&\quad = \frac{\rho}{\Delta t} \int_\Omega \mathbf{a}^{n-1} \cdot \ \mathbf{v} d\mathbf{x} - \rho \int_\Omega (\mathbf{u}^{n-1} \cdot \nabla)\mathbf{u}^{n-1} \cdot \mathbf{v} d\mathbf{x}, \quad \forall\, \mathbf{v} \in H_0^1(\Omega), \qquad (4.29)
\end{aligned}
$$

$$
\mathbf{a}^n \cdot \boldsymbol{\tau}|_\Gamma = - \left. \frac{\partial \phi^n}{\partial \boldsymbol{\tau}} \right|_\Gamma, \quad \mathbf{a}^n \cdot \mathbf{n}|_\Gamma = 0,
$$

$$
\int_{\partial P^n} \mathbf{a}^n \cdot \boldsymbol{\mu} ds = \int_{\partial P^n} \left(-\nabla \phi^n + \mathbf{V}^n + \boldsymbol{\omega}^n \wedge \mathbf{r} + \sum_{k=1}^{n} \mathbf{f}_k \dot{q}_k^n \right) \cdot \boldsymbol{\mu} ds, \quad \forall\, \boldsymbol{\mu} \in H^{-\frac{1}{2}}(\partial P^n)
$$

and force Eqs (4.16)–(4.18).

(4) Calculating the velocity

$$\mathbf{u}^n = \mathbf{a}^n + \nabla\phi^n.$$

For the discretization in space, both $\mathbf{a}^n$ and ϕ^n are approximated by bilinear finite elements on a Cartesian grid of size $N_x \times N_y$ on Ω. This allows us to use the Fast Fourier Transform to solve the linear systems resulting from Eq. (4.28) for ϕ^n and Eq. (4.29) for $\mathbf{a}^n$ (with a given $\boldsymbol{\lambda}^n$). The boundary ∂P^n is represented by N_λ segments and the Lagrange multiplier $\boldsymbol{\lambda}^n$ is approximated by linear finite elements on these segments. The linear system for $\boldsymbol{\lambda}^n$, $\mathbf{V}^n$, ω^n and $\dot{q}_k^n$ resulting from the equations in the stage 3 of the above algorithm is solved by the GMRES method. Note that the linear system (4.29) for $\mathbf{a}^n$ should be solved on each iteration of GMRES, which can make the overall resolution rather expensive. To rectify this problem we have used in situations with an important number of bubbles a Schwartz-type domain decomposition method. Namely, rather than solving the equations of the stage 3 directly, we can find iteratively the approximations for the Lagrange multiplier $\boldsymbol{\lambda}^n$ and velocity parameters $\mathbf{V}^n$, ω^n and $\dot{q}_k^n$ separately on each bubble restricting the domain to a parallelipiped (or a rectangle in 2D) containing this bubble. All the Lagrange multipliers and velocity parameters are then matched in the iterative process, in which the linear system (4.29) for $\mathbf{a}^n$ is solved on the whole domain Ω on each iteration.

5. The Coalescence Process

Coalescence is an instantaneous transition in which two bubbles, or more, in contact along a common boundary transform themselves into a single one. We choose to model this transition in considering that it takes place during a finite but small interval of time, which is suited for a numerical approach. More precisely we assume that

- When two bubbles are "sufficiently close" the coalescence process is starting.
- In a short time (numerically a small number of time steps) the small bubble empty itself into the large one.

Let V and respectively W be the volume of the large and respectively small bubble before coalescence. When they have the same size, one of the bubble, arbitrarily chosen, plays the role of the large one.

During the coalescence process their volumes are represented by

$$\tau V \quad \text{and} \quad \mu W,$$

where the parameters τ and μ are related by the following mass conservation equation:

$$\tau V + \mu W = V + W.$$

In this relation μ varies between 0 and 1 whereas τ varies between 1 and $1 + W/V$. In other words,

$$\mu = 1 + (1 - \tau)V/W$$

and, in a few time steps, the volume of the larger bubble passes from V to $V + W$ whereas the smallest one disappears.

In order to obtain a criteria telling us when two bubbles are considered as sufficiently close, we derive a method allowing to compute the distance separating them.

5.1. *Distance between two bubbles*

Let P_1 and P_2 be two domains representing the bubbles and $d(P_1, P_2)$ be the distance between them. Our algorithm is an iterative process. More precisely, we will generate two sequences of points $\mathbf{x}_k$ and $\mathbf{y}_k$ on the boundaries of P_1 and P_2, respectively, which become closer and closer to each other so that $|\mathbf{x}_k - \mathbf{y}_k|$ is expected to tend to $d(P_1, P_2)$ as $k \to \infty$.

Let us choose some points $\mathbf{G}_1 \in \mathrm{Int}(P_1)$ and $\mathbf{G}_2 \in \mathrm{Int}(P_2)$. We will assume that for any given points $\mathbf{z} \neq \mathbf{G}_1$ and $\mathbf{y}$ we are able to find the intersection of ∂P_1 with the line passing through $\mathbf{G}_1$ and $\mathbf{z}$ and to choose the point out of this intersection that is the closest to $\mathbf{y}$. We will denote the point described above by $Pr_{\partial P_1}(\mathbf{z}|\mathbf{y})$. A similar construction gives us $Pr_{\partial P_2}(\mathbf{z}|\mathbf{y})$ for any $\mathbf{z}$, $\mathbf{y} \neq \mathbf{G}_2$.

We set $\mathbf{x}_0 = Pr_{\partial P_1}(\mathbf{G}_2, \mathbf{G}_2)$ and $\mathbf{y}_0 = Pr_{\partial P_2}(\mathbf{G}_1, \mathbf{G}_1)$ as the initial approximations. At the kth iteration, having two points $\mathbf{x}_k \in \partial P_1$ and $\mathbf{y}_k \in \partial P_2$, we try to find a new point $\mathbf{x}_{k+1} \in \partial P_1$, which is closer to $\mathbf{y}_k$ than $\mathbf{x}_k$. To this end, we construct first the plane L passing through $\mathbf{x}_k$ and tangential to ∂P_1. We denote then $\mathbf{z}_k^{(0)}$ the projection of $\mathbf{y}_k$ on the plane L and set $\mathbf{x}_k^{(0)} = Pr_{\partial P_1}(\mathbf{z}_k^{(0)}, \mathbf{y}_k)$. If the distance $|\mathbf{x}_k^{(0)} - \mathbf{y}_k|$ is less than $|\mathbf{x}_k - \mathbf{y}_k|$, we accept $\mathbf{x}_k^{(0)}$ as $\mathbf{x}_{k+1}$. Otherwise, we start another inner loop

$$\mathbf{z}_k^{(l)} = \frac{\mathbf{x}_k + \mathbf{z}_k^{(l-1)}}{2}, \quad \mathbf{x}_k^{(l)} = Pr_{\partial P_1}(\mathbf{z}_k^{(0)}, \mathbf{y}_k), \quad l = 1, 2, \ldots$$

until $|\mathbf{x}_k^{(l)} - \mathbf{y}_k|$ is less than $|\mathbf{x}_k - \mathbf{y}_k|$ and then accept $\mathbf{x}_k^{(l)}$ as $\mathbf{x}_{k+1}$. To construct $\mathbf{y}_{k+1}$ we proceed in the similar way, trying to find a point $\mathbf{y}_{k+1} \in \partial P_2$, which is closer to $\mathbf{x}_k$ than $\mathbf{y}_k$.

6. Numerical Results

6.1. *The two-dimensional problem for rigid ellipses*

Let us present some results in 2D. The modeling and the numerical approach remain the same as in the 3D case. The only difference is that the barycenter is now given by a 2-vector $\mathbf{G} = (G_1, G_2)$ and the rotation is described by a single scalar parameter, which can be taken as the angle θ between the axes $\mathbf{e}_1$ and $\boldsymbol{\epsilon}_1$. The angular velocity

is reduced to the scalar $\omega(t) = \dot{\theta}(t)$ as well as the corresponding force Eq. (4.17). We take the following physical parameters: the density of the fluid $\rho = 1000\, kg/m^3$, the viscosity $\eta = 0.01\,\mathrm{Pa}\cdot\mathrm{s}$, the gravitation force $\mathbf{g} = -10\mathbf{e}_y(m/s^2)$. The time step will be fixed to $\Delta t = 10^{-3}\,\mathrm{s}$ in all our simulations.

Our first example is the motion of a massless rigid ellipse in a fluid. Hence, the system of force equations reduces to (4.16)–(4.17).

We begin with the results of simulations of a single rigid circular bubble of radius $R = 1\,\mathrm{mm}$ rising in a $2\,\mathrm{cm} \times 2\,\mathrm{cm}$ container. We make use of a Cartesian grid of $N \times N$ points to represent $\mathbf{a}$ and ϕ and a regular grid of N_λ points on $\partial P(t)$ to represent $\boldsymbol{\lambda}$. Figure 2 shows the initial configuration of the bubble, its configuration and the fluid velocity at $t = 0.1\,\mathrm{s}$ and a zoom of the bubble and the fixed grid for $N = 128$.

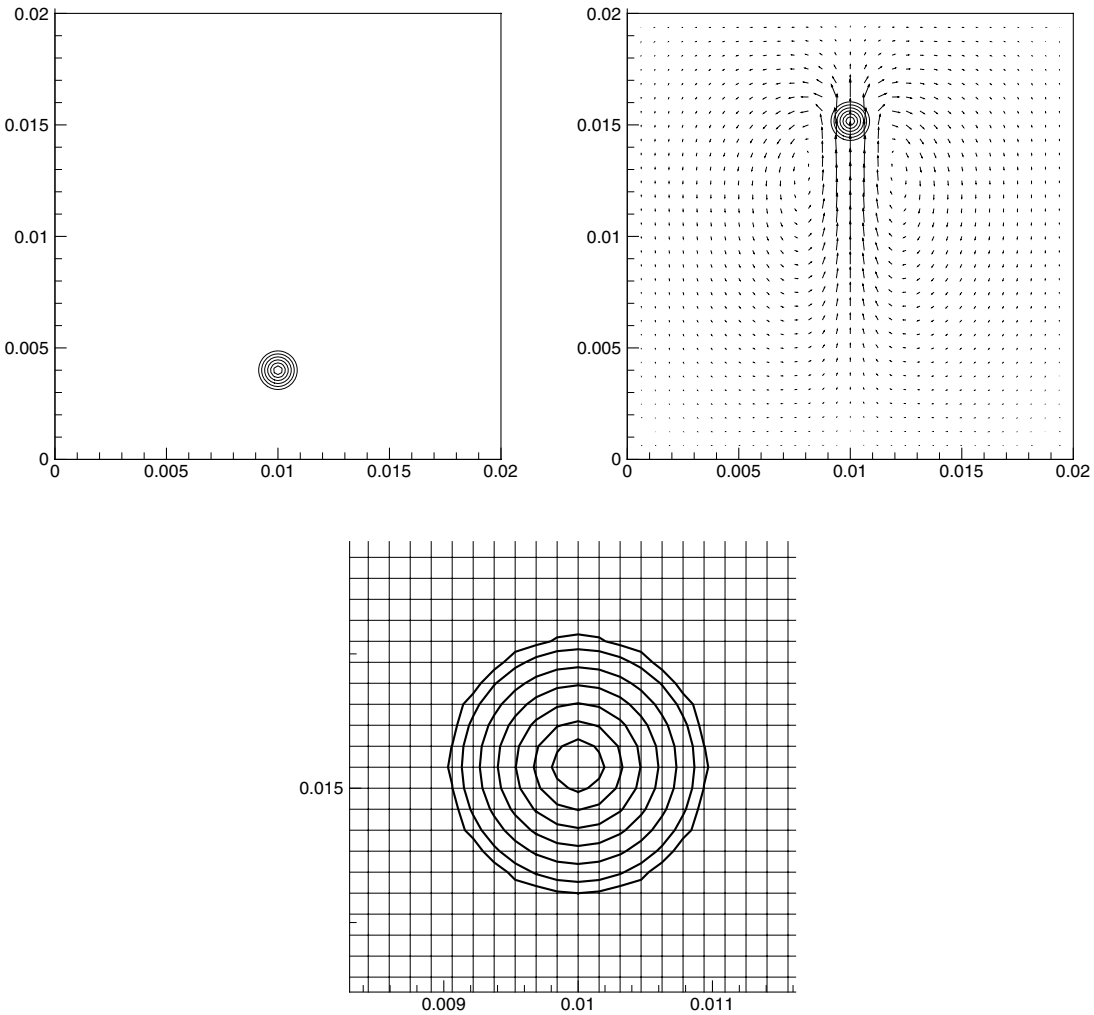


Fig. 2. One circular rigid body. Initial configuration, configuration and the fluid velocity at $t = 0.1s$ and a zoom with the grid.

We now study the influence of the discretization parameters, i.e. N and N_λ. We choose N_λ as a function of N for $N_\lambda = [k_\lambda R/h]$ where k_λ is some constant, h is the grid step corresponding to N and the brackets denote the integer part. Figures 3 and 4 present the evolution of the y-coordinate of the bubble barycenter with time. On Fig. 3, we fix $N = 128$ and vary k_λ. The ratio of the size h of the fixed grid and the size h_λ of the moving grid is also indicated. We see that the convergence (in an eyeball norm) is achieved for $k_\lambda = 4$, i.e. when h_λ is about 1.5 times greater than h. This rule for choosing N_λ will be adopted in all our subsequent simulations. On Fig. 4, we fix $k_\lambda = 4$ and vary N. The convergence can also be observed when N gets sufficiently high.

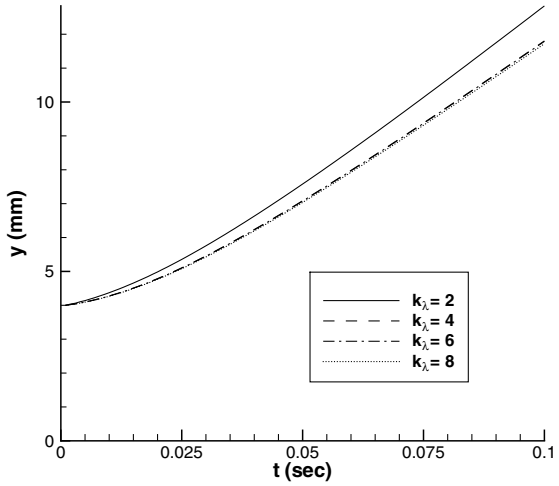


Fig. 3. One circular rigid body rising in a container; dependence on the size of the grid on $\partial P(t)$: solid line – $k_\lambda = 2$, $h_\lambda/h = 3.06$; dashed line – $k_\lambda = 4$, $h_\lambda/h = 1.54$; dash-dotted line – $k_\lambda = 6$, $h_\lambda/h = 1.02$; dotted line – $k_\lambda = 8$, $h_\lambda = 0.77/h$. The last three lines are practically superimposed.

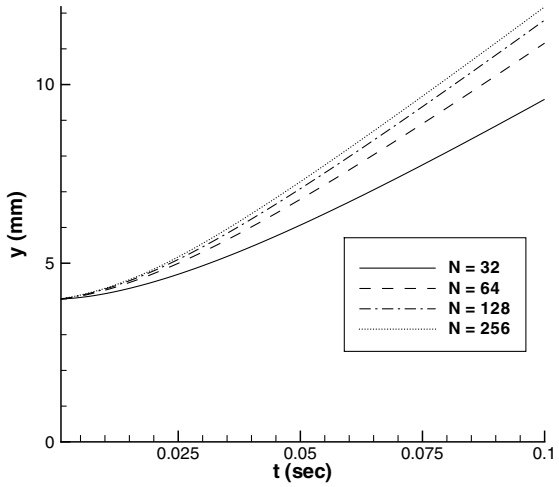


Fig. 4. One circular rigid body rising in a container; dependence on the size of the grid on Ω.

To conclude this section, we present two more complicated simulations. The first one deals with three circular bubbles of radii from 0.8 to 1.5 mm evolving in a 1 cm × 3 cm container. The Cartesian grid of 64×200 points was used to represent $\mathbf{a}$ and ϕ. The results are shown in Fig. 5 as snapshots of bubbles at some time moments increasing from left to right and from top to bottom.

Our last example in this section involves a much greater amount of bubbles and illustrates our treatment of the coalescence process. We start with a random configuration of 200 elliptic bubbles with sizes from 0.5 to 3 mm placed in a 5 cm × 10 cm container. The results are shown in Fig. 6 as snapshots of bubbles at increasing time moments. The bubbles rise so that the bigger ones swallow or sweap away the smaller ones. The Cartesian Grid consisted of 256×512 points. The domain decomposition technique described at the end of Sec. 4 was used in this case. The simulation took 501 seconds of CPU time for 300 time steps on a Pentium M760 laptop computer.

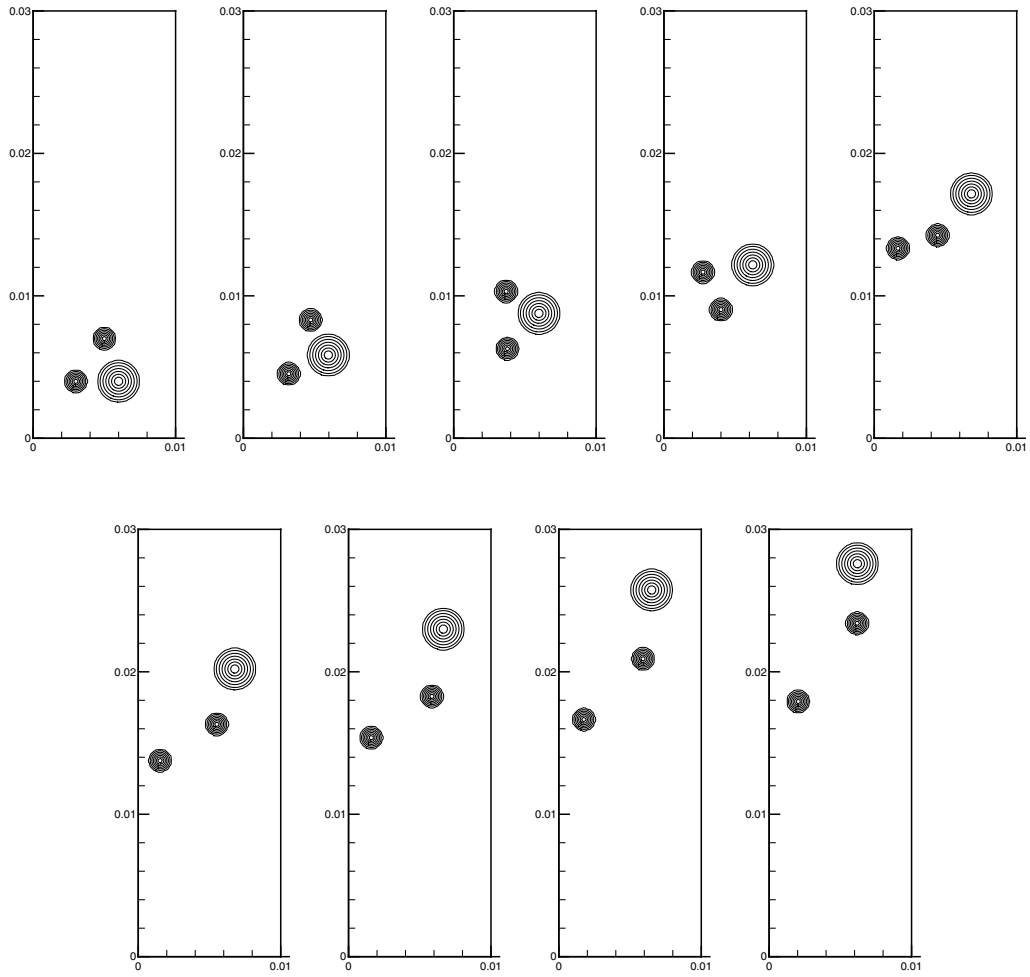


Fig. 5. Three circular rigid bodies. Snapshots at times $\{0, 2, 4, 6, 8, 10, 12, 14, 16, 18\} \cdot 10^{-2}$ s.

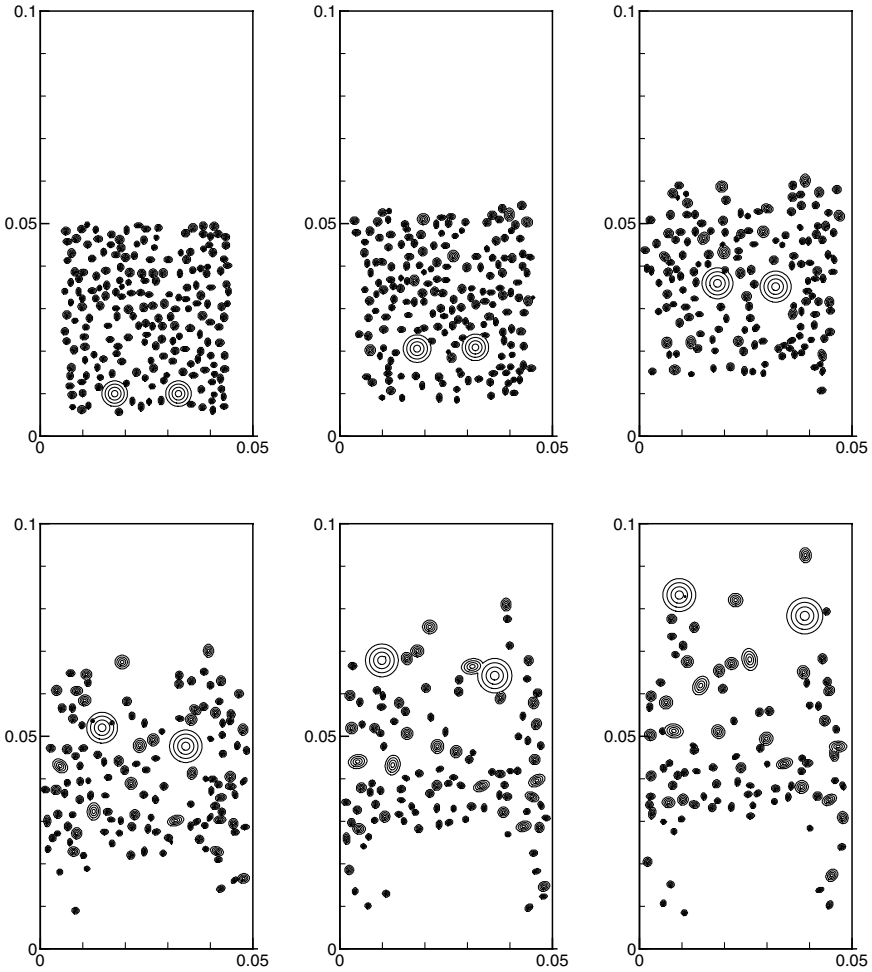


Fig. 6. Initial configuration of 200 elliptic bubbles and their evolution. Snapshots at time moments $\{0, 6, 12, 18, 24, 30\} \cdot 10^{-2}$ s.

6.2. *The two-dimensional problem for a family of deformable elliptic surfaces*

Let us now allow the ellipses from the preceding section to change their shapes in the course of time, but preserving their area. It means that we introduce a one parameter family of area preserving mappings $\mathbf{f}(\cdot, a) : \boldsymbol{\chi} \in X \to \mathbf{x} \in P_a$, X being a circle of area V centered at the origin and P_a an ellipse with axes $a\sqrt{V/\pi}$ and $1/a\sqrt{V/\pi}$:

$$x_1 = a\chi_1,$$
$$x_2 = \frac{1}{a}\chi_2.$$

Obviously, the force equations (4.16)–(4.17) remain the same as for the rigid ellipses. We should however add one more equation corresponding to (4.18) for $q_1 = a$. At

the point $\mathbf{f}(\chi, a) = \mathbf{x}$ we have clearly

$$\frac{\partial \mathbf{f}}{\partial a} = \begin{pmatrix} \chi_1 \\ -\dfrac{1}{a^2}\chi_2 \end{pmatrix} = \frac{1}{a}\begin{pmatrix} x_1 \\ -x_2 \end{pmatrix}.$$

Hence

$$\mathbf{f}_a = \sum_{j=1}^{2} \frac{\partial f^j}{\partial a}\boldsymbol{\epsilon}_j = \frac{1}{a}\mathbf{r}^*$$

with

$$\mathbf{r}^* = (\mathbf{x} - \mathbf{G}, \boldsymbol{\epsilon}_1)\boldsymbol{\epsilon}_1 - (\mathbf{x} - \mathbf{G}, \boldsymbol{\epsilon}_2)\boldsymbol{\epsilon}_2$$

Equation (3.7) for the velocity of the deformable body can be now explicitly rewritten as

$$\dot{\mathbf{x}} = \mathbf{V} + \omega \wedge \mathbf{r} + \frac{\dot{a}}{a}\mathbf{r}^* \tag{6.1}$$

and the force Eq. (4.18) reduces to

$$\begin{aligned} \int_{\partial P(t)} \mathbf{r}^* \cdot \boldsymbol{\lambda} ds = \rho \Bigg(& \frac{d}{dt}\int_{P(t)} \mathbf{r}^* \cdot \mathbf{a} d\mathbf{x} + \frac{d}{dt}\int_{\partial P(t)} \phi \mathbf{r}^* \cdot \mathbf{n} ds \\ & + \int_{P(t)} \left(\frac{\dot{a}}{a}\mathbf{r}^* + \mathbf{V}^* - 2\omega \wedge \mathbf{r}^* \right) \cdot \mathbf{u} d\mathbf{x} \\ & - \int_{P(t)} \left((\mathbf{u} \cdot \boldsymbol{\epsilon}_1)^2 - (\mathbf{u} \cdot \boldsymbol{\epsilon}_2)^2 \right) d\mathbf{x} \Bigg) \\ & + 4V\eta\frac{\dot{a}}{a} - \alpha \int_{\partial P(t)} H\mathbf{r}^* \cdot \mathbf{n} ds \end{aligned} \tag{6.2}$$

with

$$\mathbf{V}^* = (\mathbf{V}, \boldsymbol{\epsilon}_1)\boldsymbol{\epsilon}_1 - (\mathbf{V}, \boldsymbol{\epsilon}_2)\boldsymbol{\epsilon}_2.$$

We made use of the relations

$$\frac{\partial \mathbf{r}^*}{\partial t} = -\mathbf{V}^* + 2\omega \wedge \mathbf{r}^*,$$

$$\int_{P(t)} \mathbf{u} \cdot ((\mathbf{u} \cdot \nabla)\mathbf{r}^*) d\mathbf{x} = \int_{P(t)} \left((\mathbf{u} \cdot \boldsymbol{\epsilon}_1)^2 - (\mathbf{u} \cdot \boldsymbol{\epsilon}_2)^2 \right) d\mathbf{x}$$

and

$$\int_{\partial P(t)} \left(\mathbf{n} \cdot ((\mathbf{u} \cdot \nabla)\mathbf{r}^*) + \mathbf{u} \cdot \frac{\partial}{\partial \mathbf{n}}\mathbf{r}^* \right) ds = 4V\frac{\dot{a}}{a}.$$

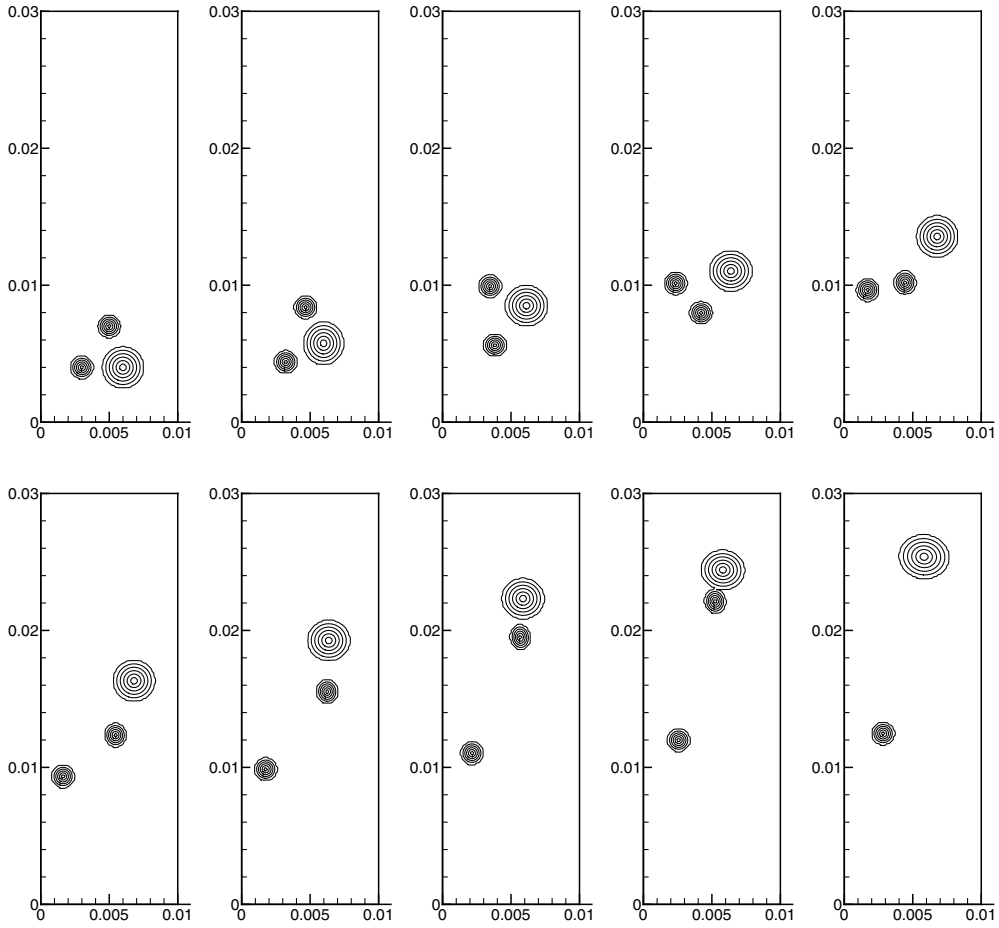


Fig. 7. Three elliptical deformable bubbles. Snapshots at times $\{0, 2, 4, 6, 8, 10, 12, 14, 16, 18\} \cdot 10^{-2}$ s.

The last computation can be easily checked by using the divergence theorem and by extending $\mathbf{u}|_{\partial P(t)}$ into $P(t)$ by $\mathbf{u} = \dot{\mathbf{x}}$ where $\dot{\mathbf{x}}$ is given by (6.1).

We now present some results for this family of surfaces. We begin with a simulation starting from the configuration of three circular bubbles used in Fig. 5. The results in the case of deformable elliptic bubbles are shown in Fig. 7. The surface tension coefficient was chosen as $\alpha = 0.05$. Comparing Figs. 5 and 7, we observe very similar bubble evolution, however a coalescence occurs in the latter case. A more detailed representation of a coalescence of two bubbles is given in Fig. 8, where two elliptic bubbles are initially chosen. The surface tension coefficient is again equal to 0.05.

Figure 9 demonstrates the influence of the surface tension. We follow here the evolution of one elliptic bubble under three different values of the surface tension coefficient α. The less α is, the flatter is the final form of the rising bubble.

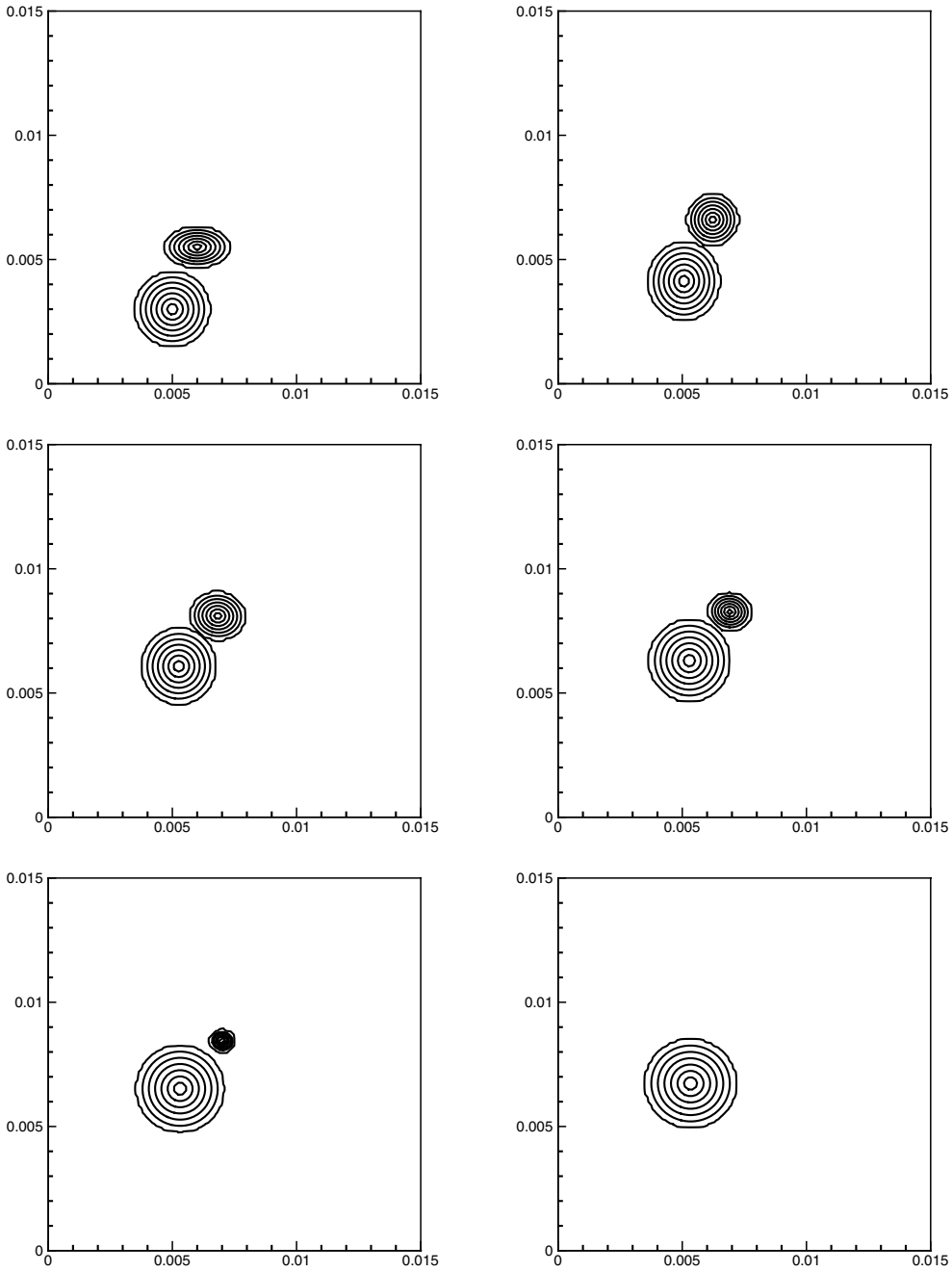


Fig. 8. Two elliptical deformable bubbles. Snapshots at times $\{0, 2, 3.8, 4, 4.2, 4.6\} \cdot 10^{-2}$ s.

6.3. *The two-dimensional problem for a family of surfaces given by polynomials of degree* 4

Let us now allow for one more degree of freedom, which we add to the ellipses used in the preceding section. Our family of mappings $\mathbf{f}$ now depends on two parameters $q_1 = a$ and $q_2 = c$, i.e. $\mathbf{f}(\cdot, a, c) : \boldsymbol{\chi} \in X \to \mathbf{x} \in P_{a,c}$, X being again a circle of

area V centered at the origin:

$$\begin{aligned} x_1 &= a\chi_1, \\ x_2 &= \frac{1}{a}\chi_2 - c\chi_1^2 + \frac{cV}{4\pi}. \end{aligned} \tag{6.3}$$

Obviously, the force equations (4.16)–(4.17) remain the same as for the rigid ellipses. It is easy to check that the force Eq. (6.2) corresponding to $q_1 = a$ remains also valid. We should add the force equation corresponding to $q_2 = c$. To do this, we compute

$$\frac{\partial \mathbf{f}}{\partial c} = \begin{pmatrix} 0 \\ -\chi_1^2 + \dfrac{V}{4\pi} \end{pmatrix} = \begin{pmatrix} 0 \\ -\dfrac{x_1^2}{a^2} + \dfrac{V}{4\pi} \end{pmatrix}.$$

Hence

$$\mathbf{f}_c = \sum_{j=1}^{2} \frac{\partial f^j}{\partial c}\boldsymbol{\epsilon}_j = \left(-\frac{1}{a^2}(\mathbf{r}, \boldsymbol{\epsilon}_1)^2 + \frac{V}{4\pi}\right)\boldsymbol{\epsilon}_2.$$

The equation for the velocity of the deformable body now reads

$$\dot{\mathbf{x}} = \mathbf{V} + \omega \wedge \mathbf{r} + \frac{\dot{a}}{a}\mathbf{r}^* + \dot{c}\left(-\frac{1}{a^2}(\mathbf{r}\cdot\boldsymbol{\epsilon}_1)^2 + \frac{V}{4\pi}\right)\boldsymbol{\epsilon}_2 \tag{6.4}$$

and the force Eq. (4.18) reduces to

$$\begin{aligned} \int_{\partial P(t)} \mathbf{f}_c \cdot \boldsymbol{\lambda} ds = \rho \Bigg(& \frac{d}{dt}\int_{P(t)} \mathbf{f}_c \cdot \mathbf{a} d\mathbf{x} + \frac{d}{dt}\int_{\partial P(t)} \phi \mathbf{f}_c \cdot \mathbf{n} ds - \int_{P(t)} \frac{\partial \mathbf{f}_c}{\partial t} \cdot \mathbf{u} d\mathbf{x} \\ & + \frac{2}{a^2}\int_{P(t)} (\mathbf{u}\cdot\boldsymbol{\epsilon}_1)(\mathbf{u}\cdot\boldsymbol{\epsilon}_2)(\mathbf{r}\cdot\boldsymbol{\epsilon}_1) d\mathbf{x} \Bigg) \\ & + \frac{2\eta}{a^2}\int_{P(t)} \mathbf{u}\cdot\boldsymbol{\epsilon}_2 d\mathbf{x} - \frac{2V\eta}{a^2}\mathbf{V}\cdot\boldsymbol{\epsilon}_2 + \frac{V^2\eta}{\pi a^3}\dot{c} - \alpha\int_{\partial P(t)} H\mathbf{f}_c \cdot \mathbf{n} ds, \end{aligned}$$

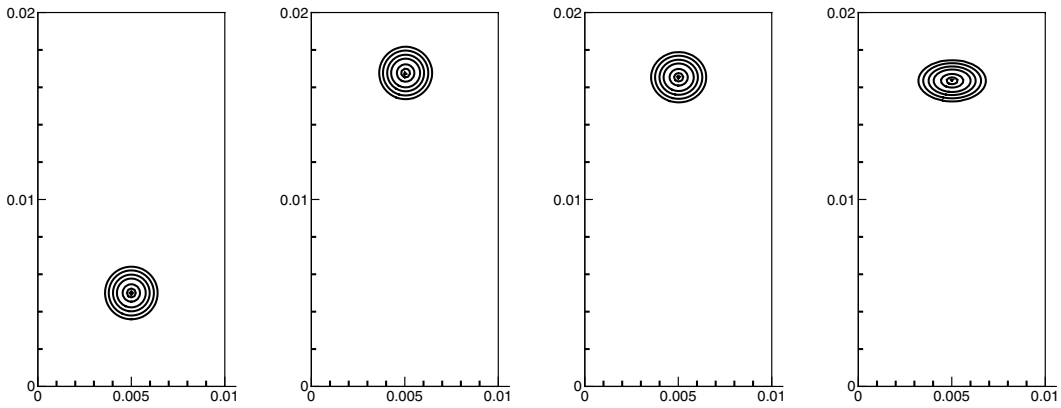


Fig. 9. One elliptical deformable bubble: influence of the surface tension. Initial configuration and final configurations at $t = 0.1\,\mathrm{s}$ and for $\alpha = 0.5$, 0.05 and 0.0005.

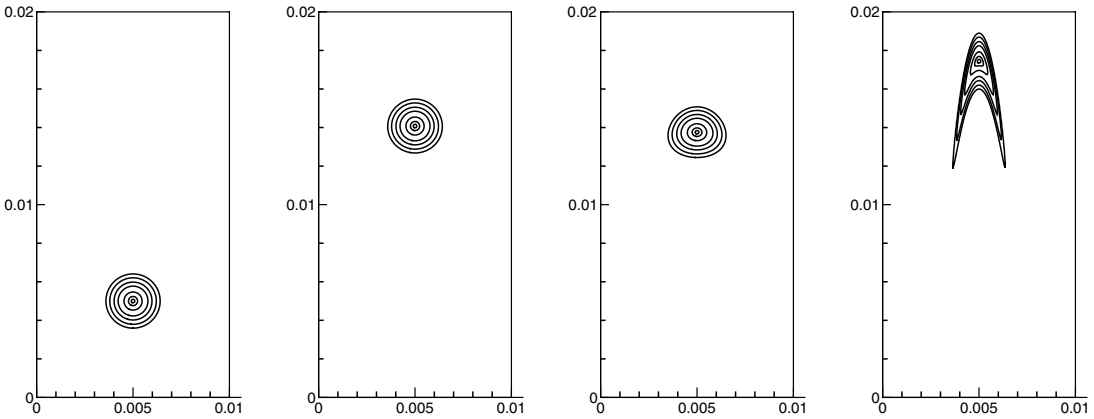


Fig. 10. A bubble described by mapping (6.3). Initial configuration and final configurations at $t = 0.10\,\mathrm{s}$ and for $\alpha = 0.5$, 0.05 and 0.0005.

where we should substitute

$$\frac{\partial \mathbf{f}_c}{\partial t} = \omega \left(\frac{1}{a^2}(\mathbf{r}, \boldsymbol{\epsilon}_1)^2 - \frac{V}{4\pi} \right) \boldsymbol{\epsilon}_1 + \frac{2(\mathbf{r}, \boldsymbol{\epsilon}_1)}{a^2} \left(\frac{\dot{a}}{a}(\mathbf{r}, \boldsymbol{\epsilon}_1) + (\mathbf{V}, \boldsymbol{\epsilon}_1) - \omega(\mathbf{r}, \boldsymbol{\epsilon}_2) \right) \boldsymbol{\epsilon}_2.$$

We made use of the relations

$$\int_{P(t)} \mathbf{u} \cdot ((\mathbf{u} \cdot \nabla)\mathbf{f}_c) d\mathbf{x} = -\frac{2}{a^2} \int_{P(t)} (\mathbf{u} \cdot \boldsymbol{\epsilon}_1)(\mathbf{u} \cdot \boldsymbol{\epsilon}_2)(\mathbf{r} \cdot \boldsymbol{\epsilon}_1) d\mathbf{x}$$

and

$$\int_{\partial P(t)} \left(\mathbf{n} \cdot ((\mathbf{u} \cdot \nabla)\mathbf{f}_c) + \mathbf{u} \cdot \frac{\partial}{\partial \mathbf{n}} \mathbf{f}_c \right) ds = -\frac{2V}{a^2} \mathbf{V} \cdot \boldsymbol{\epsilon}_2 + \frac{V^2}{\pi a^3} \dot{c}.$$

Figure 10 shows some results obtained for the family of mappings described above. We follow here the evolution of a bubble under three different values of the surface tension coefficient α. Note that the elongated form of the bubble in the experiment with the smallest value of the surface tension $\alpha = 5 \cdot 10^{-4}$ allows it to rise faster (cf. Fig. 9 as well).

6.4. *The two-dimensional problem for rigid ellipsoidal surfaces*

For the three-dimensional setting, we restrict ourselves in this paper to the generalization of the rigid body family considered in Sec. 6.1, i.e. we represent the three-dimensional bubbles by rigid ellipsoids.

We begin with the results of the simulations of a single spherical bubble of radius $R = 1\,\mathrm{mm}$ rising in a $1\,\mathrm{cm} \times 1\,\mathrm{cm} \times 1\,\mathrm{cm}$ container. In order to be able to compare our results with Stokes law for the terminal settling velocity of a sphere in unbounded viscous liquid $V_{\mathrm{term}} = 2g\rho R^2/9\eta$ we reduced the fluid velocity to $\rho = 100\,\mathrm{kg/m^3}$ in this simulation. We have then $V_{\mathrm{term}} \approx 22\,\mathrm{mm/s}$ and the Reynolds number based on the sphere radius is 0.22 so that Stokes law should be applicable. The results are shown in Fig. 11. We present here two levels of discretization: $N = 32$ and

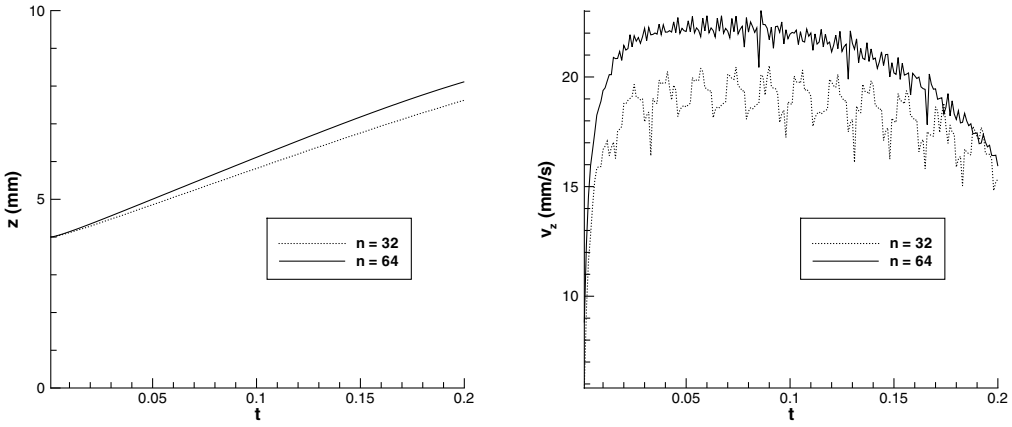


Fig. 11. Rise of a spherical rigid bubble in 3D. Evolution in time of the z-coordinate of the barycenter and z-component of its velocity.

$N = 64$, where N stands for the $N \times N \times N$ Cartesian grid in Ω. The moving grid on the sphere with 66 vertices was used in both cases. We observe that the settling value of the rising velocity becomes closer to that predicted by Stokes law V_{term} as the mesh is refined. In both simulations, after reaching the settling velocity, the bubble is impeded when it comes too close to the upper wall of the container. We note as well the presence of high oscillations in time in the velocity. This is clearly a numerical artifact which becomes less pronounced when the grids are refined. It can be mainly attributed to the imperfections of the numerical interpolation between the moving grid and the fixed one.

The snapshots of another simulation in 3D are shown in Fig. 12. We present here two ellipsoidal bubbles with axes ranging from 1 to 2 mm rising in a 1 cm $\times$ 1 cm $\times$ 1.5cm container, the fluid density being reset to $\rho = 1000\,\text{kg/m}^3$. The Cartesian grid of $64 \times 64 \times 64$ points was used in this simulation and it took 420 seconds of CPU time for 90 time steps.

7. Conclusion

The numerical results exhibited in the last section show that the model which has been derived in this paper yields on the one hand a description of the motion of a large number of bubbles in an incompressible fluid under the action of gravitation and on the other hand the velocity field within the fluid. It moreover accounts for possible coalescence between bubbles which are sufficiently close. The calculations are relatively fast: in 2D setting one can simulate up to about 2000 bubbles on a laptop computer taking about 20 seconds per time step. In 3D, we have experimented with 12 bubbles which takes about 40 secs per times step. One way to decrease the cost could be the use of domain decomposition techniques such as the Fat Boundary Method Ref. 2 for the case of rigid bodies moving in a fluid. In this method, a moving mesh is introduced not only on the bodies surfaces but also in

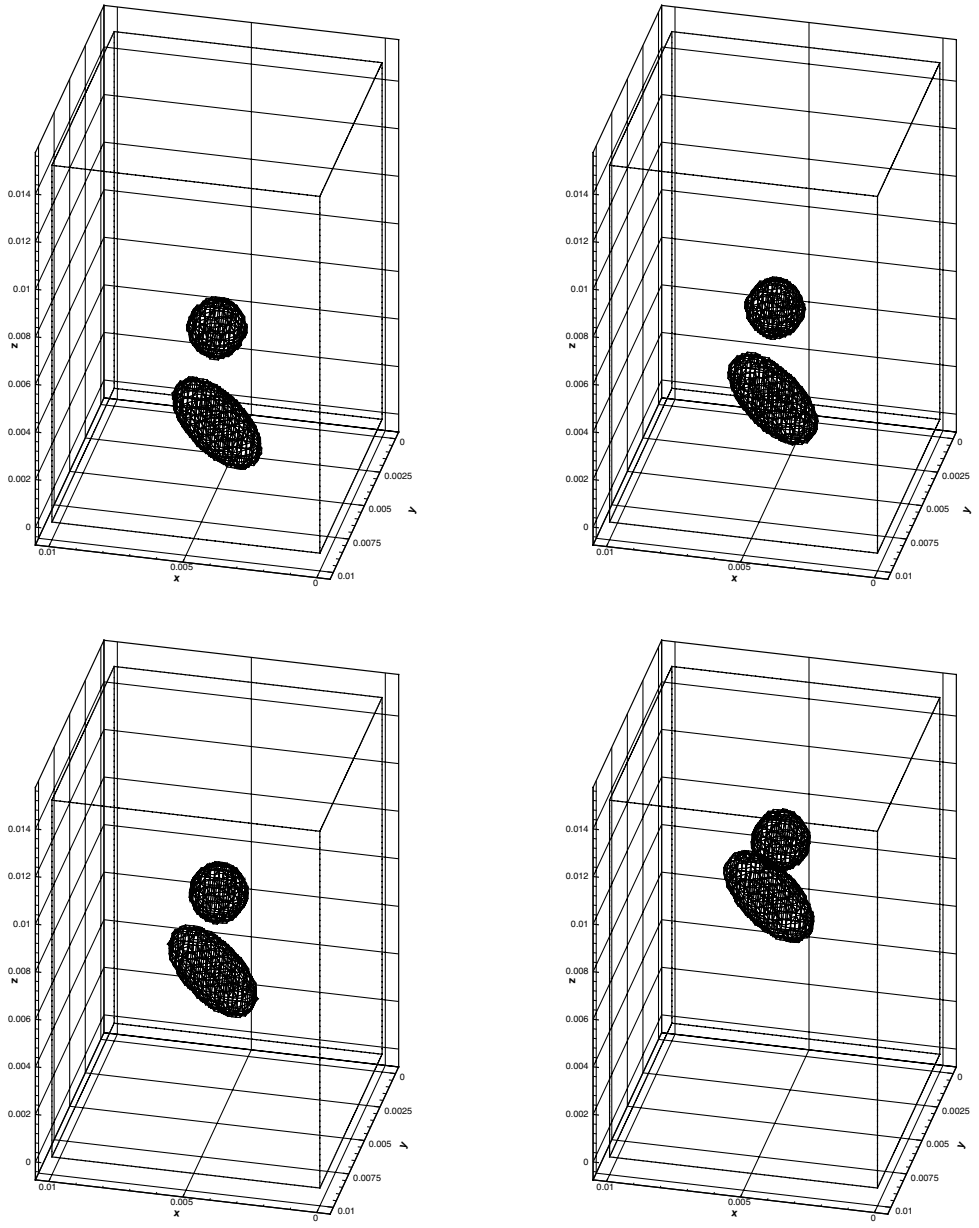


Fig. 12. Two rigid bubbles in 3D. Snapshots at times $\{0, 3, 6, 9\} \cdot 10^{-2}$ s.

some layers surrounding them. This permits to better represent the velocity around the bodies and thus to coarsen the fixed mesh on Ω.

The use of a predefined family of surfaces yields bubble shapes which are regular functions. It should be noted that with a suited family, i.e. with a sufficiently large number of parameters, it is in principle possible to reach a given degree of adequacy with respect to the exact shape.

Some difficulties still have to be overcome. The first one concerns the conditions imposed on the matching of the velocity field on the boundary between the fluid

and the bubble. This paper sticks to the dirty bubble case. For numerical reasons the clean bubble case seems more difficult to be handled than this first one.

The second problem is related to the simulation of the fluid layer on which the bubbles are freely moving after detachment. In our first attempt this physical phenomena has been described by the introduction of a short range repulsive force in the vicinity of the anode. This problem is also connected with the choice of the family of predefined surfaces used. This choice must in fact keep enough freedom to allow a displacement of the bubbles under the anode in which this last one essentially sticks to the layer.

As mentioned in the Introduction, our interest in the bubble motions arises from the fact that it enters the more general problem of the stability of the aluminum electrolysis cell, i.e. the stability of the motion of the electrolytic bath and of the aluminum already produced. The present work is thus intended to obtain a correct calculation of the velocity field which enters the equations used to derive a stability criterion, as in Ref. 4. Our approach consists in using 2D bubble simulations to calculate the velocity induced by the bubbles on the bath in the directions tangential to the anode inferior surface. This velocity is then imposed as boundary conditions to calculate the fluid mothion in the bath and melted aluminum.

Acknowledgments

This research was performed in the framework of scientific collaboration between the Chair of Analysis and Numerical Simulations, IACS-EPFL, and the company ALCAN. The authors thank Jacques Antille, Michel Flueck, Olivier Martin, Marco Picasso and Jacques Rappaz for numerous helpful discussions, insightful remarks and interest in their work. The first author was supported by CTI Project 6437.1 IWS-IW. The second author would like to acknowledge very fruitful discussions on coalescence and related problems with J. Hammerberg during his visit to the Los Alamos National Laboratory.

References

1. V. I. Arnold and B. A. Khesin, *Topological Methods in Hydrodynamics*, Applied Mathematical Sciences, Vol. 125 (Springer-Verlag, 1998).
2. S. Bertoluzza, M. Ismail and B. Maury, The fat boundary method: semi-discrete scheme and some numerical experiments, in *Domain Decomposition Methods in Science and Engineering*, Lect. Notes Comput. Sci. Eng., Vol. 40 (Springer, 2005), pp. 513–520.
3. A. Caboussat, Numerical simulation of two-phase free surface flows, *Arch. Comput. Meth. Eng.* **12** (2005) 165–210.
4. J. Descloux, M. Flueck and M. V. Romerio, A modelling of the stability of aluminium electrolysis cells, in *Nonlinear Partial Differential Equations and Their Applications. Collège de France Seminar, Vol. XIII (Paris, 1994/1996)*, Pitman Res. Notes Math. Ser., Vol. 391 (Longman, 1998), pp. 117–133.
5. E. Weinan and J.-G. Liu, Gauge method for viscous incompressible flows, *Commun. Math. Sci.* **1** (2003) 317–332.

6. S. Poncsak *et al.*, Mathematical modelling of the collective behaviour of gas bubbles under the anode, *CIM Light Metals* (1999) 139–146.
7. R. Glowinski, T. W. Pan, T. I. Hesla, D. D. Joseph and J. Périaux, A fictitious domain approach to the direct numerical simulation of incompressible viscous flow past moving rigid bodies: Application to particulate flow, *J. Comput. Phys.* **169** (2001) 363–426.
8. H. Goldstein, *Classical Mechanics*, 2nd edn. (Addison-Wesley, 1980).
9. C. W. Hirt and B. D. Nichols, Volume of fluid (vof) method for the dynamics of free boundaries, *J. Comput. Phys.* **39** (1981) 201–225.
10. P. Klouček and M. V. Romerio, The detachment of bubbles under a porous rigid surface during aluminum electrolysis, *Math. Mod. Meth. Appl. Sci.* **12** (2002) 1617–1652.
11. S. Osher and R. P. Fedkiw, *Level Set Methods and Dynamic Implicit Surfaces* (Springer-Verlag, 2003).
12. S. Poncsak, L. I. Kiss and R. T. Bui, Mathematical modelling of the growth of gas bubbles under the anode in the aluminum electrolysis cells, *TMS Light Metals* (1999) 57–72.
13. J. A. San Martín, V. Starovoitov and M. Tucsnak, Global weak solutions for the two-dimensional motion of several rigid bodies in an incompressible viscous fluid, *Arch. Rational Mech. Anal.* **161** (2002) 113–147.
14. S. A. Sethian, *Level Set Methods, Evolving Interfaces in Geometry, Fluid Mechanics, Computer Vision, and Material Science* (Cambridge Univ. Press, 1996).
15. V. N. Starovoĭtov, On the nonuniqueness of the solution of the problem of the motion of a rigid body in a viscous incompressible fluid, *Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI)* **306** (2003) 199–209; 231–232.

DUALITY-BASED ASYMPTOTIC-PRESERVING METHOD FOR HIGHLY ANISOTROPIC DIFFUSION EQUATIONS

par

PIERRE DEGOND, FABRICE DELUZET, ALEXEI LOZINSKI, JACEK NARSKI AND CLAUDIA NEGULESCU

soumis à

Communications in Mathematical Sciences

arXiv:1008.3405v1 [math.NA] 19 Aug 2010

DUALITY-BASED ASYMPTOTIC-PRESERVING METHOD FOR HIGHLY ANISOTROPIC DIFFUSION EQUATIONS

PIERRE DEGOND[†‡], FABRICE DELUZET[†‡], ALEXEI LOZINSKI[†], JACEK NARSKI[†], AND CLAUDIA NEGULESCU[§]

Abstract. The present paper introduces an efficient and accurate numerical scheme for the solution of a highly anisotropic elliptic equation, the anisotropy direction being given by a variable vector field. This scheme is based on an asymptotic preserving reformulation of the original system, permitting an accurate resolution independently of the anisotropy strength and without the need of a mesh adapted to this anisotropy. The counterpart of this original procedure is the larger system size, enlarged by adding auxiliary variables and Lagrange multipliers. This Asymptotic-Preserving method generalizes the method investigated in a previous paper [11] to the case of an arbitrary anisotropy direction field.

1. Introduction

Anisotropic problems are common in mathematical modeling of physical problems. They occur in various fields of applications such as flows in porous media [3, 20], semiconductor modeling [29], quasi-neutral plasma simulations [10], image processing [37, 38], atmospheric or oceanic flows [36] and so on, the list being not exhaustive. The initial motivation for this work is closely related to magnetized plasma simulations such as atmospheric plasma [24, 26], internal fusion plasma [4, 12] or plasma thrusters [1]. In this context, the media is structured by the magnetic field, which may be strong in some regions and weak in others. Indeed, the gyration of the charged particles around magnetic field lines dominates the motion in the plane perpendicular to magnetic field. This explains the large number of collisions in the perpendicular plane while the motion along the field lines is rather undisturbed. As a consequence the mobility of particles in different directions differs by many orders of magnitude. This ratio can be as huge as 10^{10}. On the other hand, when the magnetic field is weak the anisotropy is much smaller. As the regions with weak and strong magnetic field can coexist in the same computational domain, one needs a numerical scheme which gives accurate results for a large range of anisotropy strengths. The relevant boundary conditions in many fields of application are periodic (for instance in simulations of the tokamak plasmas on a torus) or Neumann boundary conditions (atmospheric plasma for example [5]). For these reasons we propose a strongly anisotropic model problem for wich we wish to introduce an efficient and accurate numerical scheme. This model problem reads

$$\begin{cases} -\nabla\cdot\mathbb{A}\nabla\phi^\varepsilon = f & \text{in } \Omega, \\ n\cdot\mathbb{A}\nabla\phi^\varepsilon = 0 & \text{on } \partial\Omega_N, \\ \phi^\varepsilon = 0 & \text{on } \partial\Omega_D, \end{cases} \tag{1.1}$$

where $\Omega\subset\mathbb{R}^2$ or $\Omega\subset\mathbb{R}^3$ is a bounded domain with boundary $\partial\Omega=\partial\Omega_D\cup\partial\Omega_N$ and outward normal n. The direction of the anisotropy is defined by a vector field B, where we suppose $\mathrm{div}B=0$ and $B\neq 0$. The direction of B is given by a vector field $b=B/|B|$. The anisotropy matrix is then defined as

$$\mathbb{A}=\frac{1}{\varepsilon}A_\parallel b\otimes b+(Id-b\otimes b)A_\perp(Id-b\otimes b) \tag{1.2}$$

and $\partial\Omega_D=\{x\in\partial\Omega \mid b(x)\cdot n=0\}$. The scalar field $A_\parallel>0$ and the symmetric positive definite matrix field $A_\perp$ are of order one while the parameter $0<\varepsilon<1$ can be very small, provoking thus the high anisotropy of the problem. This work extends the results of [11], where the special case of a vector field b, aligned with the z-axis, was studied. An extension of this approach is proposed in [6] to handle more

[†]Université de Toulouse, UPS, INSA, UT1, UTM, Institut de Mathématiques de Toulouse, F-31062 Toulouse, France
[‡]CNRS, Institut de Mathématiques de Toulouse UMR 5219, F-31062 Toulouse, France
[§]CMI/LATP, Université de Provence, 39 rue Frédéric Joliot-Curie 13453 Marseille cedex 13

realistic anisotropy topologies. It relies on the introduction of a curvilinear coordinate system with one coordinate aligned with the anisotropy direction. Adapted coordinates are widely used in the framework of plasma simulation (see for instance [4, 14, 31]), coordinate systems being either developped to fit particular magnetic field geometry or plasma equilibrium (Euler potentials [35], toroidal and poloidal [18, 23], quasiballooning [15], Hamada [19] and Boozer [7] coordinates). Note that the study of certain plasma regions in a tokamak have motivated the use of non-orthogonal coordinates systems [21]. In contrast with all these methods, we propose here a numerical scheme that uses coordinates and meshes independent of the anisotropy direction, like in [33]. This feature offers the capability to easily treat time evolving anisotropy directions. This is very important in the context of tokamak plasma simulation, the anisotropy being driven by the magnetic field which is time dependent.

One of the difficulties associated with the numerical solution of problem (1.1) lies in the fact that this problem becomes very ill-conditioned for small $0<\varepsilon\ll 1$. Indeed, replacing ε by zero yields an ill-posed problem as it has an infinite number of solutions (any function constant along the b field solves the problem with $\varepsilon=0$). In the discrete case the problem translates into a linear system which is ill-conditioned, as it mixes the terms of different orders of magnitude for $\varepsilon\ll 1$. As a consequence the numerical algorithm for solving this linear system gives unacceptable errors (in the case of direct solvers) or fails to converge in a reasonable time (in the case of iterative methods).

This difficulty arises when the boundary conditions supplied to the dominant $O(1/\varepsilon)$ operator lead to an ill-posed problem. This is the case for Neumann boundary conditions imposed on the part of the boundary with $b\cdot n\neq 0$ as well as for periodic boundary conditions. If instead, the boundary conditions are such that the dominant operator gives a well-posed problem, the numerical difficulty vanishes. One can resort to standard methods, as the dominant operator is sufficient to determine the limit solution. This is the case for Dirichlet and Robin boundary conditions. The problem addressed in this paper arises therefore only with specific boundary conditions. It has however a considerable impact in numerous physical problems concerning plasmas, geophysical flows, plates and shells as an example. In this paper, we will focus on Neumann boundary condition since they represent a larger range of physical applications. The periodic boundary conditions can be addressed in a very similar way.

Numerical methods for anisotropic problems have been extensively studied in the literature. Distinct methods have been developed. Domain decomposition techniques using multiple coarse grid corrections are adapted to the anisotropic equations in [17, 27]. Multigrid methods have been studied in [16, 32]. For anisotropy aligned with one or two directions, point or plane smoothers are shown to be very efficient [28]. The *hp*-finite element method is also known to give good results for singular perturbation problems [30]. All these methods have in common that they try to discretize the anisotropic PDE as it is written and then to apply purely numerical tricks to circumvent the problems related to lack of accuracy of the discrete solution or to the slow convergence of iterative algorithms. This leads to methods which are rather difficult to implement.

The approach that we pursue in this paper is entirely different: we reformulate first the original PDE in such a way that the resulting problem can be efficiently and accurately discretized by straightforward and easily implementable numerical methods for any anisotropy strength. Our scheme is related to the *Asymptotic Preserving* method introduced in [22]. These techniques are designed to give a precise solution in the various regimes with no restrictions on the computational meshes and with additional property of converging to the limit solution when $\varepsilon\to 0$. The derivation of the Asymptotic Preserving method requires identification of the limit model. In the case of Singular Perturbation problems, the original problem is reformulated in such a way that the obtained set of equations contain both the limit model and the original problem with a continuous transition between them, according to the values of ε. This reformulated system of equation sets the foundation of the AP-scheme. These Asymptotic Preserving techniques have been explored in previous studies, for instance quasi-neutral or gyro-fluid limits [9, 12], as well as anisotropic elliptic problems of the form (1.1) with vector b aligned with a

coordinate axis [11, 6].

In this paper, we present a new algorithm which extends the results of [11]. The originality of this algorithm consists in the fact, that it is applicable for variable anisotropy directions b, without additional work. The discretization mesh has not to be adapted to the field direction b, but is simply a Cartesian grid, whose mesh-size is governed by the desired accuracy, independently on the anisotropy strength ε. All this is possible by a well-adapted mathematical framework (optimally chosen spaces, introduction of Lagrange multipliers). The key idea, as in [11], is to decompose the solution ϕ into two parts: a mean part p which is constant along the field lines and the fluctuation part q consisting of a correction to the mean part needed to recover the full solution. Both parts p and q are solutions to well-posed problems for any $\varepsilon>0$. In the limit of $\varepsilon\to 0$ the AP-reformulation reduces to the so called *Limit* model (L-model), whose solution is an acceptable approximation of the P-model solution for $\varepsilon\ll 1$ (see Theorem 2.2). In [11] the Asymptotic Preserving reformulation of the original problem was obtained in two steps. Firstly, the original problem was integrated along the field lines (z-axis) leading to an ε-independent elliptic problem for the mean part p. Secondly, the mean equation was subtracted from the original problem and the ε-dependent elliptic problem for the fluctuating part q was obtained. This approach however is not applicable if the field b is arbitrary. In this paper we present a new approach. Instead of integrating the original problem along the arbitrary field lines, we choose to force the mean part p to lie in the Hilbert space of functions constant along the field lines and the fluctuating part q to be orthogonal (in L^2 sense) to this space. This is done by a Lagrange multiplier technique and requires introduction of additional variables thus enlarging the linear system to be solved. This method allows to treat the arbitrary b field case, regardless of the field topology and thus eliminates the limitations of the algorithm presented in [11]. We note that an alternative method, bypassing the need in Lagrange multipliers, is proposed in [8]. It is based on a reformulation of the original problem as a fourth order equation.

The outline of this paper is the following. Section 2 introduces the original anisotropic elliptic problem. The original problem will be referred to as the Singular-*Perturbation* model (P-model). The mathematical framework is introduced and the *Asymptotic Preserving* reformulation (AP-model) is then derived. Section 3 is devoted to the numerical implementation of the AP-formulation. Numerical results are presented for 2D and 3D test cases, for constant and variable fields b. Three methods are compared (AP-formulation, P-model and L-model) according to their precision for different values of ε. The rigorous numerical analysis of this new algorithm will be the subject of a forthcoming publication.

2. Problem definition

We consider a two or three dimensional anisotropic problem, given on a sufficiently smooth, bounded domain $\Omega\subset\mathbb{R}^d$, $d=2,3$ with boundary $\partial\Omega$. The direction of the anisotropy is defined by the vector field $b\in(C^\infty(\Omega))^d$, satisfying $|b(x)|=1$ for all $x\in\Omega$.
Given this vector field b, one can decompose now vectors $v\in\mathbb{R}^d$, gradients $\nabla\phi$, with $\phi(x)$ a scalar function, and divergences $\nabla\cdot v$, with $v(x)$ a vector field, into a part parallel to the anisotropy direction and a part perpendicular to it. These parts are defined as follows:

$$\begin{aligned}
&v_{||}:=(v\cdot b)b, \qquad v_\perp:=(Id-b\otimes b)v, \qquad \text{such that } v=v_{||}+v_\perp,\\
&\nabla_{||}\phi:=(b\cdot\nabla\phi)b,\ \nabla_\perp\phi:=(Id-b\otimes b)\nabla\phi, \text{ such that } \nabla\phi=\nabla_{||}\phi+\nabla_\perp\phi,\\
&\nabla_{||}\cdot v:=\nabla\cdot v_{||}, \quad \nabla_\perp\cdot v:=\nabla\cdot v_\perp, \qquad \text{such that } \nabla\cdot v=\nabla_{||}\cdot v+\nabla_\perp\cdot v,
\end{aligned} \tag{2.1}$$

where we denoted by $\otimes$ the vector tensor product. With these notations we can now introduce the mathematical problem, the so-called Singular Perturbation problem, whose numerical solution is the main concern of this paper.

2.1. The Singular Perturbation problem (P-model) We consider the following Singular Perturbation problem

$$(P)\ \begin{cases} -\frac{1}{\varepsilon}\nabla_{\|}\cdot\left(A_{\|}\nabla_{\|}\phi^{\varepsilon}\right)-\nabla_{\perp}\cdot(A_{\perp}\nabla_{\perp}\phi^{\varepsilon})=f & \text{in } \Omega,\\ \frac{1}{\varepsilon}n_{\|}\cdot\left(A_{\|}\nabla_{\|}\phi^{\varepsilon}\right)+n_{\perp}\cdot(A_{\perp}\nabla_{\perp}\phi^{\varepsilon})=0 & \text{on } \partial\Omega_N,\\ \phi^{\varepsilon}=0 & \text{on } \partial\Omega_D\,, \end{cases} \tag{2.2}$$

where n is the outward normal to Ω and the boundaries are defined by

$$\partial\Omega_D=\{x\in\partial\Omega \mid b(x)\cdot n=0\},\qquad \partial\Omega_N=\partial\Omega\setminus\partial\Omega_D. \tag{2.3}$$

The parameter $0<\varepsilon<1$ can be very small and is responsible for the high anisotropy of the problem. The aim is to introduce a numerical scheme, whose computational costs (simulation time and memory), for fixed precision, are independent of ε.
We shall assume in the rest of this paper the following hypothesis on the diffusion coefficients and the source terms

Hypothesis A *Let $f\in L^2(\Omega)$ and $\overset{\circ}{\partial\Omega}_D\neq\varnothing$. The diffusion coefficients $A_{\|}\in L^{\infty}(\Omega)$ and $A_{\perp}\in\mathbb{M}_{d\times d}(L^{\infty}(\Omega))$ are supposed to satisfy*

$$0<A_0\leq A_{\|}(x)\leq A_1\,,\quad \textit{f.a.a. } x\in\Omega, \tag{2.4}$$

$$A_0||v||^2\leq v^tA_{\perp}(x)v\leq A_1||v||^2,\quad \forall v\in\mathbb{R}^d \textit{ and f.a.a. } x\in\Omega. \tag{2.5}$$

As we intend to use the finite element method for the numerical solution of the P-problem, let us put (2.2) under variational form. For this let $\mathcal{V}$ be the Hilbert space

$$\mathcal{V}:=\{\phi\in H^1(\Omega)\,/\,\phi_{|\partial\Omega_D}=0\},\quad (\phi,\psi)_{\mathcal{V}}:=(\nabla_{||}\phi,\nabla_{||}\psi)_{L^2}+\varepsilon(\nabla_{\perp}\phi,\nabla_{\perp}\psi)_{L^2}\,.$$

Thus, we are seaking for $\phi^{\varepsilon}\in\mathcal{V}$, the solution of

$$a_{||}(\phi^{\varepsilon},\psi)+\varepsilon a_{\perp}(\phi^{\varepsilon},\psi)=\varepsilon(f,\psi),\quad \forall\psi\in\mathcal{V}, \tag{2.6}$$

where $(\cdot,\cdot)$ stands for the standard L^2 inner product and the continuous bilinear forms $a_{||}:\mathcal{V}\times\mathcal{V}\to\mathbb{R}$ and $a_{\perp}:\mathcal{V}\times\mathcal{V}\to\mathbb{R}$ are given by

$$a_{||}(\phi,\psi):=\int_{\Omega}A_{||}\nabla_{||}\phi\cdot\nabla_{||}\psi\,dx,\quad a_{\perp}(\phi,\psi):=\int_{\Omega}(A_{\perp}\nabla_{\perp}\phi)\cdot\nabla_{\perp}\psi\,dx. \tag{2.7}$$

Thanks to Hypothesis A and the Lax-Milgram theorem, problem (2.2) admits a unique solution $\phi^{\varepsilon}\in\mathcal{V}$ for all fixed $\varepsilon>0$.

2.2. The Limit problem (L-model) The direct numerical solution of (2.2) may be very inaccurate for $\varepsilon\ll 1$. Indeed, when ε tends to zero, the system reduces to

$$\begin{cases} -\nabla_{\|}\cdot\left(A_{\|}\nabla_{\|}\phi\right)=0 & \text{in } \Omega,\\ n_{\|}\cdot\left(A_{\|}\nabla_{\|}\phi\right)=0 & \text{on } \partial\Omega_N,\\ \phi=0 & \text{on } \partial\Omega_D. \end{cases} \tag{2.8}$$

This is an ill-posed problem as it has an infinite number of solutions $\phi \in \mathcal{G}$, where

$$\mathcal{G} = \{\phi \in \mathcal{V} \mid \nabla_{\|}\phi = 0\}, \tag{2.9}$$

is the Hilbert space of functions, which are constant along the field lines of b. This shows that the condition number of the system obtained by discretizing (2.2) tends to ∞ as $\varepsilon \to 0$ so that its solution will suffer from round-off errors.

For this reason, we should approximate (2.2) in the limit $\varepsilon \to 0$ differently. Supposing that $\phi^\varepsilon \to \phi^0$ as $\varepsilon \to 0$ we identify (at first formally) the problem satisfied by ϕ^0. From the above arguments we know that $\phi^0 \in \mathcal{G}$. Taking now test functions $\psi \in \mathcal{G}$ in (2.6), we obtain

$$\int_\Omega A_\perp \nabla_\perp \phi^\varepsilon \cdot \nabla_\perp \psi \, dx = \int_\Omega f\psi \, dx. \tag{2.10}$$

Passing to the limit $\varepsilon \to 0$ into this equation yields the variational formulation of the problem satisfied by ϕ^0 (Limit problem): find $\phi^0 \in \mathcal{G}$, the solution of

$$(L) \int_\Omega A_\perp \nabla_\perp \phi^0 \cdot \nabla_\perp \psi \, dx = \int_\Omega f\psi \, dx \ , \ \forall \psi \in \mathcal{G}, \tag{2.11}$$

which is a well posed problem. Indeed, the space $\mathcal{G} \subset \mathcal{V}$ is a Hilbert space, associated with the inner product

$$(\phi, \psi)_{\mathcal{G}} := (\nabla_\perp \phi, \nabla_\perp \psi)_{L^2}, \quad \forall \phi, \psi \in \mathcal{G}, \tag{2.12}$$

and the norm $||\cdot||_{\mathcal{G}}$ is equivalent to the H^1 norm. This is due to the Poincaré inequality, as

$$||\phi||_{L^2}^2 \le C||\nabla\phi||_{L^2}^2 = C||\nabla_{||}\phi||_{L^2}^2 + C||\nabla_\perp \phi||_{L^2}^2 = C||\nabla_\perp \phi||_{L^2}^2, \quad \forall \phi \in \mathcal{G}.$$

Hypothesis A and the Lax-Milgram lemma imply the existence and uniqueness of a solution $\phi^0 \in \mathcal{G}$ of the Limit problem (2.11).

REMARK 2.1. *Let us restrict ourselves for the moment to the simple special case (considered in a previous paper [11]) of the two dimensional domain $\Omega = (0, L_x) \times (0, L_z)$ in the (x, z) plane with a constant b-field aligned with the Z-axis:*

$$b = \begin{pmatrix} 0 \\ 1 \end{pmatrix}. \tag{2.13}$$

The functions in the space $\mathcal{G}$ are independent of z so that $\mathcal{G}$ can be identified to $H_0^1(0, L_x)$. The limit problem (2.11) *now reads: Find ϕ^0 in $H_0^1(0, L_x)$ verifying*

$$\int_0^{L_x} \bar{A}_\perp(x) \partial_x \phi^0(x) \, \partial_x \psi(x) \, dx = \int_0^{L_x} \bar{f}(x) \psi(x) \, dx, \qquad \forall \psi \in H_0^1(0, L_x),$$

where $\bar{A}_\perp(x) = (1/L_z) \int_0^{L_z} A_{\perp,11}(x, z) \, dz$ and $\bar{f}(x) = (1/L_z) \int_0^{L_z} f(x, z) \, dz$ are the mean values of $A_\perp$ and f along the field lines. The limit solution ϕ^0 thus verifies a one-dimensional elliptic equation whose coefficients are integrated along the anisotropy direction:

$$\begin{aligned} &-\partial_x \Big(\bar{A}_\perp(x) \partial_x \phi^0(x) \Big) = \bar{f}(x) \ \text{ on } (0, L_x), \\ &\phi^0(0) = \phi^0(L_x) = 0. \end{aligned} \tag{2.14}$$

We see now that $\phi^0(x)$ is a solution to the one dimensional elliptic problem so that it belongs to $H^2(0,L_x)$ provided $f \in L^2(\Omega)$. Since ϕ^0 as a function of (x,z) does not depend on z, we have also $\phi^0 \in H^2(\Omega)$. This conclusion ($\phi^0 \in H^2(\Omega)$) remains valid in the case of a cylindrical three dimensional domain $\Omega = \Omega_{xy} \times (0,L_z)$ in the (x,y,z) space with any sufficiently smooth Ω_{xy} in the (x,y) plane and the field b aligned with the Z-axis, $b = (0,0,1)^t$. Indeed, it is easy to see that $\phi^0 = \phi^0(x,y)$ solves in this case an elliptic two dimensional problem in Ω_{xy} similar to (2.14) so that we can apply the standard regularity results for the elliptic problems. These examples show that it is reasonable to suppose $\phi^0 \in H^2(\Omega)$ also in more general geometries of Ω and b. This can be indeed proved under the hypotheses in Appendix A by specifying the $(d-1)$ dimensional elliptic problem for ϕ^0. The proof being rather lengthy and technical, we prefer to postpone it to a forthcoming work [13].

2.3. The Asymptotic Preserving approach (AP-model) In this section we introduce the AP-formulation, which is a reformulation of the Singular Perturbation problem (2.2), permitting a "continuous" transition from the (P)-problem (2.2) to the (L)-problem (2.11), as $\varepsilon \to 0$. For this purpose, each function is decomposed into its mean part along the anisotropy direction (lying in the subspace $\mathcal{G}$ of $\mathcal{V}$) and a fluctuating part (cf. [11]) lying in the L^2-orthogonal complement $\mathcal{A}$ of $\mathcal{G}$ in $\mathcal{V}$, defined by

$$\mathcal{A} := \{\phi \in \mathcal{V} \ | (\phi,\psi) = 0 \ , \ \forall \psi \in \mathcal{G}\}. \tag{2.15}$$

Note that $(\cdot,\cdot)$ denote here and elsewhere the inner product of $L^2(\Omega)$.

In what follows, we need the following

Hypothesis B *The Hilbert-space $\mathcal{V}$ admits the decomposition*

$$\mathcal{V} = \mathcal{G} \oplus^{\perp} \mathcal{A}, \tag{2.16}$$

with $\mathcal{G}$ given by (2.9) and $\mathcal{A}$ given by (2.15) and where the orthogonality of the direct sum is taken with respect to the L^2-norm. Denoting by P the orthogonal projection on $\mathcal{G}$ with respect to the L^2 inner product:

$$P : \mathcal{V} \to \mathcal{G} \ \text{ such that } \ (P\phi,\psi) = (\phi,\psi) \ \ \forall \phi \in \mathcal{V}, \psi \in \mathcal{G}, \tag{2.17}$$

we shall suppose that this mapping is continuous and that we have the Poincaré-Wirtinger inequality

$$||\phi - P\phi||_{L^2(\Omega)} \leq C ||\nabla_{||}\phi||_{L^2(\Omega)}, \quad \forall \phi \in \mathcal{V}. \tag{2.18}$$

Applying the projection P to a function ϕ is nothing but a weighted average of ϕ along the anisotropy field lines of b. The space $\mathcal{G}$ is the space of averaged functions (the parallel $\mathcal{G}$radient of these averaged functions being equal to zero), whereas the space $\mathcal{A}$ is the space of the fluctuations (the $\mathcal{A}$verage of the fluctuations being equal to zero). Note that the decomposition (2.16) is not self evident and it may in fact fail on some "pathological" domains Ω. Indeed, although one can always define an L^2-orthogonal projection $\tilde{P}\phi$ on the space of functions constant along each field line, for any ϕ with square-integrable $\nabla_{||}\phi$, one cannot assure in general that $\tilde{P}\phi$ belongs to $\mathcal{V}$ for $\phi \in \mathcal{V}$ since one may lose control of the perpendicular part of the gradient of $\tilde{P}\phi$. Fortunately however, Hypothesis B is typically satisfied for the domains of practical interest. The interested reader is referred to Appendix A for an example of a set of assumptions on Ω and b which entail Hypothesis B and which resume essentially to the requirement for the field b to intersect $\partial\Omega_N$ in a uniformly non-tangential manner and for the boundary components $\partial\Omega_N$ and $\partial\Omega_D$ to be sufficiently smooth.

Let us also define the operator

$$Q : \mathcal{V} \to \mathcal{A}, \quad Q = I - P. \tag{2.19}$$

Each function $\phi \in \mathcal{V}$ can be decomposed uniquely as $\phi = p + q$, where $p = P\phi \in \mathcal{G}$ and $q = Q\phi \in \mathcal{A}$. Using this decomposition, we reformulate the Singular-Perturbation problem (2.2). Indeed, replacing $\phi^\varepsilon := p^\varepsilon + q^\varepsilon$ in problem (2.2) and taking test functions $\eta \in \mathcal{G}$ and $\xi \in \mathcal{A}$ leads to an asymptotic preserving formulation of the original problem: Find $(p^\varepsilon, q^\varepsilon) \in \mathcal{G} \times \mathcal{A}$ such that

$$\begin{cases} a_\perp(p^\varepsilon, \eta) + a_\perp(q^\varepsilon, \eta) = (f, \eta) & \forall \eta \in \mathcal{G}, \\ a_{||}(q^\varepsilon, \xi) + \varepsilon a_\perp(q^\varepsilon, \xi) + \varepsilon a_\perp(p^\varepsilon, \xi) = \varepsilon(f, \xi) & \forall \xi \in \mathcal{A}. \end{cases} \tag{2.20}$$

Contrary to the Singular Perturbation problem (2.2), setting formally $\varepsilon = 0$ in (2.20) yields the system

$$\begin{cases} a_\perp(p^0, \eta) + a_\perp(q^0, \eta) = (f, \eta), & \forall \eta \in \mathcal{G} \\ a_{||}(q^0, \xi) = 0, & \forall \xi \in \mathcal{A}, \end{cases} \tag{2.21}$$

which has a unique solution $(p^0, q^0) \in \mathcal{G} \times \mathcal{A}$, where p^0 is the unique solution of the L-problem (2.11) and $q^0 \equiv 0$. Indeed, taking $\xi = q^0$ as test function in the second equation of (2.21) yields $\nabla_{||} q^0 = 0$, which means $q^0 \in \mathcal{G}$. But at the same time, $q^0 \in \mathcal{A}$, so that $q^0 \in \mathcal{G} \cap \mathcal{A} = \{0\}$. Setting then $q^0 \equiv 0$ in the first equation of (2.21), shows that p^0 is the unique solution of the L-problem.

THEOREM 2.2. *For every $\varepsilon > 0$ the Asymptotic Preserving formulation (2.20), under Hypotheses A and B, admits a unique solution $(p^\varepsilon, q^\varepsilon) \in \mathcal{G} \times \mathcal{A}$, where $\phi^\varepsilon := p^\varepsilon + q^\varepsilon$ is the unique solution in $\mathcal{V}$ of the Singular Perturbation model (2.2).*
These solutions satisfy the bounds

$$||\phi^\varepsilon||_{H^1(\Omega)} \le C||f||_{L^2(\Omega)}, \quad ||q^\varepsilon||_{H^1(\Omega)} \le C||f||_{L^2(\Omega)}, \quad ||p^\varepsilon||_{H^1(\Omega)} \le C||f||_{L^2(\Omega)}, \tag{2.22}$$

with an ε-independent constant $C > 0$. Moreover, we have

$$\phi^\varepsilon \to \phi^0, \ p^\varepsilon \to \phi^0 \ \text{and} \quad q^\varepsilon \to 0 \quad \text{in} \quad H^1(\Omega) \ \text{as} \ \varepsilon \to 0, \tag{2.23}$$

where $\phi^0 \in \mathcal{G}$ is the unique solution of the Limit model (2.11).

Proof. The existence and uniqueness of a solution for the P-problem as well as L-problem are consequences of the Lax-Milgram theorem. The existence and uniqueness of a solution of (2.20) is then immediate by construction, remarking that the decomposition $\phi^\varepsilon = p^\varepsilon + q^\varepsilon$ is unique.
The bound $||\phi^\varepsilon||_{H^1(\Omega)} \le C||f||_{L^2(\Omega)}$ is obtained by a standard elliptic argument. Furthermore, $p^\varepsilon = P\phi^\varepsilon$ where P is the L^2-orthogonal projector on $\mathcal{G}$, which is a bounded operator in $\mathcal{V}$ by (1.4). This implies the estimates for p^ε and q^ε in (2.22). Since $p^\varepsilon \in \mathcal{G}$ and $q^\varepsilon \in \mathcal{A}$ are bounded, there exist subsequences p^{ε_n} and q^{ε_n} that weakly converge for $\varepsilon_n \to 0$ to some $p^0 \in \mathcal{G}$ and $q^0 \in \mathcal{A}$. Taking $\varepsilon = \varepsilon_n$ in (2.20) and passing to the limit $\varepsilon_n \to 0$ we identify (p^0, q^0) with the unique solution of (2.21), i.e. $p^0 = \phi^0$ is the unique solution of (2.11) and $q^0 \equiv 0$. Since the limit does not depend on the choice of the subsequence, we have the weak convergence as $\varepsilon \to 0$, i.e.

$$p^\varepsilon \rightharpoonup_{\varepsilon \to 0} p^0 \quad \text{in} \quad H^1(\Omega), \quad q^\varepsilon \rightharpoonup_{\varepsilon \to 0} 0 \quad \text{in} \quad H^1(\Omega).$$

We shall prove now that these convergences are actually strong. Introducing $e^\varepsilon = p^\varepsilon - p^0$, we have

$$a_\perp(e^\varepsilon, \eta) + a_\perp(q^\varepsilon, \eta) = 0, \ \forall \eta \in \mathcal{G}.$$

Taking now $\eta = e^\varepsilon$ in this relation and adding it to the second equation in (2.20), where we put $\xi = q^\varepsilon/\varepsilon$, yields

$$\frac{1}{\varepsilon} a_{||}(q^\varepsilon, q^\varepsilon) + a_\perp(q^\varepsilon + e^\varepsilon, q^\varepsilon + e^\varepsilon) = (f, q^\varepsilon) - a_\perp(p^0, q^\varepsilon). \tag{2.24}$$

Due to the Poincaré-Wirtinger equation (2.18), there exist a constant $C>0$ such that

$$||q||_{L^2(\Omega)} \leq C a_{||}(q,q)^{1/2}, \quad \forall q \in \mathcal{A}. \tag{2.25}$$

In combination with a Young inequality this gives $(f,q^\varepsilon) \leq ||f||_{L^2}||q^\varepsilon||_{L^2} \leq \varepsilon \frac{C^2}{2}||f||^2_{L^2} + \frac{1}{2\varepsilon}a_{||}(q^\epsilon,q^\epsilon)$. Using this in the right hand side of (2.24), we arrive at

$$\frac{1}{2\varepsilon}a_{||}(q^\varepsilon,q^\varepsilon) + a_\perp(q^\varepsilon+e^\varepsilon,q^\varepsilon+e^\varepsilon) \leq \varepsilon\frac{C^2}{2}||f||^2_{L^2} - a_\perp(p^0,q^\varepsilon).$$

Noting that $q^\varepsilon + e^\varepsilon = \phi^\varepsilon - p^0$ and $\nabla_\parallel e^\varepsilon = 0$ we can rewrite this last inequality as

$$\frac{1}{2\varepsilon}a_{||}(\phi^\varepsilon-p^0,\phi^\varepsilon-p^0) + a_\perp(\phi^\varepsilon-p^0,\phi^\varepsilon-p^0) \leq \varepsilon\frac{C^2}{2}||f||^2_{L^2} - a_\perp(p^0,q^\varepsilon).$$

Since $a_\perp(p^0,q^\varepsilon) \to 0$ as $\varepsilon \to 0$ (thanks to the weak convergence $q^\varepsilon \rightharpoonup 0$) we observe that $\phi^\varepsilon \to p^0$ strongly in $H^1(\Omega)$. Reminding again that $p^\varepsilon = P\phi^\varepsilon$ and P is bounded in the norm of $H^1(\Omega)$, we obtain also $p^\varepsilon \to Pp^0 = p^0$, which entails $q^\varepsilon \to 0$. □

REMARK 2.3. *Let us return to the simple special case discussed in remark 2.1, i.e.* $\Omega = (0,L_x)\times(0,L_z)$ *and the b-field given by (2.13). Remind that the space* $\mathcal{G}$ *can be identified in this case with the space of functions constant along the Z-axis, which means* $\mathcal{G} := \{\phi \in \mathcal{V} \,/\, \partial_z\phi = 0\}$. *The space* $\mathcal{A}$ *is orthogonal (with respect to the* L^2*-norm) to* $\mathcal{G}$ *and thus contains the functions that have zero mean value along the Z-axis, i.e.* $\mathcal{A} := \{\phi \in \mathcal{V} \,/\, \int_0^{L_z}\phi(x,z)\,dz = 0\}$. *Therefore, for* $\phi^\varepsilon = p^\varepsilon + q^\varepsilon \in \mathcal{V}$, *the function* p^ε *is the mean value of* ϕ^ε *in the direction of the field b:*

$$p^\varepsilon = \frac{1}{L_z}\int_0^{L_z}\phi^\varepsilon\,dz, \tag{2.26}$$

and q^ε *is the fluctuating part with zero mean value:*

$$q^\varepsilon = \phi^\varepsilon - \frac{1}{L_z}\int_0^{L_z}\phi^\varepsilon\,dz. \tag{2.27}$$

Hypothesis B is thus easily verified. The results obtained in this special case were presented in a previous paper [11]. In the case of an arbitrary b-field, formula (2.26) is generalized as (1.2) in Appendix A, where the length element along the b-field line is weighted by the infinitesimal cross-sectional area of the field tube around the considered b-field-line. This formula can be thus interpreted as a consequence of the co-area formula. Note that in the special case of a uniform anisotropy direction, the limit problem can easily be formulated as an elliptic problem depending on the only transverse coordinates (see equation (2.14)*). The size of the problem is thus significantly smaller than that of the initial one. This feature still occurs for non-uniform b-fields as long as adapted coordinates and meshes are used. In our case, aligned and transverse coordinates are not at our disposal and the solution of the limit problem must be searched as a function of the whole set of coordinates.*

2.4. Lagrange multiplier space The objective of this work is the numerical solution of system (2.20) and the comparison of the obtained results with those obtained by directly solving the original problem (2.2). In a general case, when the field b is not necessarily constant, the discretization of the subspaces $\mathcal{G}$ and $\mathcal{A}$, is not straightforward, as in the simpler case [11]. In order to overcome this difficulty a Lagrange multiplier technique will be used.

2.4.1. The $\mathcal{A}$ space To avoid the use of the constrained space $\mathcal{A}$, we can remark that $\mathcal{A}$ can be characterized as being the orthogonal complement (in the L^2 sense) of the $\mathcal{G}$-space. Thus, instead of (2.20), the slightly changed system will be solved: find $(p^\varepsilon, q^\varepsilon, l^\varepsilon) \in \mathcal{G} \times \mathcal{V} \times \mathcal{G}$ such that

$$\begin{cases} a_\perp(p^\varepsilon,\eta) + a_\perp(q^\varepsilon,\eta) = (f,\eta) & \forall \eta \in \mathcal{G}, \\ a_{||}(q^\varepsilon,\xi) + \varepsilon a_\perp(q^\varepsilon,\xi) + \varepsilon a_\perp(p^\varepsilon,\xi) + (l^\varepsilon,\xi) = \varepsilon(f,\xi) & \forall \xi \in \mathcal{V}, \\ (q^\varepsilon,\chi) = 0 & \forall \chi \in \mathcal{G}. \end{cases} \tag{2.28}$$

The constraint $(q^\varepsilon,\chi) = 0$, $\forall \chi \in \mathcal{G}$ is forcing the solution q^ε to belong to $\mathcal{A}$, and this property is carried over to the limit $\varepsilon \to 0$. We have thus circumvented the difficulty of discretizing $\mathcal{A}$ by introducing a new variable and enlarging the linear system.

PROPOSITION 2.4. *Problems (2.20) and (2.28) are equivalent. Indeed, $(p^\varepsilon, q^\varepsilon) \in \mathcal{G} \times \mathcal{A}$ is the unique solution of (2.20) if and only if $(p^\varepsilon, q^\varepsilon, l^\varepsilon) \in \mathcal{G} \times \mathcal{V} \times \mathcal{G}$ with $l^\varepsilon \equiv 0$ is the unique solution of (2.28).*

Proof. Let $(p^\varepsilon, q^\varepsilon) \in \mathcal{G} \times \mathcal{A}$ be the unique solution of (2.20). Then, it is immediate to show that $(p^\varepsilon, q^\varepsilon, 0)$ solves (2.28). Let now $(p^\varepsilon, q^\varepsilon, l^\varepsilon) \in \mathcal{G} \times \mathcal{V} \times \mathcal{G}$ be a solution of (2.28). Then, the last equation of (2.28) implies that $q^\varepsilon \in \mathcal{A}$. Choosing in the second equation as test function $\xi \in \mathcal{G}$, one gets

$$\varepsilon a_\perp(q^\varepsilon,\xi) + \varepsilon a_\perp(p^\varepsilon,\xi) + (l^\varepsilon,\xi) = \varepsilon(f,\xi), \quad \forall \xi \in \mathcal{G},$$

which because of the first equation in (2.28), yields $(l^\varepsilon,\xi) = 0$ for all $\xi \in \mathcal{G}$. Thus $l^\varepsilon \equiv 0$. □

2.4.2. The $\mathcal{G}$ space In order to eliminate the problems that arise when dealing with the discretization of $\mathcal{G}$, the Lagrange multiplier method will again be used. First note that

$$p \in \mathcal{G} \Leftrightarrow \begin{cases} \nabla_{||} p = 0 \\ p \in \mathcal{V} \end{cases} \Leftrightarrow \begin{cases} \displaystyle\int_\Omega A_{||} \nabla_{||} p \cdot \nabla_{||} \lambda \, dx = a_{||}(p,\lambda) = 0, \ \forall \lambda \in \mathcal{L} \\ p \in \mathcal{V}, \end{cases} \tag{2.29}$$

where $\mathcal{L}$ is a functional space that should be chosen large enough so that one could find for any $p \in \mathcal{V}$ a $\lambda \in \mathcal{L}$ with $\nabla_{||}\lambda = \nabla_{||} p$. On the other hand, the space $\mathcal{L}$ should be not too large in order to ensure the uniqueness of the Lagrange multipliers in the unconstrained system. A space that satisfies these two requirements under some quite general assumptions to be detailed later, can be defined as

$$\mathcal{L} := \{\lambda \in L^2(\Omega) \,/\, \nabla_{||}\lambda \in L^2(\Omega),\ \lambda_{|\partial\Omega_{in}} = 0\}, \quad \text{with} \quad \partial\Omega_{in} := \{x \in \partial\Omega \,/\, b(x) \cdot n < 0\}. \tag{2.30}$$

Using the characterization (2.29) of the constrained space $\mathcal{G}$, we shall now reformulate the system (2.28) as follows: Find $(p^\varepsilon, \lambda^\varepsilon, q^\varepsilon, l^\varepsilon, \mu^\varepsilon) \in \mathcal{V} \times \mathcal{L} \times \mathcal{V} \times \mathcal{V} \times \mathcal{L}$ such that

$$(AP) \begin{cases} a_\perp(p^\varepsilon,\eta) + a_\perp(q^\varepsilon,\eta) + a_{||}(\eta,\lambda^\varepsilon) = (f,\eta), & \forall \eta \in \mathcal{V}, \\ a_{||}(p^\varepsilon,\kappa) = 0, & \forall \kappa \in \mathcal{L}, \\ a_{||}(q^\varepsilon,\xi) + \varepsilon a_\perp(q^\varepsilon,\xi) + \varepsilon a_\perp(p^\varepsilon,\xi) + (l^\varepsilon,\xi) = \varepsilon(f,\xi), & \forall \xi \in \mathcal{V}, \\ (q^\varepsilon,\chi) + a_{||}(\chi,\mu^\varepsilon) = 0, & \forall \chi \in \mathcal{V}, \\ a_{||}(l^\varepsilon,\tau) = 0, & \forall \tau \in \mathcal{L}. \end{cases} \tag{2.31}$$

The advantage of the above formulation, as compared to (2.20), is that we only have to discretize the spaces $\mathcal{V}$ and $\mathcal{L}$ (at the price of the introduction of three additional variables), which is much easier than the discretization of the constrained spaces $\mathcal{G}$ and $\mathcal{A}$. More importantly, the dual formulation (2.31) does not require any change of coordinates to express the fact that p^ε is constant along the b-field lines

and that q^ε averages to zero along these lines. Therefore this formulation is particularly well adapted to time-dependent b-fields, as it does not require any operation which would have to be reinitiated as b evolves. The system (2.31) will be called the Asymptotic-Preserving formulation in the sequel.

To analyse this Asymptotic-Preserving formulation, we need the following

Hypothesis B' *The trace $\lambda_{|\partial\Omega_{in}}$ is well defined for any $\lambda\in\tilde{\mathcal{V}}$ as an element of $L^2(\partial\Omega_{in})$, with continuous dependence of the trace norm in $L^2(\partial\Omega_{in})$ on $||\lambda||_{\tilde{\mathcal{V}}}$. Moreover, the Hilbert space*

$$\tilde{\mathcal{V}}=\{\phi\in L^2(\Omega)\,/\,\nabla_{||}\phi\in L^2(\Omega)\},\quad (\phi,\psi)_{\tilde{\mathcal{V}}}:=(\phi,\psi)+(\nabla_{||}\phi,\nabla_{||}\psi), \tag{2.32}$$

admits the decomposition

$$\tilde{\mathcal{V}}=\tilde{\mathcal{G}}\oplus\mathcal{L}, \tag{2.33}$$

where $\tilde{\mathcal{G}}$ is given by

$$\tilde{\mathcal{G}}:=\{\phi\in\tilde{\mathcal{V}}\,/\,\nabla_{||}\phi=0\}, \tag{2.34}$$

and $\mathcal{L}$ is given by (2.30). The spaces $\tilde{\mathcal{G}}$ and $G=\tilde{\mathcal{G}}\cap\mathcal{V}$ are related in the following way: if $g\in\tilde{\mathcal{G}}$ is such that $\int_{\partial\Omega_{in}}\eta g d\sigma=0$ for all $\eta\in\mathcal{G}$, then $g=0$.

The decomposition (2.33) is quite natural. It tells simply that any function ϕ can be decomposed on each field line as a sum of a function that vanishes at one given point on this line and a constant (which is therefore the value of ϕ at this point). Hypothesis B' will be thus normally satisfied in cases of practical interest. For example, we prove in Appendix A that the set of assumptions on the domain Ω and the b-field which can be used to verify Hypothesis B, is also sufficient (but far from necessary) for Hypothesis B'. We are now able to show the relation between systems (2.28) and (2.31).

PROPOSITION 2.5. *Assuming Hypotheses A, B and B', problem (2.31) admits a unique solution $(p^\varepsilon,\lambda^\varepsilon,q^\varepsilon,l^\varepsilon,\mu^\varepsilon)\in\mathcal{V}\times\mathcal{L}\times\mathcal{V}\times\mathcal{V}\times\mathcal{L}$, where $(p^\varepsilon,q^\varepsilon,l^\varepsilon)\in\mathcal{G}\times\mathcal{V}\times\mathcal{G}$ is the unique solution of (2.28).*

The proof of Proposition 2.5 is based on the following two lemmas

LEMMA 2.6. *Assume Hypothesis B' and let $p\in\tilde{\mathcal{V}}$ be such that $a_{||}(p,\lambda)=0$, $\forall\lambda\in\mathcal{L}$. Then $p\in\tilde{\mathcal{G}}$.*

Proof. Take any $\eta\in\tilde{\mathcal{V}}$ and decompose $\eta=\lambda+g$ with $\lambda\in\mathcal{L}$ and $g\in\tilde{\mathcal{G}}$. We have $a_{||}(p,g)=0$, hence $a_{||}(p,\eta)=0$ for all $\eta\in\tilde{\mathcal{V}}$. This entails $\nabla_{||}p=0$, hence $p\in\tilde{\mathcal{G}}$. □

LEMMA 2.7. *Assume Hypothesis B' and let $F\in\tilde{\mathcal{V}}^*$ be such that $F(\eta)=0$ for all $\eta\in\mathcal{G}$. Then the problem of finding $\lambda\in\mathcal{L}$ such that*

$$a_{||}(\eta,\lambda)=F(\eta),\quad\forall\eta\in\tilde{\mathcal{V}}, \tag{2.35}$$

has a unique solution.

Proof. Consider the bilinear form b on $\tilde{\mathcal{V}}\times\tilde{\mathcal{V}}$

$$b(u,v)=a_{||}(u,v)+\int_{\partial\Omega_{in}}uvd\sigma$$

By Hypothesis B', this is an inner product on $\tilde{\mathcal{V}}$. Indeed, if $b(u,u)=0$ then $u\in\tilde{\mathcal{G}}\cap\mathcal{L}$ so that $u=0$. Riesz representation theorem implies that the problem of finding $\mu\in\tilde{\mathcal{V}}$ such that

$$b(\eta,\mu)=F(\eta),\quad\forall\eta\in\tilde{\mathcal{V}},$$

has a unique solution. We can now decompose $\mu=\lambda+g$ with $\lambda\in\mathcal{L}$ and $g\in\tilde{\mathcal{G}}$. This yields

$$a_{||}(\eta,\lambda)+\int_{\partial\Omega_{in}}\eta g d\sigma=F(\eta),\quad\forall\eta\in\tilde{\mathcal{V}},$$

so that, in particular, $\int_{\partial\Omega_{in}} \eta g d\sigma = 0$ for all $\eta \in \mathcal{G}$ which implies $g = 0$. We see now that λ is a solution to (2.35). The uniqueness follows easily. □

Let us now prove Proposition 2.5.

Proof of existence in Proposition 2.5. Take $(p^\varepsilon, q^\varepsilon, l^\varepsilon) \in \mathcal{G} \times \mathcal{V} \times \mathcal{G}$ as the unique solution of (2.28). Then, equations 2,3,5 in (2.31) are immediately satisfied. It remains to choose properly the Lagrange multipliers $\lambda^\varepsilon, \mu^\varepsilon \in \mathcal{L}$ to satisfy equations 1,4 in (2.31). For this, let us define $F_1, F_2 \in \tilde{\mathcal{V}}^*$ by

$$F_1(\eta) := \frac{1}{\varepsilon} a_{||}(q^\varepsilon, \eta), \quad F_2(\eta) := -(q^\varepsilon, \eta), \quad \forall \eta \in \tilde{\mathcal{V}}. \tag{2.36}$$

These functionals are indeed continuous in the norm of $\tilde{\mathcal{V}}$ since their definitions do not contain the derivatives in directions perpendicular to b. Since $F_1(\eta) = F_2(\eta) = 0$ for all $\eta \in \mathcal{G}$, Lemma 2.7 implies the existence of $\lambda^\varepsilon \in \mathcal{L}$ and $\mu^\varepsilon \in \mathcal{L}$, such that

$$a_{||}(\eta, \lambda^\varepsilon) = F_1(\eta), \quad a_{||}(\chi, \mu^\varepsilon) = F_2(\chi), \quad \forall \eta, \chi \in \tilde{\mathcal{V}}. \tag{2.37}$$

Taking $\eta, \chi \in \mathcal{V} \subset \tilde{\mathcal{V}}$ we observe (cf. the second line in (2.28) where $l^\varepsilon = 0$)

$$a_{||}(\eta, \lambda^\varepsilon) = \frac{1}{\varepsilon} a_{||}(q^\varepsilon, \eta) = (f, \eta) - a_\perp(p^\varepsilon, \eta) - a_\perp(q^\varepsilon, \eta), \quad \forall \eta \in \mathcal{V},$$

$$a_{||}(\chi, \mu^\varepsilon) = -(q^\varepsilon, \chi), \quad \forall \chi \in \mathcal{V},$$

which coincides with equations 1,4 in (2.31).

Proof of uniqueness in Proposition 2.5. Consider the solution to system (2.31) with $f = 0$. Lemma 2.6 implies then that $p^\varepsilon, l^\varepsilon \in \tilde{\mathcal{G}} \cap \mathcal{V} = \mathcal{G}$ and $(p^\varepsilon, q^\varepsilon, l^\varepsilon) \in \mathcal{G} \times \mathcal{V} \times \mathcal{G}$ verifies (2.28) with $f = 0$ so that $p^\varepsilon = q^\varepsilon = l^\varepsilon = 0$ by Proposition 2.4. Equations 1,4 in (2.31) now tell us that $\lambda^\varepsilon, \mu^\varepsilon \in \tilde{\mathcal{G}}$, but $\tilde{\mathcal{G}} \cap \mathcal{L} = \{0\}$, hence $\lambda^\varepsilon = \mu^\varepsilon = 0$. □

The presence of $1/\varepsilon$ in the formulas (2.36), (2.37) defining λ^ε indicates at a first sight that λ^ε may tend to ∞ as $\varepsilon \to 0$ which would be disastrous for an AP numerical method based on (2.31) at very small ε. Fortunately λ^ε remains bounded uniformly in ε in the cases of practical interest. It suffices to suppose that the limit solution ϕ^0 is in $H^2(\Omega)$ which is a reasonable assumption as discussed in Remark 2.1.

PROPOSITION 2.8. *Assume Hypotheses A, B, B' and $\phi^0 \in H^2(\Omega)$ where ϕ^0 is the solution to (2.11). Then λ^ε introduced in (2.31) satisfies*

$$||\nabla_{||} \lambda^\varepsilon||_{L^2} \leq C \max(||f||_{L^2}, ||\phi^0||_{H^2}) \tag{2.38}$$

with a constant C independent of ε.

Proof. We will denote all the ε-independent constants by C in this proof. We start from relation (2.24) in the proof of Theorem 2.2. Dropping the positive term $a_\perp(q^\varepsilon + e^\varepsilon, q^\varepsilon + e^\varepsilon)$ it can be rewritten as

$$\frac{1}{\varepsilon} a_{||}(q^\varepsilon, q^\varepsilon) \leq (f, q^\varepsilon) - a_\perp(\phi^0, q^\varepsilon).$$

Since $\phi^0 \in H^2(\Omega)$ we can integrate by parts in the integral defining $a_\perp(\phi^0, q^\varepsilon)$:

$$\begin{aligned} -a_\perp(\phi^0, q^\varepsilon) &= -\int_\Omega A_\perp \nabla_\perp \phi^0 \cdot \nabla_\perp q^\varepsilon dx \\ &= -\int_{\partial\Omega_N} (Id - bb^t) A_\perp \nabla_\perp \phi^0 \cdot n q^\varepsilon d\sigma + \int_\Omega (\nabla_\perp \cdot A_\perp \nabla_\perp \phi^0) q^\varepsilon dx \\ &\leq C ||\phi^0||_{H^2} \left(||q^\varepsilon||_{L^2(\partial\Omega_N)} + ||q^\varepsilon||_{L^2(\Omega)} \right) \end{aligned}$$

since $\nabla\phi^0$ has a trace on $\partial\Omega$ and its norm in $L^2(\partial\Omega_N)$ is bounded by $C||\phi^0||_{H^2}$. Thus,

$$\frac{1}{\varepsilon}||\nabla_{||}q^\varepsilon||_{L^2}^2 \le \frac{C}{\varepsilon}a_{||}(q^\varepsilon,q^\varepsilon) \le C||f||_{L^2}||q^\varepsilon||_{L^2(\Omega)} + C||\phi^0||_{H^2}\left(||q^\varepsilon||_{L^2(\partial\Omega_N)} + ||q^\varepsilon||_{L^2(\Omega)}\right).$$

By Poincaré-Wirtinger inequality (2.18) (note that $Pq^\varepsilon = 0$) and by Hypothesis B' we have

$$\max(||q^\varepsilon||_{L^2(\Omega)}, ||q^\varepsilon||_{L^2(\partial\Omega_N)}) \le C||\nabla_{||}q^\varepsilon||_{L^2}$$

so that

$$\frac{1}{\varepsilon}||\nabla_{||}q^\varepsilon||_{L^2} \le C\max(||f||_{L^2}, ||\phi^0||_{H^2}).$$

This is the same as (2.38) since $\nabla_{||}\lambda^\varepsilon = \frac{1}{\varepsilon}\nabla_{||}q^\varepsilon$ according to (2.36) and (2.37). □

REMARK 2.9. *The Limit model (2.11), reformulated using the Lagrange multiplier technique, now reads: Find* $(\phi^0, \lambda^0) \in \mathcal{V}\times\mathcal{L}$ *such that*

$$(L') \quad \begin{cases} \displaystyle\int_\Omega A_\perp\nabla_\perp\phi^0\cdot\nabla_\perp\psi\,dx + \int_\Omega A_{||}\nabla_{||}\psi\cdot\nabla_{||}\lambda^0\,dx = \int_\Omega f\psi\,dx \quad \forall\psi\in\mathcal{V} \\ \displaystyle\int_\Omega A_{||}\nabla_{||}\phi^0\cdot\nabla_{||}\kappa\,dx = 0 \qquad\qquad \forall\kappa\in\mathcal{L}. \end{cases} \tag{2.39}$$

Problem (2.39) is also well posed assuming Hypotheses A, B, B' and $\phi^0 \in H^2(\Omega)$. *Indeed, the uniqueness of the solution to (2.39) can be proved in exactly the same manner as in the proof of Proposition 2.5 above. To prove the existence of a solution, it suffices to take the limit* $\varepsilon\to 0$ *in the first two lines of (2.31). Indeed, we know by Theorem 2.2 that* $p^\varepsilon\to\phi^0$, *the solution to (2.11), and* $q^\varepsilon\to 0$ *in* $H^1(\Omega)$. *Moreover, the family* $\{\nabla_{||}\lambda^\varepsilon\}$ *is bounded in the norm of* $L^2(\Omega)$ *by Proposition 2.8. We can take therefore a weakly convergence subsequence* $\{\nabla_{||}\lambda^{\varepsilon_n}\}$ *and identify its limit with* $\{\nabla_{||}\lambda^0\}$ *with some* $\lambda^0\in\mathcal{L}$ *(cf. Lemma 2.7) to see that* $(\phi^0,\lambda^0)\in\mathcal{V}\times\mathcal{L}$ *solves (2.39).*

3. Numerical method

This section concerns the discretization of the Asymptotic Preserving formulation (2.31), based on a finite element method, and the detailed study of the obtained numerical results. The numerical analysis of the present scheme is investigated in a forthcoming work [13], in particular we are interested in the convergence of the scheme, independently of the parameter $\varepsilon > 0$.

Let us denote by $\mathcal{V}_h \subset \mathcal{V}$ and $\mathcal{L}_h \subset \mathcal{L}$ the finite dimensional approximation spaces, constructed by means of appropriate numerical discretizations (see Section 3.1 and Appendix **B**). We are thus looking for a discrete solution $(p_h^\varepsilon, \lambda_h^\varepsilon, q_h^\varepsilon, l_h^\varepsilon, \mu_h^\varepsilon) \in \mathcal{V}_h\times\mathcal{L}_h\times\mathcal{V}_h\times\mathcal{V}_h\times\mathcal{L}_h$ of the following system

$$\begin{cases} a_\perp(p_h^\varepsilon,\eta) + a_\perp(q_h^\varepsilon,\eta) + a_{||}(\eta,\lambda_h^\varepsilon) = (f,\eta), \quad \forall\eta\in\mathcal{V}_h, \\ a_{||}(p_h^\varepsilon,\kappa) = 0, \quad \forall\kappa\in\mathcal{L}_h, \\ a_{||}(q_h^\varepsilon,\xi) + \varepsilon a_\perp(q_h^\varepsilon,\xi) + \varepsilon a_\perp(p_h^\varepsilon,\xi) + (l_h^\varepsilon,\xi) = \varepsilon(f,\xi), \quad \forall\xi\in\mathcal{V}_h, \\ (q_h^\varepsilon,\chi) + a_{||}(\chi,\mu_h^\varepsilon) = 0, \quad \forall\chi\in\mathcal{V}_h, \\ a_{||}(l_h^\varepsilon,\tau) = 0, \quad \forall\tau\in\mathcal{L}_h. \end{cases} \tag{3.1}$$

Our numerical experiments indicate that the spaces $\mathcal{V}_h$ and $\mathcal{L}_h$ can be always taken of the same type and on the same mesh. The only difference between these two finite element spaces lies thus in

the incorporation of boundary conditions. In general, let X_h denote the complete finite element space (without any restrictions on the boundary) which should be H^1 conforming but otherwise arbitrarily chosen. We define then

$$\mathcal{V}_h = \{v_h \in X_h / v_h|_{\partial\Omega_D} = 0\}, \tag{3.2}$$

$$\mathcal{L}_h = \{\lambda_h \in X_h / \lambda_h|_{\partial\Omega_{in}\cup\partial\Omega_D} = 0\}. \tag{3.3}$$

While this choice of $\mathcal{V}_h$ is straight forward, the boundary conditions in $\mathcal{L}_h$ require special attention. Indeed, nothing in the definition (2.30) of space $\mathcal{L}$ on the continuous level indicates that its elements should vanish on $\partial\Omega_D$. However, this liberty on $\partial\Omega_D$ is somewhat counter-intuitive. Indeed, the Lagrange multiplier $\lambda^\varepsilon \in \mathcal{L}$ serves to impose $\nabla_{||} p^\varepsilon = 0$ for some function p^ε taken from the space $\mathcal{V}$. But, for $p \in \mathcal{V}$ the trace on $\partial\Omega_D$ is zero so that $\nabla_{||} p^\varepsilon = 0$ there without the help of a Lagrange multiplier. Of course, this argument is not valid on the continuous level since the trace of functions in $\mathcal{L}$ does not even necessarily exist. However, this may become very important on the finite element level. Indeed, we provide in Appendix B an example of a finite element setting without incorporating $\lambda_h|_{\partial\Omega_D} = 0$ into the definition of $\mathcal{L}_h$, which leads to an ill-posed system (3.1). To avoid this difficulty, we choose $\mathcal{L}_h$ as in (3.3) in all our experiments, thus obtaining well-posed problems.

3.1. Discretization Let us present the discretization in a 2D case, the 3D case being a simple generalization. The here considered computational domain Ω is a square $\Omega = [0,1]\times[0,1]$. All simulations are performed on structured meshes. Let us introduce the Cartesian, homogeneous grid

$$x_i = i/N_x \ , \ 0 \le i \le N_x, \quad y_j = j/N_y \ , \ 0 \le j \le N_y, \tag{3.4}$$

where N_x and N_y are positive even constants, corresponding to the number of discretization intervals in the x- resp. y-direction. The corresponding mesh-sizes are denoted by $h_x > 0$ resp. $h_y > 0$. Choosing a $\mathbb{Q}_2$ finite element method ($\mathbb{Q}_2$-FEM), based on the following quadratic base functions

$$\theta_{x_i} = \begin{cases} \frac{(x-x_{i-2})(x-x_{i-1})}{2h_x^2} & x \in [x_{i-2}, x_i], \\ \frac{(x_{i+2}-x)(x_{i+1}-x)}{2h_x^2} & x \in [x_i, x_{i+2}], \\ 0 & \text{else} \end{cases}, \quad \theta_{y_j} = \begin{cases} \frac{(y-y_{j-2})(y-y_{j-1})}{2h_y^2} & y \in [y_{j-2}, y_j], \\ \frac{(y_{j+2}-y)(y_{j+1}-y)}{2h_y^2} & y \in [y_j, y_{j+2}], \\ 0 & \text{else} \end{cases} \tag{3.5}$$

for even i, j and

$$\theta_{x_i} = \begin{cases} \frac{(x_{i+1}-x)(x-x_{i-1})}{h_x^2} & x \in [x_{i-1}, x_{i+1}], \\ 0 & \text{else} \end{cases}, \quad \theta_{y_j} = \begin{cases} \frac{(y_{j+1}-y)(y-y_{j-1})}{h_y^2} & y \in [y_{j-1}, y_{j+1}], \\ 0 & \text{else} \end{cases} \tag{3.6}$$

for odd i, j, we define

$$X_h := \{v_h = \sum_{i,j} v_{ij}\, \theta_{x_i}(x)\, \theta_{y_j}(y)\},$$

We then search for discrete solutions $(p_h^\varepsilon, q_h^\varepsilon, l_h^\varepsilon) \in \mathcal{V}_h \times \mathcal{V}_h \times \mathcal{V}_h$ and $(\lambda_h^\varepsilon, \mu_h^\varepsilon) \in \mathcal{L}_h \times \mathcal{L}_h$ with $\mathcal{V}_h$ and $\mathcal{L}_h$ defined by (3.2) and (3.3). This leads to the inversion of a linear system, the corresponding matrix being non-symmetric and given by

$$A = \begin{pmatrix} A_1 & A_0 & A_1 & 0 & 0 \\ A_0 & 0 & 0 & 0 & 0 \\ \varepsilon A_1 & 0 & A_0 + \varepsilon A_1 & C & 0 \\ 0 & 0 & C & 0 & A_0 \\ 0 & 0 & 0 & A_0 & 0 \end{pmatrix}. \tag{3.7}$$

ε	AP scheme		Limit model		Singular Perturbation scheme	
	L^2 error	H^1 error	L^2 error	H^1 error	L^2 error	H^1 error
10	7.2×10^{-6}	4.7×10^{-3}	5.0×10^{0}	3.51×10^{1}	7.2×10^{-6}	4.7×10^{-3}
1	7.3×10^{-7}	4.7×10^{-4}	5.0×10^{-1}	3.51×10^{0}	7.3×10^{-7}	4.7×10^{-4}
10^{-1}	1.47×10^{-7}	9.6×10^{-5}	5.0×10^{-2}	3.51×10^{-1}	1.45×10^{-7}	9.4×10^{-5}
10^{-4}	1.28×10^{-7}	8.3×10^{-5}	5.0×10^{-5}	3.61×10^{-4}	1.26×10^{-7}	8.2×10^{-5}
10^{-6}	1.28×10^{-7}	8.3×10^{-5}	5.2×10^{-7}	8.4×10^{-5}	5.9×10^{-7}	8.2×10^{-5}
10^{-10}	1.28×10^{-7}	8.3×10^{-5}	1.28×10^{-7}	8.3×10^{-5}	9.9×10^{-3}	3.12×10^{-2}
10^{-15}	1.28×10^{-7}	8.3×10^{-5}	1.28×10^{-7}	8.3×10^{-5}	7.1×10^{-1}	2.23×10^{0}

Table 3.1 – Comparison between the Asymptotic Preserving scheme, the Limit model and the Singular Perturbation model for $h=0.005$ (200 mesh points in each direction) and constant b: absolute L^2-error and H^1-error, for different ε-values.

The sub-matrices A_0, A_1 resp. C correspond to the bilinear forms $a_{||}(\cdot,\cdot)$, $a_{\perp}(\cdot,\cdot)$ resp. $(\cdot,\cdot)$, used in equations (2.31) and belong to $\mathbb{R}^{(N_x+1)(N_y+1)\times(N_x+1)(N_y+1)}$. The matrix elements are computed using the 2D Gauss quadrature formula, with 3 points in the x and y direction:

$$\int_{-1}^{1}\int_{-1}^{1} f(x,y) = \sum_{i,j=-1}^{1} \omega_i\omega_j f(x_i,y_j), \tag{3.8}$$

where $x_0=y_0=0$, $x_{\pm1}=y_{\pm1}=\pm\sqrt{\frac{3}{5}}$, $\omega_0=8/9$ and $\omega_{\pm1}=5/9$, which is exact for polynomials of degree 5.

3.2. Numerical Results

3.2.1. 2D test case, uniform and aligned b-field In this section we compare the numerical results obtained via the $\mathbb{Q}_2$-FEM, by discretizing the Singular Perturbation model (2.2), the Limit model (2.11) and the Asymptotic Preserving reformulation (2.31). In all numerical tests we set $A_{\perp}=Id$ and $A_{\parallel}=1$. We start with a simple test case, where the analytical solution is known. Let the source term f be given by

$$f=(4+\varepsilon)\pi^2\cos(2\pi x)\sin(\pi y)+\pi^2\sin(\pi y) \tag{3.9}$$

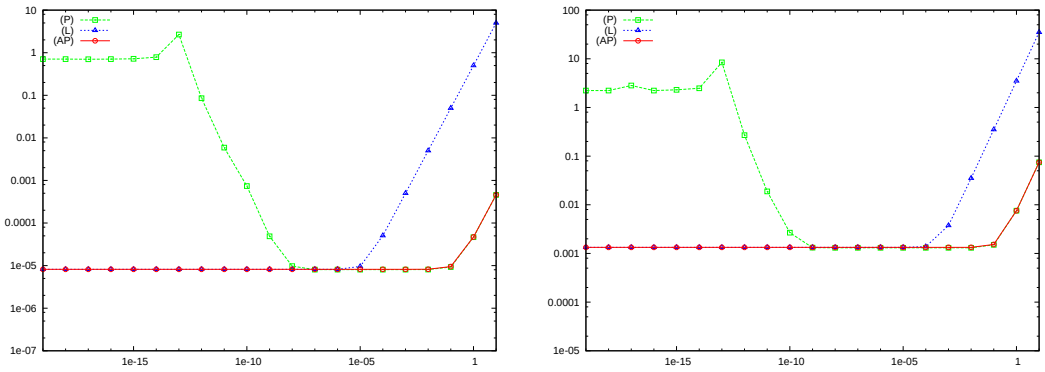
and the b field be aligned with the x-axis. Hence, the solution ϕ^ε of (2.2) and its decomposition $\phi^\varepsilon=p^\varepsilon+q^\varepsilon$ write

$$\phi^\varepsilon=\sin(\pi y)+\varepsilon\cos(2\pi x)\sin(\pi y), \tag{3.10}$$

$$p^\varepsilon=\sin(\pi y), \quad q^\varepsilon=\varepsilon\cos(2\pi x)\sin(\pi y). \tag{3.11}$$

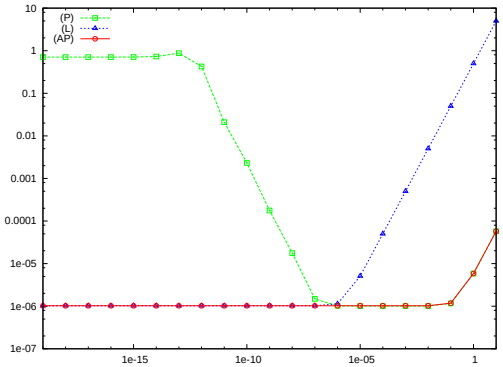
We denote by ϕ_P, ϕ_L, ϕ_A the numerical solution of the Singular Perturbation model (2.2), the Limit model (2.11) and the Asymptotic Preserving reformulation (2.31) respectively. The comparison will be done in the L^2-norm as well as the H^1-norm. The linear systems obtained after discretization of the three methods are solved using the same numerical algorithm — LU decomposition implemented in a solver MUMPS[2].

In Figure 3.1 we plotted the absolute errors (in the L^2 resp. H^1-norms) between the numerical solutions obtained with one of the three methods and the exact solution, and this, as a function of

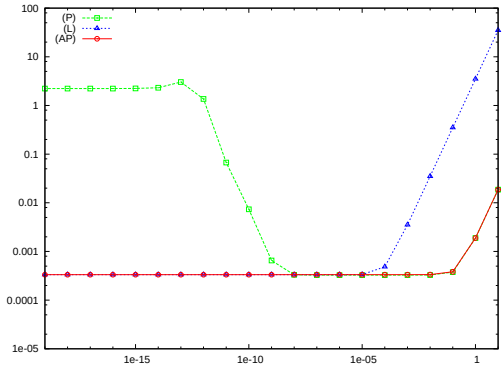


(a) L^2 error for a grid with 50×50 points.

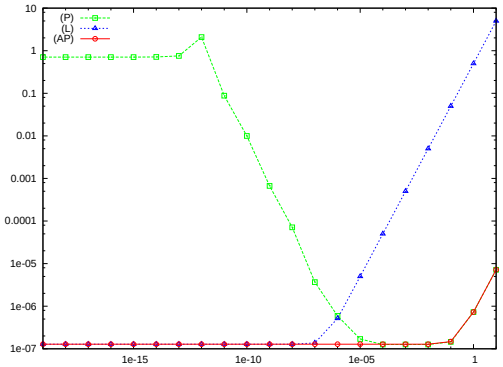
(b) H^1 error for a grid with 50×50 points.



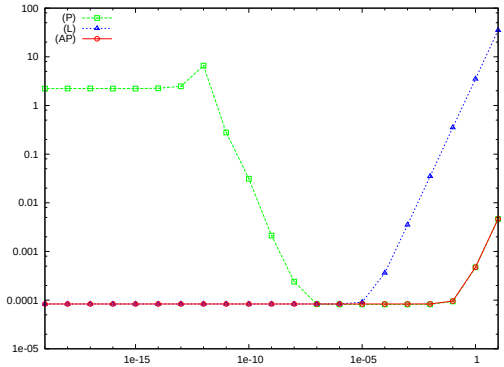
(c) L^2 error for a grid with 100×100 points.



(d) H^1 error for a grid with 100×100 points.



(e) L^2 error for a grid with 200×200 points.



(f) H^1 error for a grid with 200×200 points.

Figure 3.1 – Absolute L^2 (left column) and H^1 (right column) errors between the exact solution ϕ^ε and the computed numerical solution ϕ_A (AP), ϕ_L (L), ϕ_P (P) for the test case with constant b. The error is plotted as a function of the parameter ε and for three different mesh-sizes.

the parameter ε and for several mesh-sizes. In Table 3.1, we specified the error values for one fixed grid and several ε-values. One observes that the Singular Perturbation finite element approximation is accurate only for ε bigger than some critical value ε_P, the Limit model gives reliable results for ε

method	# rows	# non zero	time	L^2-error	H^1-error
AP	50×10^3	1563×10^3	13.212 s	1.02×10^{-6}	3.34×10^{-4}
L	20×10^3	469×10^3	5.227 s	1.14×10^{-6}	3.34×10^{-4}
P	10×10^3	157×10^3	3.707 s	1.02×10^{-6}	3.27×10^{-4}

Table 3.2 – Comparison between the Asymptotic Preserving scheme (AP), the Limit model (L) and the Singular Perturbation model (P) for $h=0.01$ (100 mesh points in each direction) and fixed $\varepsilon=10^{-6}$: matrix size, number of nonzero elements, average computational time and error in L^2 and H^1 norms.

smaller than ε_L, whereas the AP-scheme is accurate independently on ε. The order of convergence for all three methods is three in the L^2-norm and two in the H^1-norm, which is an optimal result for $\mathbb{Q}_2$ finite elements. When designing a robust numerical method one has therefore two options. The first one is to use an Asymptotic Preserving scheme, which is accurate independently on ε, but requires the solution of a bigger linear system. The second one is to design a coupling strategy that involves the solution of the Singular Perturbation formulation and the Limit problem in their respective validity domains. This is however a very delicate problem, since we observe that the critical values ε_P and ε_L are mesh dependent, namely ε_P inversely proportional to h and ε_L proportional to h. Therefore for small meshes there may exist a range of ε-values, where neither the Singular Perturbation nor the Limit model finite element approximation give accurate results. For our test case, this is even the case for meshes as big as 200×200 points, if one regards the L^2-norm. This mesh-size is generally insufficient in the case of real physical applications.

Another interesting aspect with respect to which the three methods must be compared, is the computational time and the size of the matrices involved in the linear systems. Table 3.2 shows that the Asymptotic Preserving scheme is expensive in computational time and memory requirements, as compared to the other methods. Indeed, the computational time required to solve the problem is almost four times bigger than that of the Singular Perturbation scheme. Moreover, the Asymptotic Preserving method involves matrices that have five times more rows and ten times more nonzero elements than the matrices obtained with the Singular Perturbation approximation. It is however the only scheme that provides the h-convergence regardless of ε. In order to reduce the computational costs, a coupling strategy for problems with variable ε will be proposed in a forthcoming paper. In sub-domains where $\varepsilon>\varepsilon_P$ the Singular Perturbation problem will be solved, in sub-domains where $\varepsilon<\varepsilon_L$ the Limit problem will be solved and only in the remaining part, where neither the Limit nor the Singular Perturbation model are valid, the Asymptotic Preserving formulation will be solved.

3.2.2. 2D test case, non-uniform and non-aligned b-field We now focus our attention on the original feature of the here introduced numerical method, namely its ability to treat nonuniform b fields. In this section we present numerical simulations performed for a variable field b.

First, let us construct a numerical test case. Finding an analytical solution for an arbitrary b presents a considerable difficulty. We have therefore chosen a different approach. First, we choose a limit solution

$$\phi^0=\sin\big(\pi y+\alpha(y^2-y)\cos(\pi x)\big)\,, \tag{3.12}$$

where α is a numerical constant aimed at controlling the variations of b. For $\alpha=0$, the limit solution of the previous section is obtained. The limit solution for $\alpha=2$ is shown in Figure 3.2. We set $\alpha=2$ in what follows. Since ϕ^0 is a limit solution, it is constant along the b field lines. Therefore we can

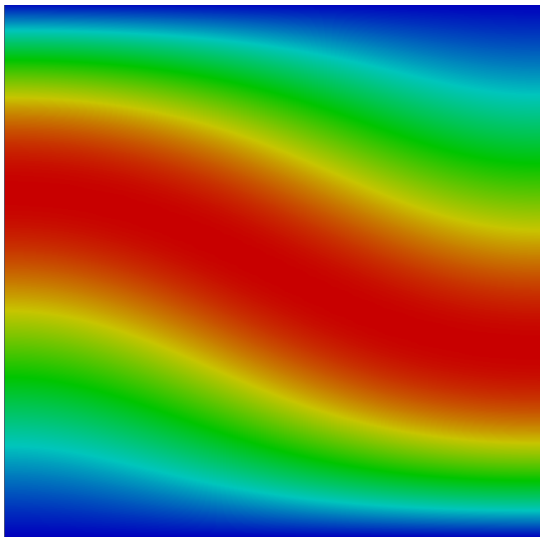


Figure 3.2 – The limit solution for the test case with variable b.

determine the b field using the following implication

$$\nabla_{\parallel}\phi^0 = 0 \quad \Rightarrow \quad b_x \frac{\partial \phi^0}{\partial x} + b_y \frac{\partial \phi^0}{\partial y} = 0, \tag{3.13}$$

which yields for example

$$b = \frac{B}{|B|}, \quad B = \begin{pmatrix} \alpha(2y-1)\cos(\pi x) + \pi \\ \pi\alpha(y^2 - y)\sin(\pi x) \end{pmatrix} \quad . \tag{3.14}$$

Note that the field B, constructed in this way, satisfies $\mathrm{div} B = 0$, which is an important property in the framework of plasma simulation. Furthermore, we have $B \neq 0$ in the computational domain. Now, we choose ϕ^ε to be a function that converges, as $\varepsilon \to 0$, to the limit solution ϕ^0:

$$\phi^\varepsilon = \sin\big(\pi y + \alpha(y^2 - y)\cos(\pi x)\big) + \varepsilon \cos(2\pi x)\sin(\pi y). \tag{3.15}$$

Finally, the force term is calculated, using the equation, i.e.

$$f = -\nabla_\perp \cdot (A_\perp \nabla_\perp \phi^\varepsilon) - \frac{1}{\varepsilon}\nabla_{\parallel} \cdot (A_{\parallel} \nabla_{\parallel} \phi^\varepsilon).$$

As in the previous section, we shall compare here the numerical solution of the Singular Perturbation model (2.2), the Limit model (2.11) and the Asymptotic Preserving reformulation (2.31), i.e. ϕ_P, ϕ_L, ϕ_A with the exact solution (3.15) . The L^2 and H^1-errors are reported on Figure 3.3 and Table 3.3. Once again the Asymptotic Preserving scheme proves to be valid for all values of ε, contrary to the other schemes. There is however a difference compared to the constant-b case. For a variable b , the threshold value ε_P seems to be independent on the mesh size and is much larger than that of the uniform b test case. This observation limits further the possible choice of coupling strategies, since even for coarse meshes there exists a range of ε-values, where neither the Singular Perturbation nor the Limit model are valid. The coupling strategy, involving all three models, remains however interesting to investigate.

ε	AP scheme		Limit model		Singular Perturbation scheme	
	L^2 error	H^1 error	L^2 error	H^1 error	L^2 error	H^1 error
10	7.2×10^{-6}	4.6×10^{-3}	5.0×10^{0}	3.50×10^{1}	7.2×10^{-6}	4.6×10^{-3}
1	7.1×10^{-7}	4.6×10^{-4}	5.0×10^{-1}	3.50×10^{0}	7.1×10^{-7}	4.6×10^{-4}
10^{-2}	2.05×10^{-7}	1.33×10^{-4}	5.0×10^{-3}	3.50×10^{-2}	2.05×10^{-7}	1.33×10^{-4}
10^{-4}	2.12×10^{-7}	1.38×10^{-4}	5.0×10^{-5}	3.77×10^{-4}	1.74×10^{-6}	1.43×10^{-4}
10^{-7}	2.17×10^{-7}	1.41×10^{-4}	2.22×10^{-7}	1.41×10^{-4}	1.68×10^{-3}	1.26×10^{-2}
10^{-10}	2.17×10^{-7}	1.41×10^{-4}	2.17×10^{-7}	1.41×10^{-4}	3.93×10^{-1}	1.35×10^{0}
10^{-15}	2.17×10^{-7}	1.41×10^{-4}	2.17×10^{-7}	1.41×10^{-4}	6.7×10^{-1}	2.32×10^{0}

Table 3.3 – Comparison between the Asymptotic preserving scheme, the Limit model and the Singular Perturbation model for $h=0.005$ (200 mesh points in each direction) and variable b: absolute L^2-error and H^1-error.

In the next test case we investigate the influence of the variations of the b field on the accuracy of the solution. We would like to answer the following question: what is the minimal number of points per characteristic length of b variations required to obtain an acceptable solution. For this, let us modify the previous test case. Let $b=B/|B|$, with

$$B=\begin{pmatrix}\alpha(2y-1)\cos(m\pi x)+\pi\\ m\pi\alpha(y^2-y)\sin(m\pi x)\end{pmatrix}, \tag{3.16}$$

m being an integer. The limit solution and ϕ^ε are chosen to be

$$\phi^0=\sin\big(\pi y+\alpha(y^2-y)\cos(m\pi x)\big), \tag{3.17}$$

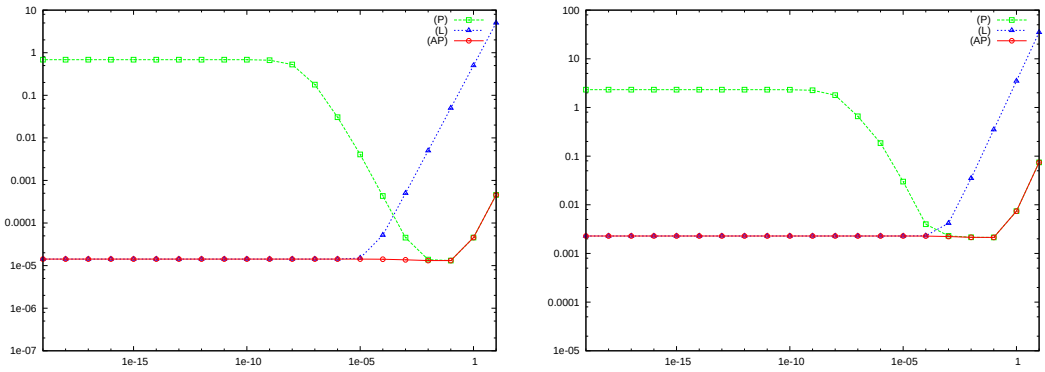
$$\phi^\varepsilon=\sin\big(\pi y+\alpha(y^2-y)\cos(m\pi x)\big)+\varepsilon\cos(2\pi x)\sin(\pi y). \tag{3.18}$$

We perform two tests: first, we fix the mesh size and vary m to find the minimal period of b for which the Asymptotic Preserving method yields still acceptable results. We define a result to be acceptable when the relative error is less then 0.01. In the second test m remains fixed and the convergence of the scheme is studied. The results are presented on Figures 3.4 and 3.5.

For $\varepsilon=1$ and 400 mesh points in each direction ($h=0.0025$) the relative error in the L^2-norm, defined as $\frac{||\phi^\varepsilon-\phi_A||_{L^2(\Omega)}}{||\phi_A||_{L^2(\Omega)}}$, is below 0.01 for all tested values of $1\leq m\leq 50$. The relative H^1-error $\frac{||\phi^\varepsilon-\phi_A||_{H^1(\Omega)}}{||\phi_A||_{H^1(\Omega)}}$ exceeds the critical value for $m>25$. For $\varepsilon=10^{-20}$ the maximal m for which the error is acceptable in both norms is 20. The minimal number of mesh points per period of b variations is 40 in the worst case, in order to obtain an 1% relative error.

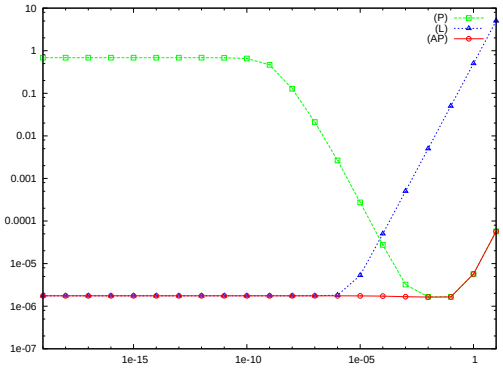
Figure 3.5 and Table 3.4 show the convergence of the Asymptotic Preserving scheme with respect to h for $m=10$ and $\varepsilon=10^{-10}$. We observe that for big values of h the error does not diminish with h. Then, for $h<0.025$ the scheme converges at a better rate then 2 for H^1-error and 3 for L^2-error. For $h<0.00625$ (160 points) the optimal convergence rate in the H^1-norm is obtained (which is 32 mesh points per period of b). The method is super-convergent in the whole tested range for the L^2-error.

These results are reassuring, as they prove that the Asymptotic Preserving scheme is precise even for strongly varying fields for relatively small mesh sizes, which was not evident. Indeed, the optimal convergence rate in the H^1-norm is obtained for 32 mesh points per b period, and an 1% relative error for 40 points. It shows that accurate results can be obtained in more complex simulations, such as tokamak plasma for example. The application of the method to bigger scale problems is the subject of an ongoing work.

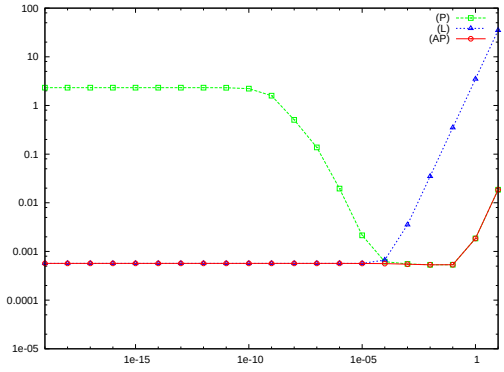


(a) L^2 error for a grid with 50×50 points.

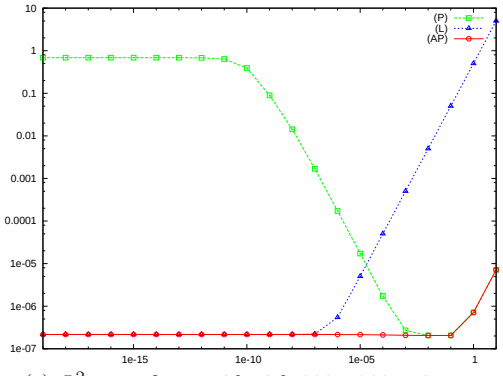
(b) H^1 error for a grid with 50×50 points.



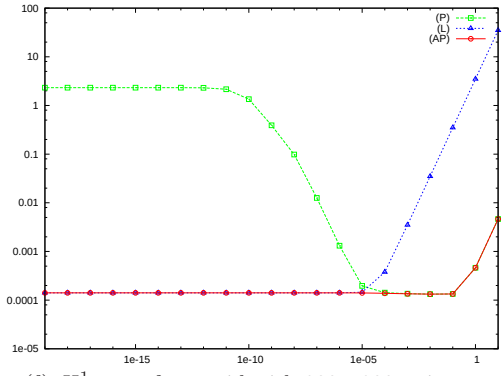
(c) L^2 error for a grid with 100×100 points.



(d) H^1 error for a grid with 100×100 points.

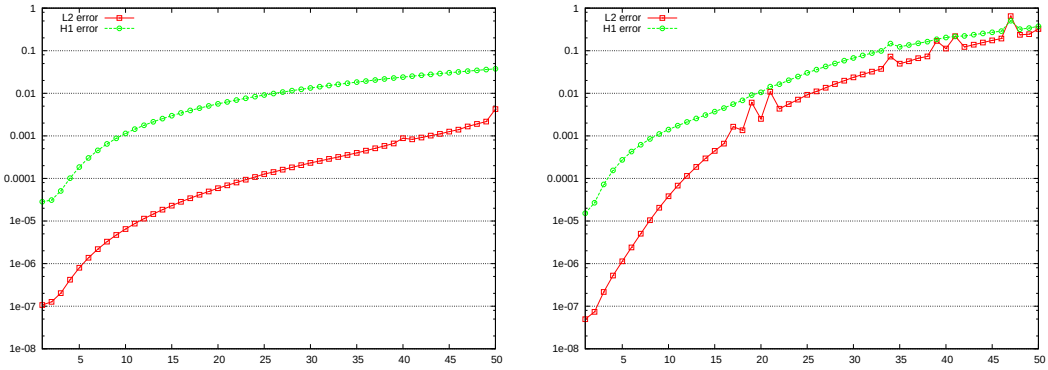


(e) L^2 error for a grid with 200×200 points.



(f) H^1 error for a grid with 200×200 points.

Figure 3.3 – Absolute L^2 (left column) and H^1 (right column) errors between the exact solution ϕ^ε and the computed solution ϕ_A (AP), ϕ_L (L), ϕ_P (P) for the test case with variable b. Plotted are the errors as a function of the small parameter ε, for three different meshes.



(a) $\varepsilon = 1$ (b) $\varepsilon = 10^{-20}$

Figure 3.4 – Relative L^2 and H^1 errors between the exact solution ϕ^ε and the computed solution ϕ_A (AP) for $h = 0.0025$ (400 points in each direction) as a function of m and for $\varepsilon = 1$ respectively 10^{-20}.

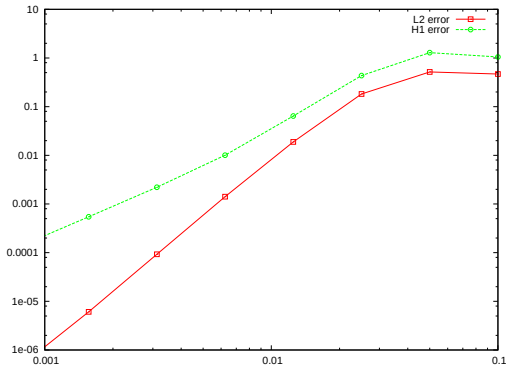


Figure 3.5 – Relative L^2 and H^1 errors between the exact solution ϕ^ε and the computed solution ϕ_A (AP) for $m = 10$ and $\varepsilon = 10^{-10}$ as a function of h.

3.2.3. 3D test case, uniform and aligned b-field Finally, we test our method on a simple $3D$ test case. Let the field b be aligned with the X-axis:

$$b = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}. \tag{3.19}$$

Let $\Omega = [0,1] \times [0,1] \times [0,1]$, and the source term f is such that the solution is given by

$$\phi^\varepsilon = \sin(\pi y)\sin(\pi z) + \varepsilon \cos(2\pi x)\sin(\pi y)\sin(\pi z), \tag{3.20}$$

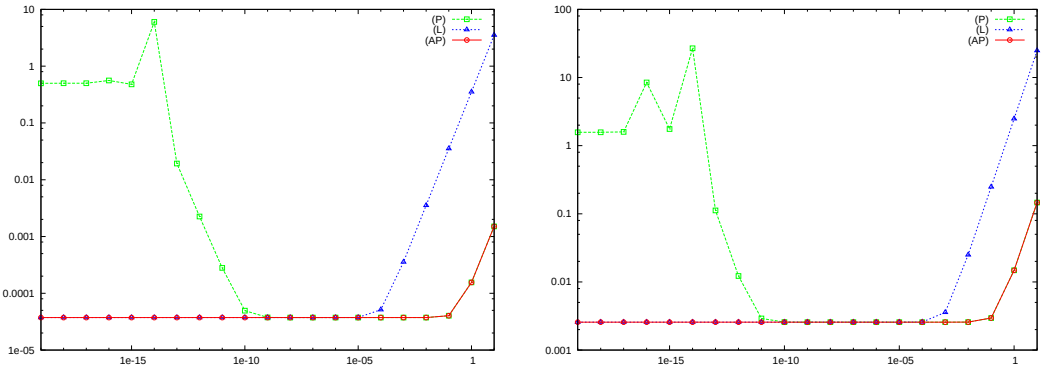
$$p^\varepsilon = \sin(\pi y)\sin(\pi z), \quad q^\varepsilon = \varepsilon \cos(2\pi x)\sin(\pi y)\sin(\pi z). \tag{3.21}$$

Numerical simulations were performed on a $30 \times 30 \times 30$ grid. Once again all three methods are compared. The L^2 and H^1-errors are given on Figure 3.6. The numerical results are equivalent with those obtained in the 2D test with constant b. Note that it is difficult to perform 3D simulations with more

h	# points per period	L^2-error	H^1-error
0.1	2	4.7×10^{-1}	1.05
0.05	4	5.2×10^{-1}	1.29
0.025	8	1.82×10^{-1}	4.3×10^{-1}
0.0125	16	1.89×10^{-2}	6.4×10^{-2}
0.00625	32	1.41×10^{-3}	1.00×10^{-2}
0.0003125	64	9.3×10^{-5}	2.21×10^{-3}
0.0015625	128	6.1×10^{-6}	5.5×10^{-4}
0.00078125	256	4.6×10^{-7}	1.36×10^{-4}

Table 3.4 – Relative L^2 and H^1 errors between the exact solution ϕ^ε and the computed solution ϕ_A (AP) for $m=10$ and $\varepsilon=10^{-10}$ as a function of h.

refined grids, due to memory requirements on standard desktop equipment as we are doing now. Every row in the matrix constructed for the Singular Perturbation model, can contain up to 125 non zero entries (for $\mathbb{Q}_2$ finite elements), while matrices associated with the Asymptotic Preserving reformulation have rows with up to 375 non zero entries. Furthermore the dimension of the latter is five times bigger. The memory requirements of the direct solver used in our simulations grow rapidly. The remedy could be to use an iterative solver with suitable preconditioner. Finding the most efficient method to inverse these matrices is however beyond the scope of this paper. In future work we will address this problem as well as a parallelization of this method.



(a) L^2 error for a grid with $30\times30\times30$ points. (b) H^1 error for a grid with $30\times30\times30$ points.

Figure 3.6 – Absolute L^2 (left column) and H^1 (right column) errors between the exact solution ϕ^ε and the computed solution ϕ_A (AP), ϕ_L (L), ϕ_P (P) for the 3D test case. The errors are plotted as a function of the anisotropy ratio ε.

4. Conclusions

The asymptotic preserving method presented in this paper is shown to be very efficient for the solution of highly anisotropic elliptic equations, where the anisotropy direction is given by an arbitrary, but smooth vector field b with non-adapted coordinates and meshes. The results presented here generalize the procedure used in [11] and have the important advantage to permit the use of Cartesian grids,

independently on the shape of the anisotropy. Moreover, the scheme is equally accurate, independently on the anisotropy strength, avoiding thus the use of coupling methods. The numerical study of this AP-scheme shall be investigated in a forthcoming paper, in particular the ε-independent convergence results shall be stated.

Another important related work consists in extending our methods to the case of anisotropy ratios ε, which are variable in Ω from moderate to very small values. This is important, for example, in plasma physics simulations as already noted in the introduction. An alternative strategy to the Asymptotic Preserving schemes whould be to couple a standard discretization in subregions with moderate ε with a limit ($\varepsilon \to 0$) model in subregions with small ε as suggested, for example, in [5, 25]. However, the limit model is only valid for $\varepsilon \ll 1$ and cannot be applied for weak anisotropies. Thus, the coupling strategy requires existence of a range of anisotropy strength where both methods are valid. This is rather undesirable since this range may not exist at all, as illustrated by our results in Fig. 3.1.

Appendix A. Decompositions $\mathcal{V} = \mathcal{G} \oplus^{\perp} \mathcal{A}$, $\tilde{\mathcal{V}} = \tilde{\mathcal{G}} \oplus \mathcal{L}$ and related estimates.

We shall show in this Appendix that all the statements in Hypotheses B and B' can be rigorously derived under some assumptions on the domain boundary $\partial\Omega$ and on the manner in which it is intersected by the field b. We assume essentially that b is tangential to $\partial\Omega$ on $\partial\Omega_D$ and that b penetrates the remaining part of the boundary $\partial\Omega_N$ at an angle that stays away from 0 on $\overline{\partial\Omega}_N$. We assume also that $\partial\Omega_N$ consists of two disjoint components for which there exist global and smooth parametrizations. This last assumption can be weakened (existence of an atlas of local smooth parametrizations should be sufficient) at the expense of lengthening the proofs. The precise set of our assumptions is the following:

Hypothesis C The boundary of Ω is the union of three components: $\partial\Omega_D$ where $b \cdot n = 0$, $\partial\Omega_{in}$ where $b \cdot n \leq -\alpha$ and $\partial\Omega_{out}$ where $b \cdot n \geq \alpha$ with some constant $\alpha > 0$. Moreover, there is a smooth system of coordinates $\xi_1, \ldots, \xi_{d-1}$ on $\partial\Omega_{in}$ meaning that there is a bounded domain $\Gamma_{in} \in \mathbb{R}^{d-1}$ and a one-to-one map $h_{in} : \Gamma_{in} \to \mathbb{R}^d$ such that $h_{in} \in C^2(\overline{\Gamma}_{in})$ and $\partial\Omega_{in}$ is the image of $h_{in}(\xi_1, \ldots, \xi_{d-1})$ as $(\xi_1, \ldots, \xi_{d-1})$ goes over Γ_{in}. The matrix formed by the vectors $(\partial h_{in}/\partial\xi_1, \ldots, \partial h_{in}/\partial\xi_1, n)$ is invertible for all $(\xi_1, \ldots, \xi_{d-1}) \in \overline{\Gamma}_{in}$. Similar assumptions hold also for $\partial\Omega_{out}$ (changing Γ_{in} to Γ_{out} and h_{in} to h_{out}).

Using this hypothesis we can introduce a system of coordinates in Ω such that the field lines of b coincide with the coordinate lines. To do this consider the initial value problem for a parametrized ordinary differential equation (ODE):

$$\frac{\partial X}{\partial \xi_d}(\xi', \xi_d) = b(X(\xi', \xi_d)), \; X(\xi', 0) = h_{in}(\xi'). \tag{1.1}$$

Here $X(\xi', \xi_d)$ is $\mathbb{R}^d$-valued and ξ' stands for $(\xi_1, \ldots, \xi_{d-1})$. For any fixed $\xi' \in \Gamma_{in}$, equation (1.1) should be understood as an ODE for a function of ξ_d. Its solution $X(\xi', \xi_d)$ goes then over the field line of b starting (as $\xi_d = 0$) at the point on the inflow boundary $\partial\Omega_{in}$, parametrized by ξ'. This field line hits the outflow boundary $\partial\Omega_{out}$ somewhere. In other words, for any $\xi' \in \Gamma_{in}$ there exists $L(\xi') > 0$ such that $X(\xi', L(\xi')) \in \partial\Omega_{out}$. The domain of definition of X is thus

$$D = \{(\xi', \xi_d) \in \mathbb{R}^d \; / \; \xi' \in \Gamma_{in} \text{ and } 0 < \xi_d < L(\xi')\}.$$

Gathering the results on parametrized ODEs, from for instance [34], we conclude that $X(\xi', \xi_d) = X(\xi_1, \ldots, \xi_d)$ is a smooth function of all its d parameters, more precisely $X \in C^2(\overline{D})$. Evidently, the map X is one-to-one from $\overline{D}$ to $\overline{\Omega}$ and thus $\xi_1, \ldots, \xi_d$ provide a system of coordinates for $\overline{\Omega}$. Moreover this system is not degenerate in the sense that the vectors $\partial X/\partial\xi_1, \ldots, \partial X/\partial\xi_d$ are linearly independent at each point of $\overline{\Omega}$. Indeed, if this was not the case, then there would exist a non trivial linear combination

$\lambda_1\partial X/\partial\xi+\cdots+\lambda_d\partial X/\partial\xi_d$ that would vanish at some point in $\overline{\Omega}$. But, ODE (1.1) implies

$$\frac{\partial}{\partial\xi_d}\sum_{i=1}^{d}\lambda_i\frac{\partial X}{\partial\xi_i}(\xi',\xi_d)=\nabla b(X(\xi',\xi_d))\cdot\sum_{i=1}^{d}\lambda_i\frac{\partial X}{\partial\xi_i}(\xi',\xi_d)$$

so that, the unique solution of this ODE, i.e. the linear combination $\sum_{i=1}^{d}\lambda_i\frac{\partial X}{\partial\xi_i}$, would vanish on the whole field line, in particular on the inflow. But this is impossible since $\frac{\partial X}{\partial\xi_i}=\frac{\partial h_{in}}{\partial\xi_i}$, $i=1,\ldots,d-1$ on the inflow, while $\frac{\partial X}{\partial\xi_d}=b$ and the vectors $\left(\frac{\partial h_{in}}{\partial\xi_1},\ldots,\frac{\partial h_{in}}{\partial\xi_{d-1}},b\right)$ are linearly independent for all $(\xi_1,\ldots,\xi_{d-1})\in\overline{\Gamma}_{in}$. We see thus that the Jacobian $J=\det(\partial X_j/\partial\xi_i)$ does not vanish on $\overline{\Omega}$ so that we can assume that $m<J<M$ everywhere on $\overline{\Omega}$ with some positive constants m and M (assuming that J is positive does not prevent the generality since if J is negative in Ω than one can replace ξ_1 by $-\xi_1$). Since $X\in C^2(\overline{\Omega})$, we have also that $J\in C^1(\overline{\Omega})$.

One also sees easily that the top of D is given by a smooth function $L(\xi')$. Indeed, $L(\xi')$ is determined for each $\xi'\in\Gamma_{in}$ from the equation $X(\xi',L(\xi'))=h_{out}(\eta)$ with some $\eta=(\eta_1,\ldots,\eta_{d-1})\in\Gamma_{out}$. We know already that this equation is solvable for $\xi_d=L(\xi')$, $\eta_1,\ldots,\eta_{d-1}$ for any $\xi'\in\Gamma_{in}$. To conclude that the solution depends smoothly on ξ' we can apply the implicit function theorem to the equation

$$F(\xi';\xi_d,\eta_1,\ldots,\eta_{d-1})=X(\xi',\xi_d)-h_{out}(\eta_1,\ldots,\eta_{d-1})=0.$$

Indeed, the R^d-valued function F is smooth and the matrix of its partial derivatives with respect to $\xi_d,\eta_1,\ldots,\eta_{d-1}$ is invertible since $\partial F/\partial\xi_d=b$ and $\partial F/\partial\eta_i=-\partial h_{out}/\partial\eta_i$ at some point at the outflow and the vectors $\partial h_{out}/\partial\eta_i$ lie in the tangent plane to $\partial\Omega_{out}$ while b is nowhere in this plane. We have moreover that $L\in C^1(\overline{\Gamma}_{in})$. Indeed, we can prove that all the derivatives of L are bounded. In order to do it, let us remark that the differential of $X(\xi',L(\xi'))$ represents a vector in the tangent plane at some point on $\partial\Omega_{out}$ so that it is perpendicular to the outward normal n. We have thus for any $i=1,\ldots,d-1$

$$0=n\cdot\left(\frac{\partial X}{\partial\xi_i}(\xi',L(\xi'))+\frac{\partial X}{\partial\xi_d}(\xi',L(\xi'))\frac{\partial L}{\partial\xi_i}(\xi')\right)=n\cdot\left(\frac{\partial X}{\partial\xi_i}(\xi',L(\xi'))+b(\xi',L(\xi'))\frac{\partial L}{\partial\xi_i}(\xi')\right)$$

so that

$$\frac{\partial L}{\partial\xi_i}(\xi')=-\frac{n\cdot\frac{\partial X}{\partial\xi_i}(\xi',L(\xi'))}{n\cdot b(\xi',L(\xi'))}$$

and this is bounded since X has bounded partial derivatives and $n\cdot b\geq\alpha$ by the hypothesis. Note also that L is strictly positive.

- We can now establish the decomposition $\mathcal{V}=\mathcal{G}\oplus^{\perp}\mathcal{A}$. Take any $\phi\in\mathcal{V}\cap C^1(\overline{\Omega})$ and introduce $p\in L^2(\Omega)$ by

$$p(x)=p(\xi',\xi_d)=p(\xi')=\frac{\int_0^{L(\xi')}\phi(\xi',t)J(\xi',t)dt}{\int_0^{L(\xi')}J(\xi',t)dt}. \tag{1.2}$$

(from now on we switch back and forth between the Cartesian coordinates $x=(x_1,\ldots,x_d)$ and the new ones $(\xi',\xi_d)=(\xi_1,\ldots,\xi_d)$). Evidently, p is constant along each field line. Moreover, p is the L^2-orthogonal projection of ϕ on the space of such functions. Indeed, if $\psi=\psi(\xi')\in L^2(\Omega)$ is any function constant along each field line then

$$\begin{aligned}\int_\Omega p\psi dx&=\int_D p\psi J d\xi=\int_{\Gamma_{in}}p(\xi')\psi(\xi')\int_0^{L(\xi')}J(\xi',\xi_d)d\xi_d d\xi'\\&=\int_{\Gamma_{in}}\int_0^{L(\xi')}\phi(\xi',\xi_d)\,\psi(\xi')\,J(\xi',\xi_d)d\xi_d d\xi'=\int_\Omega\phi\psi dx.\end{aligned}$$

Let us prove that $p\in\mathcal{V}$, i.e. that its derivatives are square integrable. The change of variable $t=L(\xi')s$ yields the function

$$p(\xi')=\frac{\int_0^1\phi(\xi',L(\xi')s)J(\xi',L(\xi')s)ds}{\int_0^1 J(\xi',L(\xi')s)ds}.$$

Now we have $\partial p/\partial\xi_d=0$ and for all $\partial p/\partial\xi_{i,i=1,\dots,d-1}$ denoting $a=a(\xi')=(\int_0^1 J(\xi',L(\xi')s)ds)^{-1}$, $\phi=\phi(\xi',L(\xi')s)$ and same for J we obtain

$$\begin{aligned}\frac{\partial p}{\partial\xi_i} &= \frac{\partial a}{\partial\xi_i}\int_0^1\phi J ds+a\int_0^1\frac{\partial\phi}{\partial\xi_i}Jds+a\int_0^1\frac{\partial\phi}{\partial\xi_d}\frac{\partial L}{\partial\xi_i}sJds\\ &\quad +a\int_0^1\phi\frac{\partial J}{\partial\xi_i}ds+a\int_0^1\phi\frac{\partial J}{\partial\xi_d}\frac{\partial L}{\partial\xi_i}sds\end{aligned}\tag{1.3}$$

Using all the previous bounds on the functions L and J and skipping the details of somewhat tedious calculations, we arrive at

$$\begin{aligned}\int_\Omega\left(\frac{\partial p}{\partial\xi_i}\right)^2dx &= \int_{\Gamma_{in}}\int_0^{L(\xi')}\left(\frac{\partial p}{\partial\xi_i}\right)^2Jd\xi_d d\xi'\\ &\le C\int_{\Gamma_{in}}\int_0^{L(\xi')}\left(\phi^2+\left(\frac{\partial\phi}{\partial\xi_i}\right)^2+\left(\frac{\partial\phi}{\partial\xi_d}\right)^2\right)Jd\xi_d d\xi'\end{aligned}$$

implying

$$\left\|\frac{\partial p}{\partial\xi_i}\right\|^2_{L^2(\Omega)}\le C\left(\|\phi\|^2_{L^2(\Omega)}+\left\|\frac{\partial\phi}{\partial\xi_i}\right\|^2_{L^2(\Omega)}+\left\|\frac{\partial\phi}{\partial\xi_d}\right\|^2_{L^2(\Omega)}\right)\le C\|\phi\|^2_{H^1(\Omega)}.$$

Thus $p\in H^1(\Omega)$, hence $p\in\mathcal{G}$ and $q=\phi-p\in\mathcal{A}$. Since the dependence of p on ϕ is continuous in the norm of $H^1(\Omega)$, a density argument shows that the decomposition $\phi=p+q$ with $p\in\mathcal{G}$ and $q\in\mathcal{A}$ exists for any $\phi\in\mathcal{V}$.

- Let us now introduce the operator P as the L^2-orthogonal projector on $\mathcal{G}$, that means

$$P:\mathcal{V}\to\mathcal{G},\quad \phi\in\mathcal{V}\longmapsto P\phi\in\mathcal{G}\quad\text{given by}\quad(1.2).$$

Then, the estimates in the preceding paragraph show that the operator P is continuous in the norm of $H^1(\Omega)$:

$$||\nabla_\perp(P\phi)||_{L^2(\Omega)}\le C||\nabla\phi||_{L^2(\Omega)},\quad\forall\phi\in\mathcal{V}\tag{1.4}$$

- We have also the following Poincaré-Wirtinger inequality:

$$||\phi-P\phi||_{L^2(\Omega)}\le C||\nabla_{||}\phi||_{L^2(\Omega)},\quad\forall\phi\in\mathcal{V}.\tag{1.5}$$

To prove this, it is sufficient to establish that $||q||_{L^2(\Omega)}\le C||\nabla_{||}q||_{L^2(\Omega)}$ for all $q\in\mathcal{A}$. We observe that

$$||q||^2_{L^2(\Omega)}=\int_{\Gamma_{in}}\int_0^{L(\xi')}q^2(\xi',\xi_d)J(\xi',\xi_d)d\xi_d d\xi'$$

and

$$||\nabla_{||}\phi||^2_{L^2(\Omega)}=\int_{\Gamma_{in}}\int_0^{L(\xi')}\left(\frac{\partial q}{\partial\xi_d}\right)^2(\xi',\xi_d)J(\xi',\xi_d)d\xi_d d\xi'.$$

The requirement $q \in \mathcal{A}$ is equivalent to

$$\int_0^{L(\xi')} q(\xi',\xi_d) J(\xi',\xi_d) d\xi_d = 0 \quad \text{f.a.a. } \xi' \in \Gamma_{in}\,. \tag{1.6}$$

We have thus to prove for every ξ'

$$\int_0^{L(\xi')} q^2(\xi',\xi_d) J(\xi',\xi_d) d\xi_d \le C^2 \int_0^{L(\xi')} \left(\frac{\partial q}{\partial \xi_d}\right)^2 (\xi',\xi_d) J(\xi',\xi_d) d\xi_d$$

provided (1.6). Fixing any ξ', making the change of integration variable $\xi_d = L(\xi')t$ and introducing the functions $u(t) = q(\xi', L(\xi')t) J(\xi', L(\xi')t)$ and $J(t) = J(\xi', L(\xi')t)$, we rewrite the last inequality as

$$\int_0^1 \frac{u^2(t)}{J(t)} dt \le \frac{C^2}{L^2(\xi')} \int_0^1 \left(\frac{u'(t)}{J(t)} - \frac{u(t)}{J^2(t)} J'(t)\right)^2 J(t) dt. \tag{1.7}$$

Since $\int_0^1 u(t)dt = 0$ we have by the standard Poincaré inequality

$$\int_0^1 u^2(t) dt \le C_P^2 \int_0^1 (u'(t))^2 dt. \tag{1.8}$$

- Let us turn to the verification of Hypothesis B'. Take any $u \in \tilde{\mathcal{V}}$. We want to prove that one can decompose $u = p + q$ with $p \in \tilde{\mathcal{G}}$ and $q \in \mathcal{L}$ and the trace of u on $\partial\Omega_{in}$ (denoted g) is in $L^2(\partial\Omega_{in})$. In the ξ-coordinates we can write a surface element of $\partial\Omega_{in}$ as $d\sigma = S(\xi')d\xi'$ with a function S smoothly depending on ξ'. We see now that for u suffuciently smooth

$$\begin{aligned} ||g||^2_{L^2(\partial\Omega_{in})} &= \int_{\Gamma_{in}} g^2(\xi') S(\xi') d\xi' \\ &\le C \int_0^1 \int_{\Gamma_{in}} \left[u^2(\xi', L(\xi')s) + \frac{1}{L(\xi')} \left(\frac{\partial u}{\partial \xi_d}\right)^2 (\xi', L(\xi')s) \right] S(\xi') d\xi' ds \\ &\qquad \text{(by a one-dimensional trace inequlity)} \\ &\le C ||u||^2_{\tilde{\mathcal{V}}} \end{aligned}$$

By density, the trace g is thus defined for any $u \in \tilde{\mathcal{V}}$ with $||g||_{L^2(\partial\Omega_{in})} \le C||u||_{\tilde{\mathcal{V}}}$. Taking $p = p(\xi') = g(\xi')$ we observe by a similar calculation that $||p||_{L^2(\Omega)} \le C||u||_{\tilde{\mathcal{V}}}$ so that $p \in \tilde{\mathcal{G}}$. By definition $q = u - p \in \mathcal{L}$.

Appendix B. On the choice of the finite element space $\mathcal{L}_h$.

Let Ω be the rectangle $(0, L_x) \times (0, L_y)$ and the anisotropy direction be constant and aligned with the y-axis: $b = (0,1)$. Let us use the $\mathbb{Q}_k$ finite elements on a Cartesian grid, i.e. take some basis function $\theta_{x_i}(x)$, $i = 0, \ldots, N_x$ and $\theta_{y_j}(y)$, $j = 0, \ldots, N_y$ and define the complete finite element space X_h (without any restrictions on the boundary) as $\text{span}\{\theta_{x_i}(x)\theta_{y_j}(y)\ 0 \le i \le N_x,\ 0 \le j \le N_y\}$. The following subspace is then used for the approximation of the unknowns $p, q, l \in \mathcal{V}$

$$\mathcal{V}_h = \{v_h \in X_h / v_h|_{\partial\Omega_D} = 0\}.$$

We want to prove that taking for the approximation of $\lambda, \mu \in \mathcal{L}$ the space $\mathcal{L}_h$ under the form

$$\mathcal{L}_h = \{\lambda_h \in X_h / \lambda_h|_{\partial\Omega_{in}} = 0\}, \tag{2.1}$$

leads to an ill posed problem (3.1).

Claim There exist $\lambda_h \in \mathcal{L}_h$, $\lambda_h \neq 0$, such that $a_{||}(\lambda_h, p_h) = 0$ for all $p_h \in \mathcal{V}_h$. In fact there are exactly $2N_y$ linearly independent functions having this property.

REMARK B.1. *In the continuous case, the equation*

$$a_{||}(p, \lambda) = 0, \quad \forall p \in \mathcal{V},$$

implies $\lambda = 0$ *by density arguments. These density arguments are lost when discretizing the spaces* $\mathcal{V}$ *resp.* $\mathcal{L}$.

Proof of the Claim. We can suppose that the basis functions $\theta_{ij}(x,y) := \theta_{x_i}(x)\theta_{y_j}(y)$ are enumerated so that $\theta_{ij}(0,y) = 0$ for all $i \geq 1$ and $\theta_{0j}(0,y) \neq 0$. Hence for all $p_h = \sum p_{ij}\theta_{ij} \in \mathcal{V}_h$, the coefficients satisfy $p_{0j} = 0$ since the part of the boundary $\{x = 0\}$ is in $\partial\Omega_D$. Let $M = (m_{ik})_{0 \leq i,k \leq N_x}$ be the mass matrix in the x-direction: $m_{ik} = \int \theta_{x_i}(x)\theta_{x_k}(x)dx$. This matrix is invertible, hence there is a vector $a \in \mathbb{R}^{N_x+1}$ that solves $Ma = e$ with $e \in \mathbb{R}^{N_x+1}$, $e = (1,0,\ldots,0)^t$. Take any fixed integer j, $1 \leq j \leq N_y$ and define $\lambda_h \in \mathcal{L}_h$ as $\lambda_h = \sum a_i \theta_{ij}$. Then, for all $p_h = \sum p_{kl}\theta_{kl} \in \mathcal{V}_h$ we have

$$\begin{aligned} a_{||}(\lambda_h, p_h) &= \sum_{i,k,l} a_i p_{kl} \int_\Omega \frac{\partial \theta_{ij}}{\partial y} \frac{\partial \theta_{kl}}{\partial y} dxdy \\ &= \sum_{i,k,l} a_i p_{kl} \int_0^{L_x} \theta_{x_i}(x)\theta_{x_k}(x)dx \int_0^{L_y} \theta'_{y_j}(y)\theta'_{y_l}(y)dy \\ &= \sum_{k,l} \delta_{k0} p_{kl} \int_0^{L_y} \theta'_{y_j}(y)\theta'_{y_l}(y)dy = 0. \end{aligned}$$

As we can do this for all (i,j), $i = 0$, $1 \leq j \leq N_y$ and in the same manner for all (i,j), $i = N_x$, $1 \leq j \leq N_y$, there are $2N_y$ linearly independent functions with the property $a_{||}(\lambda_h, p_h) = 0$ for all $p_h \in \mathcal{V}_h$.

We see now that the system (3.1) with zero right hand side $f = 0$ possesses non-zero solutions $(p_h^\varepsilon, \lambda_h^\varepsilon, q_h^\varepsilon, l_h^\varepsilon, \mu_h^\varepsilon) = (0, \lambda_h^j, 0, 0, 0)$ where λ_h^j is any of the functions constructed in the preceding paragraph. It means that (3.1) is ill posed, i.e. the corresponding matrix is singular.

Acknowledgement. This work has been supported by the Marie Curie Actions of the European Commission under the contract DEASE (MEST-CT-2005-021122), by the French 'Commissariat à l'Energie Atomique (CEA)' under contracts ELMAG (CEA-Cesta 4600156289), and GYRO-AP (Euratom-CEA V 3629.001), by the Agence Nationale de la Recherche (ANR) under contract IODISEE (ANR-09-COSI-007-02), by the 'Fondation Sciences et Technologies pour l'Aéronautique et l'Espace (STAE)' under contract PLASMAX (RTRA-STAE/2007/PF/002) and by the Scientific Council of the Université Paul Sabatier, under contract MOSITER. Support from the French magnetic fusion programme 'fédération de recherche sur la fusion par confinement magnétique' is also acknowledged. The authors would like to express their gratitude to G. Gallice and C. Tessieras from CEA-Cesta for bringing their attention to this problem, to G. Falchetto, X. Garbet and M. Ottaviani from CEA-Cadarache, for their constant support to this research programme.

REFERENCES

[1] J. Adam, J. Boeuf, N. Dubuit, M. Dudeck, L. Garrigues, D. Gresillon, A. Heron, G. Hagelaar, V. Kulaev, N. Lemoine, et al. Physics, simulation and diagnostics of Hall effect thrusters. *Plasma Physics and Controlled Fusion*, 50:124041, 2008.

[2] P. R. Amestoy, I. S. Duff, J.-Y. L'Excellent, and J. Koster. A fully asynchronous multifrontal solver using distributed dynamic scheduling. *SIAM J. Matrix Anal. Appl.*, 23(1):15–41, 2001.

[3] S. F. Ashby, W. J. Bosl, R. D. Falgout, S. G. Smith, A. F. Tompson, and T. J. Williams. A Numerical Simulation of Groundwater Flow and Contaminant Transport on the CRAY T3D and C90 Supercomputers. *International Journal of High Performance Computing Applications*, 13(1):80–93, 1999.
[4] M. Beer, S. Cowley, and G. Hammett. Field-aligned coordinates for nonlinear simulations of tokamak turbulence. *Physics of Plasmas*, 2(7):2687, 1995.
[5] C. Besse, P. Degond, F. Deluzet, J. Claudel, G. Gallice, and C. Tessieras. A model hierarchy for ionospheric plasma modeling. *Math. Models Methods Appl. Sci.*, 14(3):393–415, 2004.
[6] C. Besse, F. Deluzet, C. Negulescu, and C. Yang. Three dimensional simulation of ionsphoric plasma disturbences. in preparation.
[7] A. H. Boozer. Establishment of magnetic coordinates for a given magnetic field. *Physics of Fluids*, 25(3):520–521, 1982.
[8] S. Brull, P. Degond, and F. Deluzet. Numerical degenerate elliptic problems and their applications to magnetized plasma simulations. in preparation.
[9] P. Crispel, P. Degond, and M.-H. Vignal. An asymptotic preserving scheme for the two-fluid Euler-Poisson model in the quasineutral limit. *J. Comput. Phys.*, 223(1):208–234, 2007.
[10] P. Degond, F. Deluzet, L. Navoret, A.-B. Sun, and M.-H. Vignal. Asymptotic-preserving particle-in-cell method for the vlasov-poisson system near quasineutrality. *J. Comput. Phys.*, 229(16):5630–5652, 2010.
[11] P. Degond, F. Deluzet, and C. Negulescu. An asymptotic preserving scheme for strongly anisotropic elliptic problems. *Multiscale Model. Simul.*, 8(2):645–666, 2009/10.
[12] P. Degond, F. Deluzet, A. Sangam, and M.-H. Vignal. An asymptotic preserving scheme for the Euler equations in a strong magnetic field. *J. Comput. Phys.*, 228(10):3540–3558, 2009.
[13] F. Deluzet, A. Lozinski, J. Narski, and C. Negulescu. Error estimates for an asymptotic-preserving method for highly anisotropic diffusion equations. in preparation.
[14] W. D. D'haeseleer, W. N. G. Hitchon, J. D. Callen, and J. L. Shohet. *Flux coordinates and magnetic field structure. A guide to a fundamental tool of plasma theory.* Springer Series in Computational Physics. Springer-Verlag, Berlin, 1991.
[15] A. M. Dimits. Fluid simulations of tokamak turbulence in quasiballooning coordinates. *Phys. Rev. E*, 48(5):4070–4079, Nov 1993.
[16] M. W. Gee, J. J. Hu, and R. S. Tuminaro. A new smoothed aggregation multigrid method for anisotropic problems. *Numer. Linear Algebra Appl.*, 16(1):19–37, 2009.
[17] L. Giraud and R. S. Tuminaro. Schur complement preconditioners for anisotropic problems. *IMA J. Numer. Anal.*, 19(1):1–18, 1999.
[18] V. Grandgirard, Y. Sarazin, X. Garbet, G. Dif-Pradalier, P. Ghendrih, N. Crouseilles, G. Latu, E. Sonnendrücker, N. Besse, and P. Bertrand. Computing ITG turbulence with a full-f semi-lagrangian code. *Communications in Nonlinear Science and Numerical Simulation*, 13(1):81 – 87, 2008.
[19] S. Hamada. Hydromagnetic equilibria and their proper coordinates. *Nucl. Fusion*, 2:23–37, 1962.
[20] T. Y. Hou and X.-H. Wu. A multiscale finite element method for elliptic problems in composite materials and porous media. *J. Comput. Phys.*, 134(1):169–189, 1997.
[21] Y. Igitkhanov, M. Mikhailov, and W. Feneberg. Fluid Description of Edge Plasma Transport in a Non-Orthogonal Coordinate System. *Contributions to Plasma Physics*, 34(2-3):398–403, 1994.
[22] S. Jin. Efficient asymptotic-preserving (AP) schemes for some multiscale kinetic equations. *SIAM J. Sci. Comput.*, 21(2):441–454, 1999.
[23] E. Kaveeva and V. Rozhansky. Poloidal and toroidal flows in tokamak plasma near magnetic islands. *Technical physics letters*, 30(7):538–540, 2004.
[24] M. Kelley, W. Swartz, and J. Makela. Mid-latitude ionospheric fluctuation spectra due to secondary EÕB instabilities. *Journal of Atmospheric and Solar-Terrestrial Physics*, 66(17):1559–1565, 2004.
[25] M. Keskinen. Nonlinear theory of the E× B instability with an inhomogeneous electric field. *Journal of Geophysical Research*, 89(A6):3913–3920, 1984.
[26] M. Keskinen, S. Ossakow, and B. Fejer. Three-dimensional nonlinear evolution of equatorial ionospheric spread-F bubbles. *Geophys. Res. Lett*, 30(16):4–1–4–4, 2003.
[27] B. N. Khoromskij and G. Wittum. Robust Schur complement method for strongly anisotropic elliptic equations. *Numer. Linear Algebra Appl.*, 6(8):621–653, 1999.
[28] I. Llorente and N. Melson. Robust multigrid smoothers for three dimensional elliptic equations with strong anisotropies. ICASE Technical Report: TR-98-37, 1998.
[29] T. Manku and A. Nathan. Electrical properties of silicon under nonuniform stress. *Journal of Applied Physics*, 74(3):1832–1837, 1993.
[30] J. M. Melenk. *hp-finite element methods for singular perturbations*, volume 1796 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2002.
[31] K. Miyamoto. *Controlled fusion and plasma physics*. Chapman & Hall, 2007.
[32] Y. Notay. An aggregation-based algebraic multigrid method. Report GANMN 08-02, Université Libre de Bruxelles, Brussels, Belgium, 2008.
[33] M. A. Ottaviani. An alternative approach to field-aligned coordinates for plasma turbulence simulations.

arXiv:1002.0748, 2010.
[34] L. S. Pontryagin. *Ordinary differential equations*. Addison-Wesley Publishing Co., Inc., Reading, Mass.-Palo Alto, Calif.-London, 1962.
[35] D. Stern. Geomagnetic Euler potentials. *J. geophys. Res*, 72(15):3995–4005, 1967.
[36] A. M. Tréguier. Modélisation numérique pour l'océanographie physique. *Ann. Math. Blaise Pascal*, 9(2):345–361, 2002.
[37] W.-W. Wang and X.-C. Feng. Anisotropic diffusion with nonlinear structure tensor. *Multiscale Model. Simul.*, 7(2):963–977, 2008.
[38] J. Weickert. *Anisotropic diffusion in image processing*. European Consortium for Mathematics in Industry. B. G. Teubner, Stuttgart, 1998.

AN ANISOTROPIC ERROR ESTIMATOR FOR THE CRANK-NICOLSON METHOD : APPLICATION TO A PARABOLIC PROBLEM

par

ALEXEI LOZINSKI, MARCO PICASSO AND VIRABOUTH PRACHITTHAM

paru dans

SIAM Journal of Scientific Computing

SIAM J. SCI. COMPUT.
Vol. 31, No. 4, pp. 2757–2783
© 2009 Society for Industrial and Applied Mathematics

AN ANISOTROPIC ERROR ESTIMATOR FOR THE CRANK–NICOLSON METHOD: APPLICATION TO A PARABOLIC PROBLEM*

ALEXEI LOZINSKI†, MARCO PICASSO‡, AND VIRABOUTH PRACHITTHAM‡

Abstract. In this paper we derive two a posteriori upper bounds for the heat equation. A continuous, piecewise linear finite element discretization in space and the Crank–Nicolson method for the time discretization are used. The error due to the space discretization is derived using anisotropic interpolation estimates and a postprocessing procedure. The error due to the time discretization is obtained using two different continuous, piecewise quadratic time reconstructions. The first reconstruction is developed following G. Akrivis, C. Makridakis, and R. H. Nochetto [*Math. Comp.*, 75 (2006), pp. 511–531], while the second one is new. Moreover, in the case of isotropic meshes only, upper and lower bounds are provided as in [R. Verfürth, *Calcolo*, 40 (2003), pp. 195–212]. An adaptive algorithm is developed. Numerical studies are reported for several test cases and show that the second error estimator is more efficient than the first one. In particular, the second error indicator is of optimal order with respect to both the mesh size and the time step when using our adaptive algorithm.

Key words. anisotropic finite elements, a posterori error estimates, Crank–Nicolson

AMS subject classifications. 65M15, 65M50, 65M60

DOI. 10.1137/080715135

1. Introduction. A posteriori error analysis is at the base of refinement/coarsening procedures in mesh adaptivity techniques. An impressive amount of work is available for a huge number of problems, as evidenced by the reviews [AO00, Ver96, BR03, BS01, Han05] and the references therein. The goal of a posteriori error estimates is to derive an easily computable bound of the error in order to ensure global control of the solution. The theory is particularly well developed in the case of elliptic problems; see, for instance, [AO93, AN98, Ver94, BV00]. In the case of parabolic problems, most papers deal with the Euler implicit time discretization (see, for instance, [Pic98, BBM05, CF04]), or higher order discontinuous Galerkin methods [EJ91, EJ95]. However, little attention was paid to the popular Crank–Nicolson method for parabolic problems. In [Ver03], the θ-scheme was considered for the time discretization of the heat equation and a posteriori bounds were derived applying standard energy techniques for a continuous, piecewise linear approximation in time. However, when considering the case $\theta = 1/2$, which corresponds to the Crank–Nicolson scheme, the error estimator corresponding to the energy norm was of suboptimal order with respect to the time step. In order to restore the appropriate second order of convergence, it was suggested in [AMN06] to work with a continuous, piecewise quadratic polynomial function in time rather than a piecewise linear one. The so-called Crank–Nicolson reconstruction was then introduced for a semidiscrete time discretization of a general

*Received by the editors February 5, 2008; accepted for publication (in revised form) March 30, 2009; published electronically June 25, 2009.
http://www.siam.org/journals/sisc/31-4/71513.html
†Université Paul Sabatier, Institut de Mathématiques de Toulouse, 118 route de Narbonne, 31062 Toulouse Cedex 9, France (alexei.lozinski@math.univ-toulouse.fr).
‡Institut d'Analyse et Calcul Scientifique, EPFL-SB-IACS-ASN, École Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland (marco.picasso@epfl.ch, virabouth.prachittham@epfl.ch).
The work of the third author was supported by the Swiss National Science Foundation.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

parabolic problem. This approach was generalized in [AMN07] to Runge–Kutta and Galerkin methods.

In the present paper, we are interested in the fully discrete situation, taking the example of the linear heat equation $\partial u/\partial t - \Delta u = f$ discretized in space by continuous piecewise linear finite elements and in time by the Crank–Nicolson method. We extend the results of [AMN06] to the fully discrete case by introducing piecewise quadratic (in time) reconstructions of the numerical solution and using them to obtain a posteriori error estimates. We consider two reconstructions. The first one is a direct transposition of the reconstruction from [AMN06] to our fully discrete setting, whereas the second one is new. It is based on a finite difference approximation of $\partial^2 u/\partial t^2$ rather than an approximation of $\partial f/\partial t + \Delta(\partial u/\partial t)$ and thus formally resembles the reconstruction from [AMN07]. A posteriori upper bounds are derived for both reconstructions and are used to construct an adaptive algorithm for both the time step and the space step. Numerical experiments show that the first error estimator does not lead to an optimal rate of convergence with respect to the time step when using our adaptive algorithm, especially in situations when the error is mainly due to the space discretization. On the other side, the second error estimator always provides a fair representation of the true error.

Another feature of our work is to use finite elements on anisotropic triangulations, that is to say, meshes having triangles with possibly large aspect ratio. The goal is to reduce, as much as possible, the number of degrees of freedom required to reach a given level of accuracy. For this reason, anisotropic a posteriori error estimates have recently received much attention [Kun00, Kun02, KN03, KV00, AL96, Ape99, Sie96, BJP04], with the goal to derive error estimates for which the involved constants do not depend on the triangle's aspect ratio. Here we pursue the anisotropic approach developed in [Pic98, Pic03, Pic06], which are, in turn, based on the anisotropic interpolation estimates derived in [FP03, FP01]. The results in this paper are thus presented in the anisotropic framework, and the numerical tests are done on examples that lead naturally to highly stretched meshes. Note, however, that our anisotropic a posteriori error bounds can also be translated in the isotropic setting, i.e., when the triangulation satisfies the minimum angle condition.

The outline of the paper is as follows. In the next section, we present the model problem and its time and space discretization. We will then introduce in section 3 some definitions and notations relative to the mesh anisotropy. The a posteriori error estimates for the two reconstructions are presented in section 4. Section 5 is devoted to the description of an adaptive algorithm in space and time. A numerical study is carried out for several test cases in section 6.

2. The heat equation and its discretization. Consider a polygonal domain Ω of $\mathbb{R}^2$ with boundary $\partial\Omega$. Given a final time $T > 0$, a function $f : \Omega \times (0,T) \to \mathbb{R}$, and an initial condition $u^0 : \Omega \to \mathbb{R}$, we consider the following problem: find $u : \Omega \times (0,T) \to \mathbb{R}$ such that

$$\left\{ \begin{array}{ll} \dfrac{\partial u}{\partial t} - \Delta u = f & \text{in } \Omega \times (0,T), \\ u = 0 & \text{on } \partial\Omega \times (0,T), \\ u(\cdot,0) = u^0 & \text{in } \Omega. \end{array} \right. \tag{2.1}$$

For the sake of simplicity, homogeneous Dirichlet boundary conditions are considered. However, the analysis can be extended to mixed Dirichlet–Neumann boundary

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

conditions; see Example 6.2 at the end of the paper. We suppose henceforth that $f \in L^2(0,T;H^{-1}(\Omega))$, $u^0 \in L^2(\Omega)$, and seek (see, for example, [DL92]) a solution $u \in W$ with

$$W = \{w \in L^2(0,T;H_0^1(\Omega)) \text{ and } \partial w/\partial t \in L^2(0,T;H^{-1}(\Omega))\} \tag{2.2}$$

such that $u(\cdot,0) = u^0$ and

$$\left\langle \frac{\partial u}{\partial t}, v \right\rangle + \int_\Omega \nabla u \cdot \nabla v \; dx = \langle f, v\rangle \quad \forall v \in H_0^1(\Omega) \text{ and a.e. } t \in (0,T), \tag{2.3}$$

where $< \cdot, \cdot >$ denotes the duality pairing between $H^{-1}(\Omega)$ and $H_0^1(\Omega)$. In order to describe the time discretization corresponding to (2.3), we introduce a partition of the interval $[0,T]$ into subintervals $I_n = [t^{n-1}, t^n]$, $n = 1, \ldots, N$, such that $0 = t^0 < t^1 < \cdots < t^N = T$, and we denote the time steps by $\tau_n = t^n - t^{n-1}$. For any $0 < h < 1$, let $\mathcal{T}_h$ be a conforming triangulation of $\overline{\Omega}$ into triangles K (not necessarily satisfying the minimum angle condition) with diameter h_K less than h. We define by V_h the usual finite element space of continuous, piecewise linear functions on $\mathcal{T}_h$:

$$V_h = \left\{ v_h \in \mathcal{C}^0(\overline{\Omega});\ v_{h|K} \in P_1;\ \forall K \in \mathcal{T}_h \right\},$$

and we set

$$V_h^0 = V_h \cap H_0^1(\Omega).$$

We suppose that $f \in C^0(0,T;L^2(\Omega))$ and $u^0 \in C^0(\overline{\Omega})$. We set $f^n(\cdot) = f(\cdot, t^n)$, and we compute $u_h^0 = r_h u^0$, where r_h is the Lagrange interpolant corresponding to V_h^0. The Crank–Nicolson scheme consists of seeking $u_h^n \in V_h^0$ for all $n = 1, \ldots, N$ such that $\forall v_h \in V_h^0$,

$$\int_\Omega \frac{u_h^n - u_h^{n-1}}{\tau_n} v_h \; dx + \frac{1}{2}\int_\Omega \left(\nabla u_h^n + \nabla u_h^{n-1}\right) \cdot \nabla v_h \; dx = \frac{1}{2}\int_\Omega (f^n + f^{n-1}) v_h \; dx. \tag{2.4}$$

Throughout the paper the following notations will be used for $n = 1, \ldots, N$:

$$\partial_n u_h = \frac{u_h^n - u_h^{n-1}}{\tau_n}, \quad u_h^{n-1/2} = \frac{1}{2}(u_h^n + u_h^{n-1})$$

and

$$\partial_n f = \frac{f^n - f^{n-1}}{\tau_n}, \quad f^{n-1/2} = \frac{1}{2}(f^n + f^{n-1}).$$

With these notations, we can rewrite (2.4) as

$$\int_\Omega \partial_n u_h \; v_h \; dx + \int_\Omega \nabla u_h^{n-1/2} \cdot \nabla v_h \; dx = \int_\Omega f^{n-1/2} v_h \; dx, \tag{2.5}$$

for all $v_h \in V_h^0$. As in [Pic98], we introduce the continuous, piecewise linear approximation in time defined for all $t \in I_n$ by

$$\begin{aligned} u_{h\tau}(x,t) &= \frac{t - t^{n-1}}{\tau_n} u_h^n + \frac{t^n - t}{\tau_n} u_h^{n-1} \\ &= u_h^{n-1/2} + (t - t^{n-1/2}) \partial_n u_h, \end{aligned} \tag{2.6}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

where $t^{n-1/2} = (t^n + t^{n-1})/2$. So for all $v_h \in V_h^0$, (2.4) or (2.5) can be rewritten as

$$\int_\Omega \partial_n u_h \; v_h \; dx + \int_\Omega \nabla u_{h\tau} \cdot \nabla v_h \; dx = \int_\Omega f^{n-1/2} v_h \; dx + (t - t^{n-1/2}) \int_\Omega \nabla \partial_n u_h \cdot \nabla v_h \; dx. \tag{2.7}$$

3. Anisotropic finite elements. In order to describe the mesh anisotropy, we introduce some definitions and properties taken from [FP03, FP01]. Alternative notations can be found using the framework of [Kun00, KV00]. For any triangle K of $\mathcal{T}_h$, we consider $T_K : \widehat{K} \to K$, the affine application that maps the triangle $\widehat{K}$ into K. Let $M_K \in \mathbb{R}^{2\times 2}$ and $t_K \in \mathbb{R}^2$ be the matrix and the vector defining such a map, and we have

$$x = T_K(\hat{x}) = M_K \hat{x} + t_K \quad \forall \hat{x} \in \mathbb{R}^2.$$

Since M_K is invertible, it admits a singular value decomposition

$$M_K = R_K^T \Lambda_K P_K,$$

where R_K and P_K are orthogonal and Λ_K is diagonal with positive entries. In the following we set

$$\Lambda_K = \begin{pmatrix} \lambda_{1,K} & 0 \\ 0 & \lambda_{2,K} \end{pmatrix} \quad \text{and} \quad R_K = \begin{pmatrix} r_{1,K}^T \\ r_{2,K}^T \end{pmatrix} \tag{3.1}$$

with the choice $\lambda_{1,K} \geq \lambda_{2,K}$. We refer to section 2 of [Pic06] for examples of such a transformation.

Let us now recall some results on interpolation on anisotropic meshes proved in [FP01, FP03]. The classical minimum angle condition is not required in this context. However, for each vertex, the number of neighboring vertices should be bounded from above, uniformly with respect to the mesh size h. There is another restriction on the mesh (see [Pic03] for a rigorous definition and illustrations) that prevents, loosely speaking, the stretching directions $r_{1,K}$, $r_{2,K}$ from changing too abruptly between the adjacent triangles of the mesh. In the rest of this paper we suppose that the family $\mathcal{T}_h$ meets the above-mentioned restrictions. In practice, the BL2D anisotropic mesh generator [BL96] that we have used meets these restrictions.

PROPOSITION 3.1 ([FP01, FP03]). *Let $I_h : H_0^1(\Omega) \to V_h^0$ be the Clément interpolation operator* [Clé75]. *There is a constant C independent of the mesh size and aspect ratio such that, for any $v \in H^1(\Omega)$ and any $K \in \mathcal{T}_h$, we have*

$$\|v - I_h v\|_{L^2(K)} + \lambda_{2,K} \|\nabla(v - I_h v)\|_{L^2(K)} + \lambda_{2,K}^{1/2} \|v - I_h v\|_{L^2(\partial K)} \leq C \omega_K(v). \tag{3.2}$$

Here $\omega_K(v)$ is defined by

$$\omega_K^2(v) = \lambda_{1,K}^2 \left(r_{1,K}^T G_K(v) r_{1,K} \right) + \lambda_{2,K}^2 \left(r_{2,K}^T G_K(v) r_{2,K} \right),$$

$\lambda_{i,K}$ and $r_{i,K}$ are given by (3.1), *and $G_K(v)$ is the following 2×2 matrix:*

$$G_K(v) = \sum_{T \in \triangle_K} \begin{pmatrix} \displaystyle\int_T \left(\frac{\partial v}{\partial x_1} \right)^2 dx & \displaystyle\int_T \frac{\partial v}{\partial x_1} \frac{\partial v}{\partial x_2} dx \\ \displaystyle\int_T \frac{\partial v}{\partial x_1} \frac{\partial v}{\partial x_2} dx & \displaystyle\int_T \left(\frac{\partial v}{\partial x_2} \right)^2 dx \end{pmatrix},$$

where $\triangle_K$ represents the set of triangles of $\mathcal{T}_h$ having a common vertex with K.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Remark 3.2. The reader should note that similar interpolation error estimates can be found in [Kun00, KV00]. For instance, estimate (3.2) is similar to estimates (11) and (12) in [Kun00] with different notations ($\lambda_{1,K} r_{1,K}$ plays the role of p_1 in [Kun00]).

4. A posteriori error estimates.

4.1. Piecewise quadratic reconstructions of the numerical solution. It was observed in [AMN06] that a direct use of $u_{h\tau}$ in an a posteriori error analysis would lead to a suboptimal estimate. It was proposed there to work rather with a piecewise quadratic time reconstruction of the numerical solution. We now recall briefly the construction of [AMN06] made for a Crank–Nicolson discretization of an abstract evolutionary equation

$$\frac{\partial u}{\partial t} + Au = f(t)$$

posed in a Hilbert space V, where A is a positive definite, self-adjoint, linear operator on V with the dense domain in V. We discretize the last equation in time as

$$\frac{u^n - u^{n-1}}{\tau_n} + A\frac{u^n + u^{n-1}}{2} = \frac{f(t^n) + f(t^{n-1})}{2} \tag{4.1}$$

and consider the linear interpolation

$$u_\tau(t) = u^{n-1} + (t - t^{n-1})\frac{u^n - u^{n-1}}{\tau_n}, \quad t \in I_n$$

and the quadratic one

$$\widehat{u}_\tau(t) = u_\tau(t) + \frac{1}{2}(t-t^{n-1})(t-t^n)\left(\frac{f(t^n) - f(t^{n-1})}{\tau_n} - A\frac{u^n - u^{n-1}}{\tau_n}\right), \quad t \in I_n; \tag{4.2}$$

see (3.5) in [AMN06]. The latter reconstruction allows an a posteriori error estimate to be obtained for (4.1), which is of optimal second order.

We return now to the heat equation (2.3) and its discretization (2.4) or equivalently (2.5). Comparing (2.5) and (4.1), we can interpret the operator A in (2.5) as the finite-dimensional approximation of $-\Delta$ in (2.5). This analogy allows us to introduce the following quadratic reconstruction:

$$\widehat{u}_{h\tau}(x,t) = u_{h\tau}(x,t) + \frac{1}{2}(t - t^{n-1})(t - t^n)w_h^n, \quad t \in I_n,\ 1 \le n \le N, \tag{4.3}$$

where $u_{h\tau}$ is defined in (2.6) and $w_h^n \in V_h^0$ is defined by

$$\int_\Omega w_h^n v_h \, dx = \int_\Omega (\partial_n f \; v_h - \nabla \partial_n u_h \cdot \nabla v_h) \, dx \quad \forall v_h \in V_h^0. \tag{4.4}$$

This reconstruction is the analog of (4.2) in the case of the heat equation discretized in both space and time. We will refer to $\widehat{u}_{h\tau}$ as the two-point reconstruction since it involves only u_h^n and u_h^{n-1}. We will use it to construct an upper bound for the error analogous to that of [AMN06]. However, considering the numerical results obtained with the time error estimator based on $\widehat{u}_{h\tau}$, we cannot conclude that it gets the optimal rate of convergence with respect to the time step when using our adaptive

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

algorithm. That is why we propose an alternative quadratic reconstruction in the next paragraph.

As a motivation, we observe that w_h^n in (4.3) is formally an approximation of $\partial f/\partial t + \Delta \partial u/\partial t$ in the time slab I_n, and the latter is equal to $\partial^2 u/\partial t^2$ by (2.1). It seems natural then to try to replace w_h^n in (4.3) by a finite difference approximation of $\partial^2 u/\partial t^2$. We introduce thus

$$\widetilde{u}_{h\tau}(x,t) = u_{h\tau}(x,t) + \frac{1}{2}(t-t^{n-1})(t-t^n)\partial_n^2 u_h, \quad t \in I_n,\ 2 \le n \le N, \tag{4.5}$$

where

$$\partial_n^2 u_h = \frac{\dfrac{u_h^n - u_h^{n-1}}{\tau_n} - \dfrac{u_h^{n-1} - u_h^{n-2}}{\tau_{n-1}}}{(\tau_n + \tau_{n-1})/2}. \tag{4.6}$$

Note that $\widetilde{u}_{h\tau}$ is again continuous, piecewise quadratic in time. We will refer to it as the three-point quadratic reconstruction since it involves u_h^n, u_h^{n-1}, and u_h^{n-2}. Note that $u_{h\tau}$, $\widehat{u}_{h\tau}$, and $\widetilde{u}_{h\tau}$ coincide at all times $t^1, \dots, t^N$.

Remark 4.1. The reconstruction $\widetilde{u}_{h\tau}$ restricted to the time interval I_n is the unique quadratic polynomial, which coincides with u_h^{n-2}, u_h^{n-1}, u_h^n at time t^{n-2}, t^{n-1}, t^n, respectively. Indeed, denoting the latter by $P_n(x,t)$, we observe that it is equal to $u_{h\tau}$ at $t = t^{n-1}$ and $t = t^n$ and $u_{h\tau}$ is linear in time on I_n. Hence, necessarily

$$P_n(x,t) = u_{h\tau}(x,t) + C_n(x)(t-t^{n-1})(t-t^n)$$

for some $C_n \in V_h^0$. Then we find C_n deriving P_n twice in t and taking into account that $P_n(\cdot, t^{n-2}) = u_h^{n-2}$. We thus find that $C_n = \partial_n^2 u_h/2$ and recover (4.5).

Remark 4.2. The requirement that w_h^n vanish on the boundary can seem inconsistent with the nature of the finite element problem for it, which does not contain any derivatives and thus does not require boundary conditions. We note, however, that w_h^n can be interpreted as an approximation of $\partial f/\partial t + \Delta \partial u/\partial t = \partial^2 u/\partial t^2$. As u is equal to zero on $\partial\Omega$, its second derivative in time is also equal to zero there. Consequently, it is reasonable to define w_h^n as an element of V_h^0.

4.2. A posteriori error estimates on anisotropic meshes. We will now present the two error indicators based on $\widehat{u}_{h\tau}$ and $\widetilde{u}_{h\tau}$ and used subsequently in our adaptive algorithm. In both cases, a standard energy technique is used that leads to combined error indicators in space and time. The estimator in space is similar to the one considered in [Pic03, Pic06]. In what follows we keep the notations of sections 2 and 4.1 and set $e = u - u_{h\tau}$, $\widehat{e} = u - \widehat{u}_{h\tau}$, and $\widetilde{e} = u - \widetilde{u}_{h\tau}$. We announce now our main results.

THEOREM 4.3. *Let $\widehat{f}$ be the linear interpolant of f defined by*

$$\widehat{f}(\cdot,t) = \frac{t-t^{n-1}}{\tau_n} f^n + \frac{t^n - t}{\tau_n} f^{n-1}, \quad t \in I_n,\ 1 \le n \le N,$$

and suppose that the mesh is such that there exists a constant c independent of the time step, mesh size, and aspect ratio such that

$$\lambda_{1,K}^2 \left(r_{1,K}^T G_K(\widehat{e}) r_{1,K}\right) \le c \lambda_{2,K}^2 \left(r_{2,K}^T G_K(\widehat{e}) r_{2,K}\right) \quad \forall K \in \mathcal{T}_h. \tag{4.7}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Then there is a constant C independent of the time step, mesh size, and aspect ratio such that

$$\int_0^T \|\nabla e\|^2_{L^2(\Omega)}\,dt + \|e(\cdot,T)\|^2_{L^2(\Omega)} \le \|e(\cdot,0)\|^2_{L^2(\Omega)} + C\sum_{n=1}^{N}\sum_{K\in\mathcal{T}_h}\left\{\int_{t^{n-1}}^{t^n}\left(\|f-\partial_n u_h+\Delta u_{h\tau}\|_{L^2(K)} + \frac{1}{2\lambda_{2,K}^{1/2}}\|[\nabla u_{h\tau}\cdot\mathbf{n}]\|_{L^2(\partial K)}\right)\omega_K(\widehat{e})\,dt + \int_{t^{n-1}}^{t^n}\left\|f-\widehat{f}\right\|^2_{L^2(K)}dt + \frac{\tau_n^5}{120}\|\nabla w_h^n\|^2_{L^2(K)} + \frac{\lambda_{2,K}^2\tau_n^3}{12}\|w_h^n\|^2_{L^2(K)}\right\}.$$

Here $[\cdot]$ denotes the jump of the bracketed quantity across an internal edge, $[\cdot]=0$ for an edge on the boundary $\partial\Omega$, and $\mathbf{n}$ is the unit edge normal (in arbitrary direction).

THEOREM 4.4. *Let*

$$\widetilde{f} = f^{n-1/2} + (t-t^{n-1/2})\frac{f^n-f^{n-2}}{\tau_n+\tau_{n-1}}, \quad t\in I_n,\ 2\le n\le N, \tag{4.8}$$

and suppose that the mesh is such that there exists a constant c independent of the time step, mesh size, and aspect ratio such that

$$\lambda_{1,K}^2\left(r_{1,K}^T G_K(\widetilde{e})r_{1,K}\right) \le c\lambda_{2,K}^2\left(r_{2,K}^T G_K(\widetilde{e})r_{2,K}\right) \quad \forall K\in\mathcal{T}_h. \tag{4.9}$$

Then there is a constant C independent of the time step, mesh size, and aspect ratio such that

$$\int_{t^1}^T \|\nabla e\|^2_{L^2(\Omega)}\,dt + \|e(\cdot,T)\|^2_{L^2(\Omega)} \le \left\|e(\cdot,t^1)\right\|^2_{L^2(\Omega)} + C\sum_{n=2}^{N}\sum_{K\in\mathcal{T}_h}\left\{\int_{t^{n-1}}^{t^n}\left(\|f-\partial_n u_h+\Delta u_{h\tau}\|_{L^2(K)} + \frac{1}{2\lambda_{2,K}^{1/2}}\|[\nabla u_{h\tau}\cdot\mathbf{n}]\|_{L^2(\partial K)}\right)\omega_K(\widetilde{e})\,dt + \int_{t^{n-1}}^{t^n}\left\|f-\widetilde{f}\right\|^2_{L^2(K)}dt + \left\{\frac{\tau_{n-1}^2\tau_n^3}{48}+\frac{\tau_n^5}{120}\right\}\left\|\nabla\partial_n^2 u_h\right\|^2_{L^2(K)} + \frac{\lambda_{2,K}^2\tau_n^3}{12}\left\|\partial_n^2 u_h\right\|^2_{L^2(K)}\right\}.$$

Remark 4.5. The bounds in Theorems 4.3 and 4.4 are not traditional a posteriori error estimates since they involve $\omega_K(\widehat{e})$ or $\omega_K(\widetilde{e})$ and hence the gradient of the exact solution u. A way to approximate it by a computable quantity was proposed in [Pic03, Pic06]. The resulting error estimator was proved very efficient for several problems, in particular, the Poisson equation and the Euler discretization of the heat equation. We will also apply this technique in constructing our error estimators and the adaptive algorithm for the Crank–Nicolson scheme in this paper.

Remark 4.6. As we will see in section 5, our anisotropic adaptive algorithm ensures assumptions (4.7) and (4.9) to be fulfilled with $c=1$. A similar assumption has been made in [Pic06] in order to prove a lower bound in the framework of the Laplace problem.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

We will prove here only Theorem 4.4 because the proof of Theorem 4.3 can follow the same lines with even less technicalities. We first need the following result.

PROPOSITION 4.7. *Let $\widetilde{u}_{h\tau}$ be defined by* (4.5) *and $\widetilde{f}$ by* (4.8)*, then we have*

$$\int_\Omega \frac{\partial \widetilde{u}_{h\tau}}{\partial t} v_h \, dx + \int_\Omega \nabla u_{h\tau} \cdot \nabla v_h \, dx = \int_\Omega \widetilde{f} v_h \, dx + (t - t^{n-1/2}) \frac{\tau_{n-1}}{2} \int_\Omega \nabla \partial_n^2 u_h \cdot \nabla v_h \, dx,$$

for all $v_h \in V_h^0$ and for any $t \in I_n$, $2 \le n \le N$.

Proof. We choose any $2 \le n \le N$ and $t \in I_n$ in this proof. We have from (4.5) that

$$\frac{\partial \widetilde{u}_{h\tau}}{\partial t} = \partial_n u_h + (t - t^{n-1/2}) \partial_n^2 u_h. \tag{4.10}$$

Then, using (2.7), we have for all $v_h \in V_h^0$

$$\int_\Omega \frac{\partial \widetilde{u}_{h\tau}}{\partial t} v_h \, dx + \int_\Omega \nabla u_{h\tau} \cdot \nabla v_h \, dx = \int_\Omega f^{n-1/2} \, v_h \, dx + (t - t^{n-1/2}) \int_\Omega \left\{ \partial_n^2 u_h \, v_h + \nabla \partial_n u_h \cdot \nabla v_h \right\} dx. \tag{4.11}$$

We invoke now (2.4) at the time t^{n-1}

$$\int_\Omega \frac{u_h^{n-1} - u_h^{n-2}}{\tau_{n-1}} \, v_h \, dx + \int_\Omega \frac{\nabla (u_h^{n-1} + u_h^{n-2})}{2} \cdot \nabla v_h \, dx = \int_\Omega \frac{f^{n-1} + f^{n-2}}{2} \, v_h \, dx,$$

and we subtract it from (2.4) to obtain for all $v_h \in V_h^0$

$$\int_\Omega \left\{ \partial_n^2 u_h \, v_h + \nabla \left(\frac{u_h^n - u_h^{n-2}}{\tau_n + \tau_{n-1}} \right) \cdot \nabla v_h \right\} dx = \frac{1}{\tau_n + \tau_{n-1}} \int_\Omega (f^n - f^{n-2}) v_h \, dx.$$

Thus

$$\int_\Omega \left\{ \partial_n^2 u_h \, v_h + \nabla \partial_n u_h \cdot \nabla v_h \right\} dx = \frac{1}{\tau_n + \tau_{n-1}} \int_\Omega (f^n - f^{n-2}) v_h \, dx + \frac{\tau_{n-1}}{2} \int_\Omega \nabla \partial_n^2 u_h \cdot \nabla v_h \, dx. \tag{4.12}$$

It suffices now to insert (4.12) into (4.11) to obtain the result. ∎

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Proof of Theorem 4.4. We choose any $2 \le n \le N$ and $t \in I_n$. Using (2.7), (4.10), and Proposition 4.7, we obtain for all $v \in H_0^1(\Omega)$ and all $v_h \in V_h^0$

$$\begin{aligned}
&\int_\Omega \frac{\partial \widetilde{e}}{\partial t} v \, dx + \int_\Omega \nabla e \cdot \nabla v \, dx \\
&= \int_\Omega (f - \partial_n u_h)\, v \, dx - \int_\Omega \nabla u_{h\tau} \cdot \nabla v \, dx - (t - t^{n-1/2}) \int_\Omega \partial_n^2 u_h \; v \, dx \\
&= \int_\Omega (f - \partial_n u_h)\, (v - v_h) \, dx - \int_\Omega \nabla u_{h\tau} \cdot \nabla (v - v_h) \, dx \\
&\quad + \int_\Omega \left(f - \frac{\partial \widetilde{u}_{h\tau}}{\partial t} \right) v_h \, dx - \int_\Omega \nabla u_{h\tau} \cdot \nabla v_h \, dx - (t - t^{n-1/2}) \int_\Omega \partial_n^2 u_h (v - v_h) dx \\
&= \int_\Omega (f - \partial_n u_h)\, (v - v_h) \, dx - \int_\Omega \nabla u_{h\tau} \cdot \nabla (v - v_h) \, dx \\
&\quad - (t - t^{n-1/2}) \frac{\tau_{n-1}}{2} \int_\Omega \nabla \partial_n^2 u_h \cdot \nabla v_h \, dx + \int_\Omega (f - \widetilde{f}) \; v_h \, dx \\
&\quad - (t - t^{n-1/2}) \int_\Omega \partial_n^2 u_h \; (v - v_h) \, dx.
\end{aligned}$$

Then taking $v_h = I_h v$ and integrating by parts on each triangle K, we obtain

(4.13)

$$\begin{aligned}
&\int_\Omega \frac{\partial \widetilde{e}}{\partial t} v \, dx + \int_\Omega \nabla e \cdot \nabla v \, dx = \\
&\quad \sum_{K \in \mathcal{T}_h} \left\{ \int_K (f - \partial_n u_h + \Delta u_{h\tau})\, (v - I_h v) \, dx - \frac{1}{2} \int_{\partial K} [\nabla u_{h\tau} \cdot \mathbf{n}]\, (v - I_h v) \, dx \right. \\
&\quad - (t - t^{n-1/2}) \frac{\tau_{n-1}}{2} \int_K \nabla \partial_n^2 u_h \cdot \nabla I_h v \, dx + \int_K (f - \widetilde{f}) \; I_h v \, dx \\
&\quad \left. - (t - t^{n-1/2}) \int_K \partial_n^2 u_h \; (v - I_h v) \, dx \right\}.
\end{aligned}$$

Taking $v = \widetilde{e}$, rewriting $\nabla e \cdot \nabla \widetilde{e} = \frac{1}{2}|\nabla e|^2 + \frac{1}{2}|\nabla \widetilde{e}|^2 - \frac{1}{2}|\nabla e - \nabla \widetilde{e}|^2$ with

$$\|\nabla(e - \widetilde{e})\|^2_{L^2(\Omega)} = \|\nabla(\widetilde{u}_{h\tau} - u_{h\tau})\|^2_{L^2(\Omega)} = \frac{1}{4}(t - t^{n-1})^2 (t - t^n)^2 \|\nabla \partial_n^2 u_h\|^2_{L^2(\Omega)}, \tag{4.14}$$

and using the Cauchy–Schwarz, Young, and Poincaré inequalities together with Proposition 3.1, we have

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt} \int_\Omega |\widetilde{e}|^2 \, dx + \frac{1}{2} \int_\Omega |\nabla e|^2 \, dx + \frac{1}{2} \int_\Omega |\nabla \widetilde{e}|^2 \, dx \\
&\le \sum_{K \in \mathcal{T}_h} \left\{ C_1 \left[\left(\|f - \partial_n u_h + \Delta u_{h\tau}\|_{L^2(K)} + \frac{1}{2\lambda_{2,K}^{1/2}} \|[\nabla u_{h\tau} \cdot \mathbf{n}]\|_{L^2(\partial K)} \right) \omega_K(\widetilde{e}) \right. \right. \\
&\qquad \left. + |t - t^{n-1/2}| \left\| \partial_n^2 u_h \right\|_{L^2(K)} \omega_K(\widetilde{e}) \right] \\
&\qquad + \left\{ \frac{p}{8} (\tau_{n-1})^2 (t - t^{n-1/2})^2 + \frac{1}{8} (t - t^{n-1})^2 (t - t^n)^2 \right\} \left\| \nabla \partial_n^2 u_h \right\|^2_{L^2(K)} \\
&\qquad \left. + \frac{p}{2} \left\| f - \widetilde{f} \right\|^2_{L^2(K)} + \frac{1 + C_2^2}{2p} \|\nabla I_h \widetilde{e}\|^2_{L^2(K)} \right\},
\end{aligned}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

where C_1 is the constant of Proposition 3.1, C_2 is the constant in the Poincaré inequality, and $p > 0$ is arbitrary. Error equidistribution inequality (4.9) combined with Proposition 3.1 implies that

$$\omega_K(\widetilde{e}) \le C_3 \lambda_{2,K} \left\|\nabla \widetilde{e}\right\|_{L^2(K)} \quad \text{and} \quad \left\|\nabla I_h \widetilde{e}\right\|_{L^2(K)} \le C_4 \left\|\nabla \widetilde{e}\right\|_{L^2(K)}. \tag{4.15}$$

We have then

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\int_\Omega |\widetilde{e}|^2\,dx + \frac{1}{2}\int_\Omega |\nabla e|^2\,dx + \frac{1}{2}\int_\Omega |\nabla \widetilde{e}|^2\,dx \\
&\le \sum_{K\in\mathcal{T}_h}\left\{ C_1\left[\left(\left\|f-\partial_n u_h + \Delta u_{h\tau}\right\|_{L^2(K)} + \frac{1}{2\lambda_{2,K}^{1/2}}\left\|[\nabla u_{h\tau}\cdot\mathbf{n}]\right\|_{L^2(\partial K)}\right)\omega_K(\widetilde{e})\right.\right. \\
&\quad \left.\left. + \frac{p}{2}\lambda_{2,K}^2 (t-t^{n-1/2})^2 \left\|\partial_n^2 u_h\right\|_{L^2(K)}^2\right] + \frac{C_1 C_3^2}{2p}\left\|\nabla\widetilde{e}\right\|_{L^2(K)}^2 \right. \\
&\quad + \left\{\frac{p}{8}\tau_{n-1}^2 (t-t^{n-1/2})^2 + \frac{1}{8}(t-t^{n-1})^2(t-t^n)^2\right\}\left\|\nabla\partial_n^2 u_h\right\|_{L^2(K)}^2 \\
&\quad \left. + \frac{p}{2}\left\|f-\widetilde{f}\right\|_{L^2(K)}^2 + \frac{1+C_2^2}{2p}\left\|\nabla I_h\widetilde{e}\right\|_{L^2(K)}^2\right\}.
\end{aligned} \tag{4.16}$$

Finally, use the second inequality of (4.15) in (4.16), choose $p = C_1C_3^2 + C_4^2(1+C_2^2)$, and integrate (4.16) between $t = t^{n-1}$ and $t = t^n$ to obtain

$$\begin{aligned}
&\int_{t^{n-1}}^{t^n} \left\|\nabla e\right\|_{L^2(\Omega)}^2\,dt + \left\|\widetilde{e}(\cdot,t^n)\right\|_{L^2(\Omega)}^2 \le \left\|\widetilde{e}(\cdot,t^{n-1})\right\|_{L^2(\Omega)}^2 \\
&+ C\sum_{K\in\mathcal{T}_h}\left\{\int_{t^{n-1}}^{t^n}\left(\left\|f-\partial_n u_h+\Delta u_{h\tau}\right\|_{L^2(K)} + \frac{1}{2\lambda_{2,K}^{1/2}}\left\|[\nabla u_{h\tau}\cdot\mathbf{n}]\right\|_{L^2(\partial K)}\right)\omega_K(\widetilde{e})\,dt\right. \\
&+ \int_{t^{n-1}}^{t^n}\left\|f-\widetilde{f}\right\|_{L^2(K)}^2\,dt + \left\{\frac{\tau_{n-1}^2\tau_n^3}{48} + \frac{\tau_n^5}{120}\right\}\left\|\nabla\partial_n^2 u_h\right\|_{L^2(K)}^2 \\
&\left. + \frac{\lambda_{2,K}^2\tau_n^3}{12}\left\|\partial_n^2 u_h\right\|_{L^2(K)}^2\right\},
\end{aligned}$$

where $C = \max(1, 2C_1, pC_1, p)$. Summing up this inequality on n and noting that $\widetilde{e}(t^n) = e(t^n)\ \forall n$ leads to the final result. □

4.3. A posteriori error indicators. As already noted in the previous section, the upper bounds derived in Theorems 4.3 and 4.4 are not traditional a posteriori error estimates since they involve $\omega_K(\widehat{e})$ and $\omega_K(\widetilde{e})$ and hence the exact unknown solution u. Therefore, following [Pic03, Pic06], we introduce some Zienkiewicz–Zhu postprocessing in order to estimate these unknown quantities. From the theoretical viewpoint, on general unstructured meshes, the Zienkiewicz–Zhu error estimator [ZZ87, ZZ92] is proved to be equivalent to the true error; see, for instance, [Rod94, Car04] for isotropic meshes and [KN03] for anisotropic meshes. The asymptotic exactness of this estimator can be justified whenever superconvergence occurs, that is, on parallel [Rod94, AO00] or mildly structured meshes [XZ04]. Zienkiewicz–Zhu postprocessing for anisotropic meshes was used in [KN03] to approximate the so-called matching function and in

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

[Pic03, Pic06] to build triangulations with large aspect ratio. The numerical experiments showed that the Zienkiewicz–Zhu error estimator performs surprisingly well, even on general unstructured anisotropic meshes. Here, we follow this approach and use this postprocessing method to derive our space error indicator. The interested reader should note that alternative gradient recovery methods based on least square fitting have been proposed in [NZ04, ZN05] and have proved to be even more robust than Zienkiewicz–Zhu postprocessing.

Our Zienkiewicz–Zhu error estimator is defined to be the difference between $\nabla u_{h\tau}$ and an approximate $L^2(\Omega)$ projection onto V_h:

$$\eta^{ZZ}(u_{h\tau}) = \begin{pmatrix} \eta_1^{ZZ}(u_{h\tau}) \\ \eta_2^{ZZ}(u_{h\tau}) \end{pmatrix} = \begin{pmatrix} (I-\Pi_h)\left(\dfrac{\partial u_{h\tau}}{\partial x_1}\right) \\ (I-\Pi_h)\left(\dfrac{\partial u_{h\tau}}{\partial x_2}\right) \end{pmatrix},$$

where $\Pi_h(\nabla u_{h\tau}) \in V_h$ is defined by its values at each vertex P as

$$\Pi_h(\nabla u_{h\tau})(P) = \begin{pmatrix} \Pi_h\left(\dfrac{\partial u_{h\tau}}{\partial x_1}\right)(P) \\ \Pi_h\left(\dfrac{\partial u_{h\tau}}{\partial x_2}\right)(P) \end{pmatrix} = \frac{1}{\displaystyle\sum_{\substack{K\in\mathcal{T}_h \\ P\in K}} |K|} \begin{pmatrix} \displaystyle\sum_{\substack{K\in\mathcal{T}_h \\ P\in K}} |K| \left(\dfrac{\partial u_{h\tau}}{\partial x_1}\right)_{|K} \\ \displaystyle\sum_{\substack{K\in\mathcal{T}_h \\ P\in K}} |K| \left(\dfrac{\partial u_{h\tau}}{\partial x_2}\right)_{|K} \end{pmatrix}.$$

Then, our error indicator is obtained by replacing $G_K(\widehat{e})$ in $\omega_K(\widehat{e})$ and $G_K(\widetilde{e})$ in $\omega_K(\widetilde{e})$ by $\breve{G}_K(u_{h\tau})$ defined for any $v_h \in V_h$ by

$$\breve{G}_K(v_h) = \begin{pmatrix} \displaystyle\int_K \left(\eta_1^{ZZ}(v_h)\right)^2 \, dx & \displaystyle\int_K \eta_1^{ZZ}(v_h)\eta_2^{ZZ}(v_h) \, dx \\ \displaystyle\int_K \eta_1^{ZZ}(v_h)\eta_2^{ZZ}(v_h) \, dx & \displaystyle\int_K \left(\eta_2^{ZZ}(v_h)\right)^2 \, dx \end{pmatrix}.$$

We define the anisotropic space error estimator η^A as

$$\eta^A = \left(\sum_{n=1}^{N} \sum_{K\in\mathcal{T}_h} \left(\eta_{K,n}^A(u_{h\tau})\right)^2\right)^{1/2},$$

where the contributions $\eta_{K,n}^A$ are defined on each triangle K of $\mathcal{T}_h$ and each time interval I_n as

$$\begin{aligned} \left(\eta_{K,n}^A(u_{h\tau})\right)^2 = \int_{t^{n-1}}^{t^n} & \left(\|f - \partial_n u_h + \Delta u_{h\tau}\|_{L^2(K)} + \frac{1}{2\lambda_{2,K}^{1/2}} \|[\nabla u_{h\tau}\cdot\mathbf{n}]\|_{L^2(\partial K)} \right) \\ & \times \left(\lambda_{1,K}^2 \left(r_{1,K}^T \breve{G}_K(u_{h\tau}) r_{1,K} \right) + \lambda_{2,K}^2 \left(r_{2,K}^T \breve{G}_K(u_{h\tau}) r_{2,K} \right) \right)^{1/2} dt. \end{aligned} \tag{4.17}$$

We introduce now two time error estimators: $\widehat{\eta}^T$ corresponding to the two-point reconstruction $\widehat{u}_{h\tau}$ (cf. Theorem 4.3) and $\widetilde{\eta}^T$ corresponding to the three-point reconstruction $\widetilde{u}_{h\tau}$ (cf. Theorem 4.4) defined, respectively, by

$$\widehat{\eta}^T = \left(\sum_{n=1}^{N} \left(\widehat{\eta}_n^T(u_{h\tau})\right)^2\right)^{1/2} \quad \text{and} \quad \widetilde{\eta}^T = \left(\sum_{n=2}^{N} \left(\widetilde{\eta}_n^T(u_{h\tau})\right)^2\right)^{1/2}.$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

The contributions $\widehat{\eta}_n^T$ and $\widetilde{\eta}_n^T$ are computed on each time interval I_n, $n \geq 1$, via

(4.18)
$$\left(\widehat{\eta}_n^T(u_{h\tau})\right)^2 = \int_{t^{n-1}}^{t^n} \left\| f - \widehat{f} \right\|_{L^2(\Omega)}^2 \, dt + \frac{\tau_n^5}{120} \|\nabla w_h^n\|_{L^2(\Omega)}^2 + \sum_{K \in \mathcal{T}_h} \frac{\lambda_{2,K}^2 \tau_n^3}{12} \|w_h^n\|_{L^2(K)}^2 ,$$

and for $n \geq 2$,

(4.19)
$$\left(\widetilde{\eta}_n^T(u_{h\tau})\right)^2 = \int_{t^{n-1}}^{t^n} \left\| f - \widetilde{f} \right\|_{L^2(\Omega)}^2 \, dt + \left\{ \frac{\tau_{n-1}^2 \tau_n^3}{48} + \frac{\tau_n^5}{120} \right\} \left\| \nabla \partial_n^2 u_h \right\|_{L^2(\Omega)}^2 + \sum_{K \in \mathcal{T}_h} \frac{\lambda_{2,K}^2 \tau_n^3}{12} \left\| \partial_n^2 u_h \right\|_{L^2(K)}^2 .$$

In our implementation, all the integrals between t^{n-1} and t^n are approximated by the midpoint rule.

In order to measure the quality of our estimators, the estimated error is compared to the true error introducing the so-called effectivity index, ei. Thus, we define the following effectivity indices in both space

$$ei^A = \frac{\eta^A}{\|e\|_{L^2(0,T;H^1(\Omega))}}$$

and time

$$\widehat{ei}^T = \frac{\widehat{\eta}^T}{\|e\|_{L^2(0,T;H^1(\Omega))}}, \ \widetilde{ei}^T = \frac{\widetilde{\eta}^T}{\|e\|_{L^2(t^1,T;H^1(\Omega))}}, \ ei^T = \frac{\eta^T}{\|e\|_{L^2(0,T;H^1(\Omega))}}.$$

We will also check the behavior of the Zienkiewicz–Zhu error estimator. We thus introduce the corresponding global estimator and the effectivity index

$$\eta^{ZZ} = \left(\sum_{n=1}^{N} \sum_{K \in \mathcal{T}_h} \int_{t^{n-1}}^{t^n} \int_K |\eta^{ZZ}(u_{h\tau})|^2 \, dx \, dt \right)^{1/2} \quad \text{and} \quad ei^{ZZ} = \frac{\eta^{ZZ}}{\|e\|_{L^2(0,T;H^1(\Omega))}}.$$

4.4. A posteriori error estimates on isotropic meshes. We deviate now from the framework of anisotropic meshes to the more traditional case of isotropic ones; i.e., we consider in this subsection a family of triangulations $\mathcal{T}_h$ satisfying the minimum angle condition, so that $\lambda_{1,K} \simeq \lambda_{2,K} \simeq h_K \leq h$ for all $K \in \mathcal{T}_h$. A much more complete theory of a posteriori error estimation can be constructed for the Crank–Nicolson method in this case. Indeed, the error indicator in space can be rewritten in a closed form (no longer involving the unknown exact solution). Moreover, upper and lower bounds of optimal order can be proved for the space-time error estimator provided the error is measured with respect to the quadratic reconstructions $\widehat{u}_{h\tau}$ or $\widetilde{u}_{h\tau}$, and we choose a slightly stronger norm than before, namely,

$$|| \cdot ||_W^2 = || \cdot ||_{L^2(0,T;H_0^1(\Omega))}^2 + \left\| \frac{\partial \cdot}{\partial t} \right\|_{L^2(0,T;H^{-1}(\Omega))}^2 .$$

Note that this is the natural norm of the space W introduced in (2.2), and we have $||u||_{L^\infty(0,T;L^2(\Omega))} \leq ||u||_W$ for all $u \in W$. Such a norm was used in [Ver03] to obtain upper and lower bounds for the θ-scheme. More precisely, it was proved that the error

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

$||u - u_{h\tau}||_W$ was bounded above and below by some error estimator being of order $O(h+\tau)$, even for the Crank–Nicolson scheme $\theta = 1/2$. Here, we can prove that the errors $||\hat{e}||_W = ||u - \hat{u}_{h\tau}||_W$ or $||\widetilde{e}||_W = ||u - \widetilde{u}_{h\tau}||_W$ are bounded above and below by an error estimator which is of optimal order $O(h+\tau^2)$.

Let us outline this theory in the case of the three-point reconstruction $\widetilde{u}_{h\tau}$, keeping in mind that similar statements can be made about the two-point reconstruction $\widehat{u}_{h\tau}$. Consider the isotropic space error indicator corresponding to (4.17),

$$\left(\eta^{iso}_{K,n}(u_{h\tau})\right)^2 = \int_{t^{n-1}}^{t^n} \left(h_K^2 \left\| f - \partial_n u_h + \Delta u_{h\tau} \right\|^2_{L^2(K)} + h_K \left\| [\nabla u_{h\tau} \cdot \mathbf{n}] \right\|^2_{L^2(\partial K)} \right) dt,$$

and the time error indicator (4.19), in which we keep only the second order finite difference in time of the approximated solution:

$$\left(\widetilde{\eta}^{T,iso}_n(u_{h\tau})\right)^2 = \left\{ \frac{\tau^2_{n-1}\tau^3_n}{48} + \frac{\tau^5_n}{120} \right\} \left\| \nabla \partial^2_n u_h \right\|^2_{L^2(\Omega)}.$$

Then we have the global upper bound of the error

$$(4.20)\quad ||\widetilde{e}||^2_{L^2(t^1,T;H^1_0(\Omega))} + \left\| \frac{\partial \widetilde{e}}{\partial t} \right\|^2_{L^2(t^1,T;H^{-1}(\Omega))} \le ||e(\cdot,t^1)||^2_{L^2(\Omega)}$$
$$+ C \sum_{n=2}^{N} \left\{ \sum_{K\in\mathcal{T}_h} (\eta^{iso}_{K,n})^2 + (\widetilde{\eta}^{T,iso}_n)^2 + \int_{t^{n-1}}^{t^n} \left\| f - \widetilde{f} \right\|^2_{H^{-1}(\Omega)} dt + \sum_{K\in\mathcal{T}_h} \frac{h_K^2 \tau_n^3}{12} \left\| \partial^2_n u_h \right\|^2_{L^2(K)} \right\}$$

and the local in time lower bound for all $n \ge 2$

$$(4.21)\quad \frac{1}{3}(\widetilde{\eta}^{T,iso}_n)^2 \le ||\widetilde{e}||^2_{L^2(t^{n-1},t^n;H^1_0(\Omega))} + \left\| \frac{\partial \widetilde{e}}{\partial t} \right\|^2_{L^2(t^{n-1},t^n;H^{-1}(\Omega))} + \int_{t^{n-1}}^{t^n} \left\| f - \widetilde{f} \right\|^2_{H^{-1}(\Omega)} dt.$$

We no longer have to assume the error equidistribution (4.9), but the constant C now depends on the mesh aspect ratio. The last term in (4.20) is dominated by $(\widetilde{\eta}^{T,iso}_n)^2$ provided $h = O(\tau^2)$, and thus it can be neglected in practice.

The proof of (4.20) is essentially the same as that of Theorem 4.4. Indeed, using standard estimates for the Clément interpolation I_h on isotropic meshes in (4.13), we arrive at

$$(4.22)\quad \int_\Omega \frac{\partial \widetilde{e}}{\partial t} v \, dx + \int_\Omega \nabla e \cdot \nabla v \, dx$$
$$\le \left\{ C \sum_{K\in\mathcal{T}_h} \left(h_K^2 \left\| f - \partial_n u_h + \Delta u_{h\tau} \right\|^2_{L^2(K)} + h_K \left\| [\nabla u_{h\tau} \cdot \mathbf{n}] \right\|^2_{L^2(\partial K)} \right) \right.$$
$$+ (t - t^{n-1/2})^2 \, \frac{\tau^2_{n-1}}{4} ||\nabla \partial^2_n u_h||^2_{L^2(\Omega)} + ||f - \widetilde{f}||^2_{H^{-1}(\Omega)}$$
$$\left. + (t - t^{n-1/2})^2 \sum_{K\in\mathcal{T}_h} h_K^2 ||\partial^2_n u_h||^2_{L^2(K)} \right\}^{\frac{1}{2}} \times ||v||_{H^1(\Omega)}, \quad \forall v \in H^1_0(\Omega),\ t \in [t^{n-1}, t^n].$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

TABLE 4.1
Convergence results using uniform isotropic meshes and constant time steps, case (a).

h	τ	$\left(\int_0^T \int_\Omega \lvert\nabla e\rvert^2\right)^{1/2}$	ei^{ZZ}	ei^{A}	$\widetilde{ei}^T$	$\widehat{ei}^T$
0.0125	0.025	0.17	0.10	0.25	17.23	14.05
0.00625	0.025	0.17	0.05	0.125	17.35	14.15
0.003125	0.025	0.17	0.024	0.06	17.39	14.17
0.0125	0.0125	0.047	0.44	1.09	16.25	13.22
0.00625	0.0125	0.043	0.24	0.60	17.85	14.53
0.003125	0.0125	0.042	0.12	0.30	18.42	14.92

Setting $v = \widetilde{e}$, integrating in time and noting (4.14) yields estimate (4.20) for $||\widetilde{e}||^2_{L^2(t^1,T;H^1_0(\Omega))}$. In a similar manner, we can take v in (4.22) such that

$$\int_\Omega \frac{\partial \widetilde{e}}{\partial t} v = \left\| \frac{\partial \widetilde{e}}{\partial t} \right\|_{H^{-1}(\Omega)} ||v||_{H^1(\Omega)}$$

and simplify by $||v||_{H^1(\Omega)}$. We thus obtain estimate (4.20) for $\|\partial\widetilde{e}/\partial t\|_{L^2(t^1,T;H^{-1}(\Omega))}$. On the other hand, the identity in Proposition 4.7 can be rewritten as

$$\frac{1}{2}\left(-(t-t^{n-1/2})\tau_{n-1} - (t-t^{n-1})(t-t^n)\right) \int_\Omega \nabla \partial_n^2 u_h \cdot \nabla v_h \, dx$$
$$= \int_\Omega \frac{\partial \widetilde{e}}{\partial t} v_h \, dx + \int_\Omega \nabla \widetilde{e} \cdot \nabla v_h \, dx - \int_\Omega (f - \widetilde{f}) v_h \, dx.$$

This is valid for any $v_h \in V_h^0$. In particular, taking $v_h = \partial_n^2 u_h$ and integrating from t^{n-1} to t^n we obtain the lower bound (4.21).

4.5. A numerical study of the error estimators with uniform time steps and mesh size. We study here the effectivity indices corresponding to the two error estimators $\widehat{\eta}^T$ and $\widetilde{\eta}^T$ on several test cases for which the error comes from either the space discretization or the time discretization, or from both of them. We consider the problem (2.1) with $\Omega = (0,1) \times (0,1)$, the unit square, $T = 1$, and the exact solution u given by

$$\begin{aligned}
&\text{case (a)} \qquad u(x,y,t) = sin(15\pi t)\ sin(\pi x)\ sin(\pi y),\\
&\text{case (b)} \qquad u(x,y,t) = sin(\pi t/2)\ sin(10\pi x)\ sin(10\pi y),\\
&\text{case (c)} \qquad u(x,y,t) = sin(\pi t)\ sin(\pi x)\ sin(\pi y).
\end{aligned}$$

Note that in case (a) the error should be mainly due to the time discretization, while in case (b) it should be mainly due to space discretization. Case (c) provides an example in which the error comes from both time and space discretization. The numerical results are reported in Tables 4.1–4.3. Isotropic meshes and constant time steps are used in all the experiments of this section.

Referring to Table 4.1, we observe that, the computed error in the test case (a) is mainly due to the time discretization. Indeed, for a given time step, the error does not depend on the space step h, and for constant h, the error is divided by four when the time step τ is divided by two. Moreover the two time error estimators $\widetilde{\eta}^T$ and $\widehat{\eta}^T$ behave as the true error.

Referring to Table 4.2, case (b), the error is now mainly due to space discretization. We observe that for constant h, the error does not depend on the time step τ,

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

TABLE 4.2
Convergence results using uniform isotropic meshes and constant time steps, case (b).

h	τ	$\left(\int_0^T \int_\Omega \lvert\nabla e\rvert^2\right)^{1/2}$	ei^{ZZ}	ei^A	$\widetilde{ei}^T$	$\widehat{ei}^T$
0.00625	0.05	1.11	1.01	2.47	0.48	0.61
0.00625	0.025	1.11	1.01	2.47	0.12	0.16
0.00625	0.0125	1.11	1.01	2.47	0.03	0.04
0.003125	0.05	0.56	1.00	2.46	0.96	1.21
0.003125	0.025	0.56	1.00	2.46	0.24	0.31
0.003125	0.0125	0.56	1.00	2.46	0.06	0.078

TABLE 4.3
Convergence results using uniform isotropic meshes and constant time steps, case (c).

h	τ	$\left(\int_0^T \int_\Omega \lvert\nabla e\rvert^2\right)^{1/2}$	ei^{ZZ}	ei^A	$\widetilde{ei}^T$	$\widehat{ei}^T$
0.025	0.05	0.044	1.00	2.46	0.52	0.42
0.00625	0.025	0.011	1.00	2.45	0.51	0.43
0.0015625	0.0125	0.0028	1.00	2.45	0.51	0.49
0.025	0.0125	0.044	1.00	2.46	0.032	0.027
0.00625	0.00625	0.011	1.00	2.45	0.032	0.034
0.0015625	0.003125	0.0028	1.00	2.45	0.032	0.067

that the space effectivity index stays close to 2.5, and that the Zienkiewicz–Zhu error estimator is asymptotically exact. Thus, when the error is mainly due to the space discretization we can see that the space error estimator η^A behaves as the true error.

Referring to Table 4.3, case (c), the error comes now from both time and space discretizations. We observe that the error in the $L^2(0,T;H^1(\Omega))$ norm is $O\left(h+\tau^2\right)$, that the space error estimator and the two time error estimators, $\widehat{\eta}^T$ and $\widetilde{\eta}^T$, are equivalent to the true error, and that the Zienkiewicz–Zhu error estimator is asymptotically exact. Thus, using uniform time steps and mesh size, we observe that the two time error estimators, $\widehat{\eta}^T$ and $\widetilde{\eta}^T$, provide a good representation of the true error.

5. Adaptive algorithm. We now propose a time and space adaptive finite element algorithm. We will describe this algorithm while using (4.17) and (4.19). Since the time error estimator needs a solution u_h^{n-2}, we do not change the first time step. For $n \geq 2$, the idea is to build successive triangulations $\mathcal{T}_h^n$ with possibly large aspect ratio and to choose appropriate time steps τ_n so that the relative estimated error in space and time in the $L^2(0,T;H^1(\Omega))$ norm is close to a preset tolerance TOL, for example,

$$0.875\ TOL\ \leq \frac{\left((\eta^A)^2+(\widetilde{\eta}^T)^2\right)^{1/2}}{\left(\displaystyle\int_0^T \int_\Omega |\nabla u_{h\tau}|^2\ dx\ dt\right)^{1/2}} \leq 1.125\ TOL. \tag{5.1}$$

In doing so, we are beyond the scope of the theory developed in the previous section since the mesh $\mathcal{T}_h$ was not allowed to vary in time there. A more rigorous adaptive procedure would have to include the error due to the interpolations from $\mathcal{T}_h^{n-1}$ to $\mathcal{T}_h^n$. We do not attempt to develop such a theory here and conjecture that the interpolation error can be neglected provided the total number of remeshings does not depend on the prescribed tolerance TOL, which will be the case in the numerical experiments.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Thus, in order to satisfy (5.1) we require that, for all $n \geq 1$, the error indicator in space is such that

$$\frac{0.875^2\ TOL^2}{2} \int_{t^{n-1}}^{t^n} \int_\Omega |\nabla u_{h\tau}|^2\ dx\ dt \leq \sum_{K \in \mathcal{T}_h} \left(\eta^A_{K,n}(u_{h\tau})\right)^2$$
$$\leq \frac{1.125^2\ TOL^2}{2} \int_{t^{n-1}}^{t^n} \int_\Omega |\nabla u_{h\tau}|^2\ dx\ dt, \tag{5.2}$$

and, for all $n \geq 2$, the error indicator in time is such that

$$\frac{0.875^2\ \ TOL^2}{2} \int_{t^{n-1}}^{t^n} \int_\Omega |\nabla u_{h\tau}|^2\ dx\ dt \leq \sum_{K \in \mathcal{T}_h} \left(\tilde{\eta}^T_{K,n}(u_{h\tau})\right)^2$$
$$\leq \frac{1.125^2\ TOL^2}{2} \int_{t^{n-1}}^{t^n} \int_\Omega |\nabla u_{h\tau}|^2\ dx\ dt. \tag{5.3}$$

Note that we do not take into account the error in time corresponding to the first time step.

If (5.2) is not satisfied, the BL2D anisotropic mesh generator [BL96] is invoked to construct another mesh based on the space error indicator η^A. The P_1 interpolation between the previous mesh $\mathcal{T}_h^{n-1}$ and the new mesh $\mathcal{T}_h^n$ is carried out by the BL2D mesh generator. Thus, BL2D provides us an interpolated solution $r_h^n(u_h^{n-1})$ of u_h^{n-1} on the new mesh, where r_h^n is the Lagrange interpolant operator on $\mathcal{T}_h^n$. Then, after each remeshing, we seek $u_h^n \in V_h^0$ such that $\forall v_h \in V_h^0$,

$$\int_\Omega \frac{u_h^n - r_h^n(u_h^{n-1})}{\tau_n} v_h\ dx + \frac{1}{2} \int_\Omega \left(\nabla u_h^n + \nabla r_h^n(u_h^{n-1})\right) \cdot \nabla v_h\, dx = \frac{1}{2} \int_\Omega (f^n + f^{n-1}) v_h\, dx. \tag{5.4}$$

Since BL2D requires the data to be given at the mesh vertices rather than triangles, we calculate at each vertex P the anisotropic error estimator defined as

$$\eta^A_{i,P,n}\left(u_{h\tau}\right) = \left(\sum_{\substack{K \in \mathcal{T}_h \\ P \in K}} \left(\eta^A_{i,K,n}(u_{h\tau})\right)^4 \right)^{1/4},$$

where $\eta^A_{i,K,n}$ is the local error estimator on triangle K in the direction $r_{i,K}$ defined as

$$\left(\eta^A_{i,K,n}(u_{h\tau})\right)^2 = \int_{t^{n-1}}^{t^n} \left(\|f - \partial_n u_h + \Delta u_{h\tau}\|_{L^2(K)} + \frac{1}{2\lambda_{2,K}^{1/2}} \|[\nabla u_{h\tau} \cdot \mathbf{n}]\|_{L^2(\partial K)} \right)$$
$$\times \left(\lambda^2_{i,K} \left(r^T_{i,K} \breve{G}_K(u_{h\tau}) r_{i,K} \right) \right)^{1/2} dt$$

and a gradient matrix $\breve{G}_{P,n}$ such that

$$\breve{G}_{P,n}(u_{h\tau}) = \sum_{\substack{K \in \mathcal{T}_h \\ P \in K}} \left(\left(\eta^A_{1,K,n}(u_{h\tau})\right)^4 + \left(\eta^A_{2,K,n}(u_{h\tau})\right)^4 \right) \breve{G}_K(u_{h\tau}).$$

The eigenvectors of matrix $\breve{G}_P$ provide the desired direction of the anisotropy at vertex P, and the directional error estimators $\eta^A_{i,P,n}$ are used to prescribe the new

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Set $\mathcal{T}_h^0$, u_h^0, $n = 1$, $t = 0$	Initializations
Do while $t < T$	Time loop
$t := t + \tau_n$	Increment the current time step
Compute u_h^n on mesh $\mathcal{T}_h^{n-1}$	
Do for all triangles K of $\mathcal{T}_h^{n-1}$	
Compute $r_{1,K}$, $r_{2,K}$, $\lambda_{1,K}$, $\lambda_{2,K}$	Directions and amplitudes of stretching
Compute $\breve{G}_K$	Approximated error gradient matrix
Compute $\breve{\eta}_{K,n}^T$	Time error estimator
Compute $\eta_{K,n}^A$	Space error estimator
End Do	
If (5.2) and (5.3) are satisfied	The mesh and the time step are correct
$\mathcal{T}_h^n := \mathcal{T}_h^{n-1}$	Same mesh
$n := n + 1$	
Else	
If (5.2) is not satisfied	Mesh adaptation
Do for all vertices P	
Compute $\breve{G}_{P,n}$, $\eta_{1,P,n}^A$, $\eta_{2,P,n}^A$	Averaged values on vertices
Set directions of anisotropy $r_{1,P}$ and $r_{2,P}$ to eigenvectors of $\breve{G}_P$	
If $\eta_{1,P,n}^A$ is too small (resp. too large) the mesh size in the first direction of anisotropy should be increased (resp. decreased)	Coarsening (resp. refinement) in the direction $r_{1,P}$.
If $\eta_{2,P,n}^A$ is too small (resp. too large) the mesh size in the second direction of anisotropy should be increased (resp. decreased)	Coarsening (resp. refinement) in the direction $r_{2,P}$
End Do	
Build new anisotropic mesh $\mathcal{T}_h^n$ using BL2D	
If (5.3) is not satisfied	Time adaptation
If $\sum_{K\in\mathcal{T}_h} \left(\eta_{K,n}^T(u_{h\tau})\right)^2$ is too small (resp. too large) τ_n should be increased (resp. decreased)	
$\mathcal{T}_h^{n-1} := \mathcal{T}_h^n$	
$t := t - \tau_n$	
End If	
End Do	

FIG. 5.1. *Adaptive algorithm.*

stretching amplitudes. In practice, the error is equidistributed in the two directions of maximum and minimum stretching. This ensures conditions (4.7) and (4.9) to be fulfilled with $c = 1$. Once an acceptable mesh has been constructed, we check (5.3) and modify the current time step if necessary. Our adaptive algorithm is summarized in Figure 5.1. More technical details are available in [Pic03] in a similar context. Note that we use the same algorithm with (4.17) and (4.18).

In order to monitor the mesh anisotropy, we define the maximum aspect ratio and the mean aspect ratio respectively defined by

$$ar = \max_{K\in\mathcal{T}_h} \frac{\lambda_{1,K}}{\lambda_{2,K}} \quad \text{and} \quad \overline{ar} = \frac{\displaystyle\sum_{K\in\mathcal{T}_h} \frac{\lambda_{1,K}}{\lambda_{2,K}}}{\displaystyle\sum_{K\in\mathcal{T}_h} 1}.$$

6. A numerical study of the adaptive algorithm. We apply here the adaptive algorithm described in Figure 5.1 to two test cases exhibiting increasing levels of

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

TABLE 6.1
Example 6.1. *Convergence results using uniform isotropic meshes and constant time steps.*

h	τ	$\left(\int_0^T \int_\Omega \lvert\nabla e\rvert^2\right)^{1/2}$	ei^{ZZ}	ei^{A}	$\widetilde{ei}^T$	$\widehat{ei}^T$
0.025	0.05	0.22	0.99	2.50	17.51	14.32
0.00625	0.025	0.055	0.99	2.50	18.15	14.79
0.0015625	0.0125	0.014	1.00	2.52	18.28	14.93
0.025	0.0125	0.22	1.04	2.59	1.15	0.94
0.00625	0.00625	0.056	1.00	2.52	1.14	0.93
0.0015625	0.003125	0.014	1.00	2.52	1.14	0.93

anisotropy. For the first test case we start all our adaptive simulations on an isotropic 10×10 mesh and for the second test case with an anisotropic 100×2 mesh. Moreover, we choose the time step $\tau_1 = 0.05$. We monitor the absolute error ϵ_{abs} in the $L^2(0,T;H^1(\Omega))$ norm, the relative error ϵ_{rel} in the same norm, the number of nodes nb_n, maximum, and mean aspect ratio, all computed for the mesh at the final time T. We also report the number of time steps nb_τ needed to reach the final time and the number of remeshings nb_m occurred.

Example 6.1. Set $\Omega = (0,1) \times (0,1)$, the unit square, $T = 1$, and choose u^0 and f so that the solution u of (2.1) is given by

$$u(x,y,t) = e^{-100r^2(x,y,t)}, \tag{6.1}$$

where

$$r^2(x,y,t) = (x - 0.3 - 0.4\,\beta(t))^2 + (y - 0.3 - 0.4\,\beta(t))^2$$

and

$$\beta(t) = 0.5 + 0.5 \tanh\left(\frac{t - 0.5}{0.2}\right).$$

Thus, u is a Gaussian function, whose center moves from point $(0.3, 0.3)$ at $t = 0$ to point $(0.7, 0.7)$ at $t = 1$ with a transport velocity $0.4\beta'(t)(1,1)^T$ peaking at $t = 0.5$. Before starting our adaptive algorithm, we first want to study the effectivity indices corresponding to the two time error indicators $\widehat{\eta}^T$ and $\widetilde{\eta}^T$ when using uniform time steps and mesh size. In order to reduce the error due to the first time step, we replace ∇u_h^0, in (2.4), by $r_h \nabla u^0$. We have reported in Table 6.1 the results obtained when $h = 10\tau^2$ and $h = 160\tau^2$. We can observe that the error in the $L^2(0,T;H^1(\Omega))$ norm is $O(h + \tau^2)$, that the space error estimator and the two time error estimators, $\widehat{\eta}^T$ and $\widetilde{\eta}^T$, are equivalent to the true error as their effectivity indices tend to a constant value, and that the Zienkiewicz–Zhu error estimator is asymptotically exact. Moreover, when $h = 10\tau^2$, the time effectivity indices $\widetilde{ei}^T$ and $\widehat{ei}^T$ are similar to those of Table 4.1. On the other hand, when $h = 160\tau^2$ the error due to the time discretization is divided by 16 so as the time effectivity indices $\widetilde{ei}^T$ and $\widehat{ei}^T$. In Figure 6.1, the values of w_h^n and $\partial_n^2 u_h$ are compared to $\partial^2 u/\partial t^2$ at time $t = 0.5$ along the axis $y = x$ and with $h = \tau = 0.00625$. We observe that $\partial_n^2 u_h$ provides a smoother approximation of $\partial^2 u/\partial t^2$ than w_h^n. Indeed, we can notice slight oscillations when approaching $\partial^2 u/\partial t^2$ with w_h^n.

We now use the adaptive algorithm with first the three-point time error estimator (4.19) and the anisotropic space error estimator (4.17). We have reported in

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

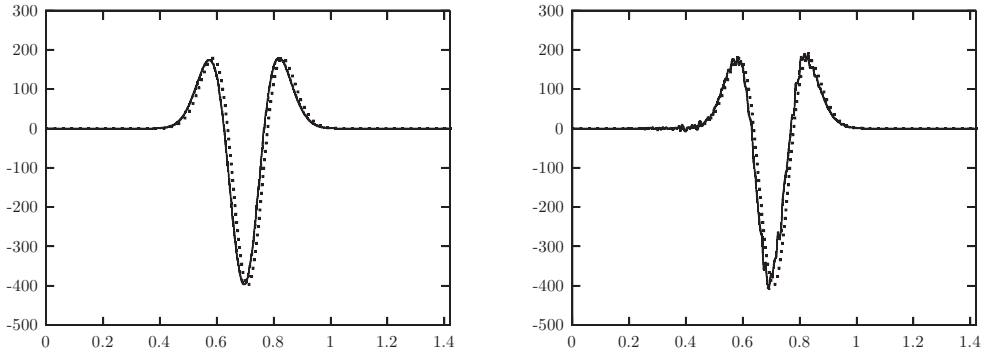


FIG. 6.1. *Example* 6.1. $\partial_n^2 u_h$ *(solid line, left) and and* w_h^n *(solid line, right) compared to* $\partial^2 u/\partial t^2$ *(dotted line) at time* $t = 0.5$ *and with* $h = \tau = 0.00625$.

TABLE 6.2

Example 6.1. *True error and effectivity indices of the adapted solution at final time* $T = 1$. *The three-point time error estimator* (4.19) *is used. Top: solving* (2.4) *if* $\mathcal{T}_h^{n-1} = \mathcal{T}_h^n$ *and* (5.4) *if* $\mathcal{T}_h^{n-1} \neq \mathcal{T}_h^n$. *Bottom: solving* (6.2) *if* $\mathcal{T}_h^{n-1} = \mathcal{T}_h^n$ *and* (6.3) *if* $\mathcal{T}_h^{n-1} \neq \mathcal{T}_h^n$.

TOL	ϵ_{rel}	ϵ_{abs}	ei^{ZZ}	ei^A	$\widetilde{ei}^T$	nb_n	nb_τ	nb_m	$\overline{ar}$	ar
0.25	0.085	0.15	0.88	2.11	2.07	854	48	18	2.41	8.34
0.125	0.043	0.76	0.88	2.12	2.02	3232	67	22	2.53	10.59
0.0625	0.022	0.038	0.88	2.11	2.02	12610	96	27	2.78	18.49
0.03125	0.011	0.019	0.87	2.09	1.95	46253	136	31	2.78	18.58
0.25	0.078	0.134	0.99	2.37	2.30	853	48	18	2.42	9.84
0.125	0.039	0.066	0.99	2.38	2.28	3189	67	23	2.54	10.47
0.0625	0.02	0.034	1.00	2.38	2.23	11714	96	27	2.79	14.82
0.03125	0.01	0.017	1.00	2.38	2.24	48062	135	30	2.92	18.72

Table 6.2 some numerical results with several values of tolerance TOL. We observe in Table 6.2 (top) that the error is divided by two each time the tolerance TOL is, and that both the time error indicator $\widetilde{\eta}^T$ and the space error indicator η^A are of optimal order. Indeed, we note that $\widetilde{\eta}^T$ is of optimal order as the number of time steps, nb_τ, is approximatively multiplied by $\sqrt{2}$ when TOL is divided by two and that η^A is also of optimal order as the number of nodes, nb_n, at final time, is approximatively multiplied by four when TOL is divided by two. Moreover, both the time and space effectivity indices tend to a constant value, which shows that $\widetilde{\eta}^T$ and η^A are equivalent to the true error. However, we remark that the Zienkiewicz–Zhu error estimator is not asymptotically exact. We have a value of ei^{ZZ} around 0.88, whereas we expect this value to be close to one. This discrepancy can be attributed to the interpolation error between two successive meshes that is not taken into account in our theoretical estimates. Therefore, in order to recover its asymptotical convergence we decide to replace, if $\mathcal{T}_h^{n-1}$ is identical to $\mathcal{T}_h^n$, ∇u_h^{n-1}, in (2.4), by its Z-Z recovery, $\Pi_h^n\left(\nabla(u_h^{n-1})\right)$ and if $\mathcal{T}_h^{n-1}$ is different from $\mathcal{T}_h^n$, $\nabla r_h^n(u_h^{n-1})$, in (5.4), by its Z-Z recovery, $\Pi_h^n\left(\nabla r_h^n(u_h^{n-1})\right)$. Thus, if $\mathcal{T}_h^{n-1}$ and $\mathcal{T}_h^n$ are identical, we seek $u_h^n \in V_h^0$ such that $\forall v_h \in V_h^0$,

(6.2)
$$\int_\Omega \frac{u_h^n - u_h^{n-1}}{\tau_n} v_h \, dx + \frac{1}{2}\int_\Omega \left(\nabla u_h^n + \Pi_h^n\left(\nabla(u_h^{n-1})\right)\right)\cdot \nabla v_h \, dx = \frac{1}{2}\int_\Omega (f^n + f^{n-1}) v_h \, dx,$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

and if $\mathcal{T}_h^{n-1}$ and $\mathcal{T}_h^n$ are different, we seek $u_h^n \in V_h^0$ such that $\forall v_h \in V_h^0$,

$$\int_\Omega \frac{u_h^n - r_h^n(u_h^{n-1})}{\tau_n} v_h \, dx + \frac{1}{2}\int_\Omega \left(\nabla u_h^n + \Pi_h^n\left(\nabla r_h^n(u_h^{n-1})\right)\right)\cdot \nabla v_h \, dx = \frac{1}{2}\int_\Omega (f^n + f^{n-1}) v_h \, dx. \tag{6.3}$$

We have reported in Table 6.2 (bottom) the corresponding results. We now observe that the Zienkiewicz–Zhu error estimator is asymptotically exact as its effectivity index is close to one. All the previous observations remain unchanged, especially the good behavior of both the time and space error indicators. We have reported in Figure 6.2 the corresponding numerical simulation when the tolerance $TOL = 0.125$.

In Figure 6.3, we plot the evolution of the number of nodes and time step size corresponding to the numerical simulation presented in Figure 6.2. We observe that the number of nodes remains almost constant whereas the time step size decreases until $t = 0.5$ and increases until the final time $T = 1$. The evolution of the time step size thus fits the transport velocity. The three-point time error estimator $\widetilde{\eta}^T$ seems then to be a good approximation of the true error. Finally, we plot in Figure 6.4 the value of $\partial_n^2 u_h$ when solving (2.4) if $\mathcal{T}_h^{n-1} = \mathcal{T}_h^n$ and (5.4) if $\mathcal{T}_h^{n-1} \neq \mathcal{T}_h^n$ and when solving (6.2) if $\mathcal{T}_h^{n-1} = \mathcal{T}_h^n$ and (6.3) if $\mathcal{T}_h^{n-1} \neq \mathcal{T}_h^n$ compared to $\partial^2 u/\partial t^2$ at time $t = 0.5$ along the axis $y = x$ and when $TOL = 0.125$. We note a better approximation of $\partial^2 u/\partial t^2$ when solving (6.2) if $\mathcal{T}_h^{n-1} = \mathcal{T}_h^n$ and (6.3) if $\mathcal{T}_h^{n-1} \neq \mathcal{T}_h^n$.

We now use the time error estimator (4.18) instead of (4.19) in the adaptive algorithm, and we report in Table 6.3 (top) some numerical results with several values of the tolerance TOL. The results obtained show that the good behavior observed in the case of constant time and space steps is not preserved when using our adaptive algorithm. Indeed, we can see that the time error indicator $\widehat{\eta}^T$ is not of optimal order of convergence since the number of time steps, nb_τ, is not multiplied by $\sqrt{2}$ when we divide TOL by two. For a tolerance $TOL = 0.125$, the number of the time steps was 67 when using the time error indicator $\widetilde{\eta}^T$, whereas it becomes 1860 for the time error indicator $\widehat{\eta}^T$. This significant difference shows that $\widehat{\eta}^T$ tends to dramatically over-predict the true error. In Figure 6.5 (left) we plot the evolution of the time step size when $TOL = 0.125$. We can see that the time step evolution is completely different and shows an irregular progression with very small time steps around a value of 0.005.

In order to reduce this overprediction, we decide to solve (6.2) if $\mathcal{T}_h^{n-1} = \mathcal{T}_h^n$ and (6.3) if $\mathcal{T}_h^{n-1} \neq \mathcal{T}_h^n$. We have reported in Table 6.3 (bottom) the numerical results for several values of TOL. We see that the number of time steps, nb_τ, is not multiplied by $\sqrt{2}$ when we divide TOL by two so the optimal order of convergence is not recovered. We plot in Figure 6.5 (right) the evolution of the time step size when $TOL = 0.125$. We note the same irregular profile, but its general aspect seems to be closer to the one presented in Figure 6.3. Indeed, neglecting these oscillations we roughly have the same evolution as in Figure 6.3. We suspect that this irregular profile is due to the interpolation error after each remeshing. Thus, we introduce the relative two-point time error estimator $\widehat{\eta}_{rel}^n$ and the relative error ϵ_{rel}^n defined, respectively, by

$$\widehat{\eta}_{rel}^n = \frac{\sum_{K\in\mathcal{T}_h}\left(\widehat{\eta}_{K,n}^T(u_{h\tau})\right)^2}{\int_{t^{n-1}}^{t^n}\int_\Omega |\nabla u_{h\tau}|^2 \, dx \, dt} \quad \text{and} \quad \epsilon_{rel}^n = \frac{\int_{t^{n-1}}^{t^n}\int_\Omega |\nabla e|^2 \, dx \, dt}{\int_{t^{n-1}}^{t^n}\int_\Omega |\nabla u_{h\tau}|^2 \, dx \, dt}, \tag{6.4}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

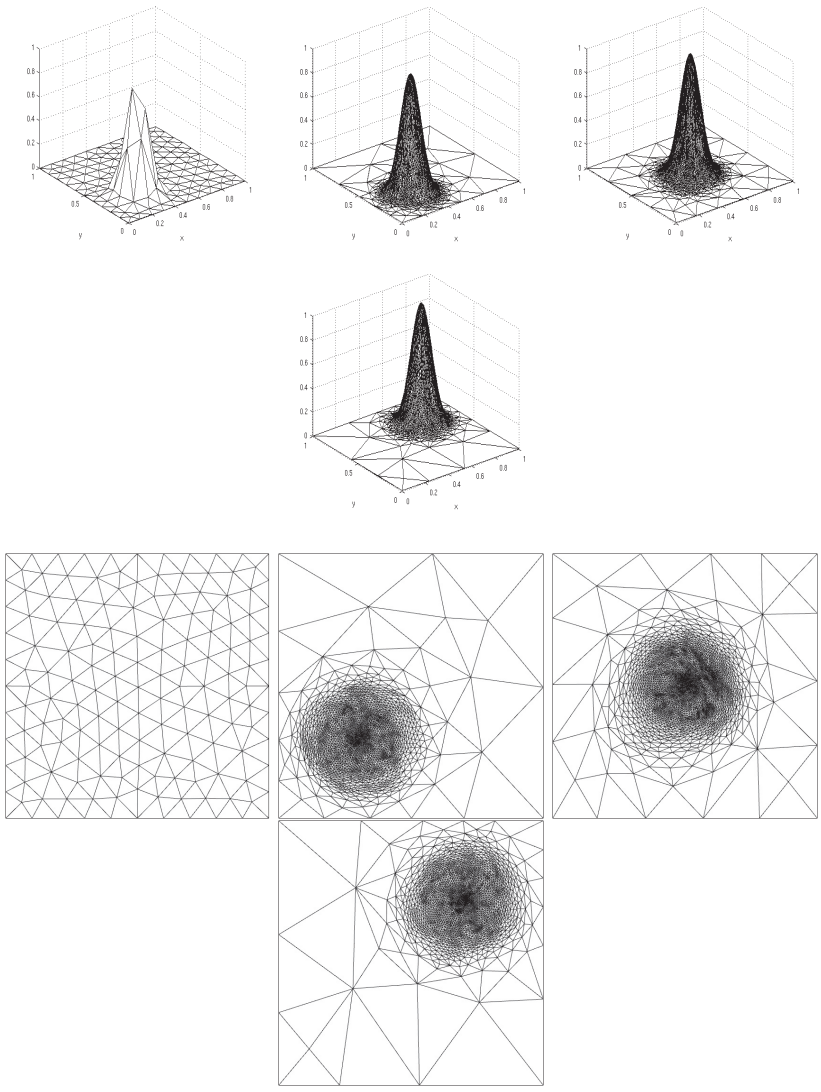


FIG. 6.2. *Example* 6.1. *Adapted meshes and solutions obtained with the time error estimator* (4.19) *and* $TOL = 0.125$ *when solving* (6.2) *if* $\mathcal{T}_h^{n-1} = \mathcal{T}_h^n$ *and* (6.3) *if* $\mathcal{T}_h^{n-1} \neq \mathcal{T}_h^n$. *From left to right: time* $t = 0$, 0.05, 0.5, *and* 1 *(*151, 2674, 3333, *and* 3189 *nodes, respectively).*

and we plot in Figure 6.6 the evolution of these two quantities for a tolerance $TOL = 0.125$ when solving (6.2) if $\mathcal{T}_h^{n-1} = \mathcal{T}_h^n$ and (6.3) if $\mathcal{T}_h^{n-1} \neq \mathcal{T}_h^n$. We can see that each remeshing results in a jump of the two-point time error estimator that causes the irregular evolution. In Figure 6.7 we plot w_h^n compared to $\partial^2 u/\partial t^2$ at time $t = 0.5$ along the axis $y = x$ and when $TOL = 0.125$. Two implementations of our adaptive algorithm are reported here: first when solving (2.4) if $\mathcal{T}_h^{n-1} = \mathcal{T}_h^n$ and (5.4) if $\mathcal{T}_h^{n-1} \neq \mathcal{T}_h^n$, and second when solving (6.2) if $\mathcal{T}_h^{n-1} = \mathcal{T}_h^n$ and (6.3) if $\mathcal{T}_h^{n-1} \neq \mathcal{T}_h^n$. We can see that w_h^n does not at all approach $\partial^2 u/\partial t^2$ if the Zienkiewicz–Zhu recovery method is not used. The huge oscillations of w_h^n observed in this case can explain why the time error indicator $\widehat{\eta}^T$ extensively overpredicts the true error. A much better behavior

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

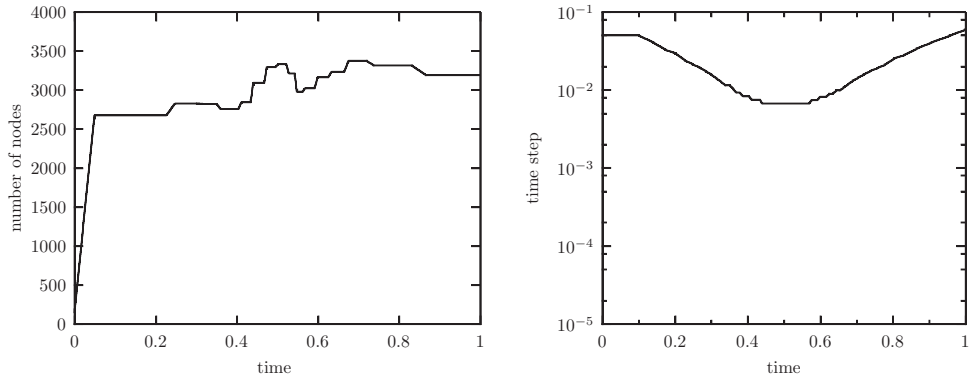


FIG. 6.3. *Example* 6.1. *Number of nodes (left) and time step size (right) with respect to time* t *when solving* (6.2) *if* $\mathcal{T}_h^{n-1} = \mathcal{T}_h^n$ *and* (6.3) *if* $\mathcal{T}_h^{n-1} \neq \mathcal{T}_h^n$. *The time error estimator is given by* (4.19) *and* $TOL = 0.125$.

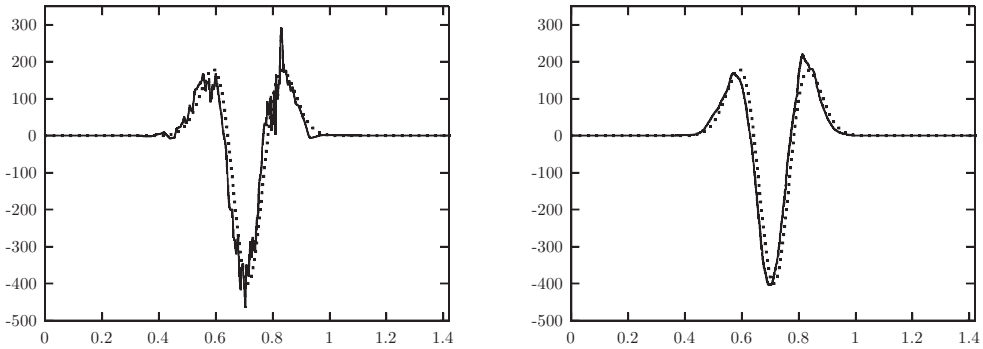


FIG. 6.4. *Example* 6.1. $\partial_n^2 u_h$ *(solid line) compared to* $\partial^2 u/\partial t^2$ *(dotted line) at time* $t = 0.5$ *and when* $TOL = 0.125$. *Left: We solve* (2.4) *if* $\mathcal{T}_h^{n-1} = \mathcal{T}_h^n$ *and* (5.4) *if* $\mathcal{T}_h^{n-1} \neq \mathcal{T}_h^n$. *Right: We solve* (6.2) *if* $\mathcal{T}_h^{n-1} = \mathcal{T}_h^n$ *and* (6.3) *if* $\mathcal{T}_h^{n-1} \neq \mathcal{T}_h^n$.

TABLE 6.3

Example 6.1. *True error and effectivity indices of the adapted solution at final time* $T = 1$. *The two-point time error estimator* (4.18) *is used. Top: Solving* (2.4) *if* $\mathcal{T}_h^{n-1} = \mathcal{T}_h^n$ *and* (5.4) *if* $\mathcal{T}_h^{n-1} \neq \mathcal{T}_h^n$. *Bottom: Solving* (6.2) *if* $\mathcal{T}_h^{n-1} = \mathcal{T}_h^n$ *and* (6.3) *if* $\mathcal{T}_h^{n-1} \neq \mathcal{T}_h^n$.

TOL	ϵ_{rel}	ϵ_{abs}	ei^{ZZ}	ei^{A}	$\widehat{ei}^{T}$	nb_n	nb_τ	nb_m	$\overline{ar}$	ar
0.25	0.074	0.128	1.00	2.39	2.25	885	620	15	2.38	9.52
0.125	0.038	0.066	1.00	2.40	2.14	2871	1860	19	2.70	16.45
0.0625	0.019	0.033	1.00	2.40	2.20	11924	4882	22	2.71	12.68
0.03125	0.01	0.017	1.00	2.39	2.20	48038	23851	26	2.95	19.51
0.25	0.078	0.133	1.00	2.40	2.28	852	71	15	2.44	10.66
0.125	0.038	0.066	1.00	2.41	2.20	3125	127	18	2.67	12.53
0.0625	0.019	0.033	1.00	2.41	2.17	12347	260	23	2.89	16.86
0.03125	0.0096	0.0166	1.00	2.40	2.18	47513	1695	25	3.01	20.57

of w_h^n is observed if we solve (6.2) when $\mathcal{T}_h^{n-1} = \mathcal{T}_h^n$ and (6.3) when $\mathcal{T}_h^{n-1} \neq \mathcal{T}_h^n$. However, this approximation also suffers from important spurious oscillations, and it is still not good enough to recover the optimal order of convergence of the time error estimator $\widehat{\eta}^T$.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

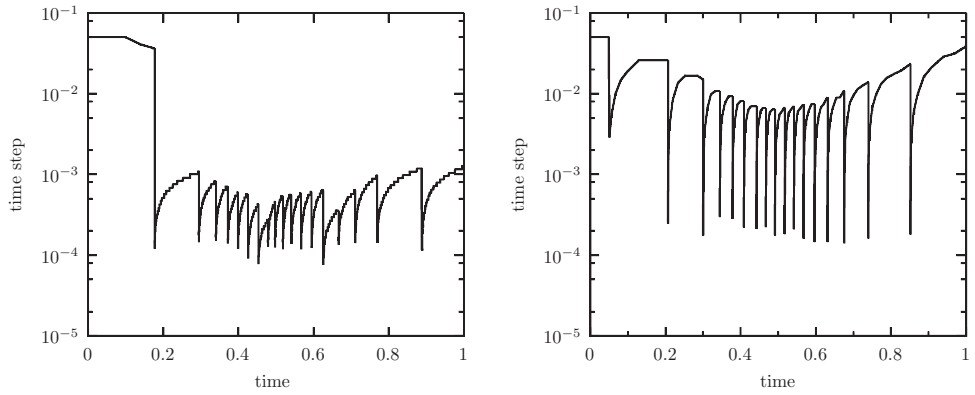


FIG. 6.5. *Example* 6.1. *Time step size with respect to time t when solving* (2.4) *if* $\mathcal{T}_h^{n-1} = \mathcal{T}_h^n$ *and* (5.4) *if* $\mathcal{T}_h^{n-1} \neq \mathcal{T}_h^n$ *(left) and when solving* (6.2) *if* $\mathcal{T}_h^{n-1} = \mathcal{T}_h^n$ *and* (6.3) *if* $\mathcal{T}_h^{n-1} \neq \mathcal{T}_h^n$ *(right). The time error estimator is given by* (4.18) *and* $TOL = 0.125$.

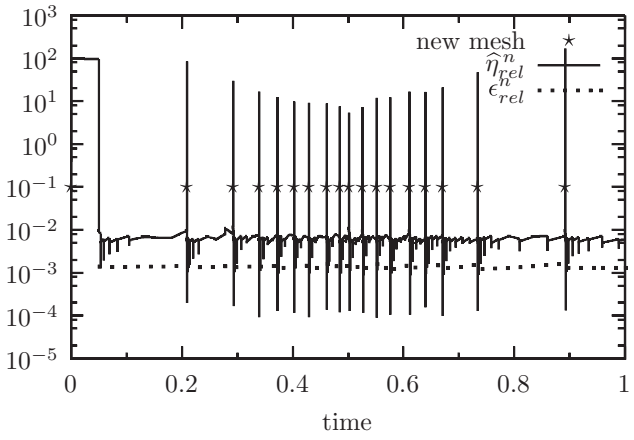


FIG. 6.6. *Example* 6.1. *Evolution of* $\widehat{\eta}_{rel}^n$ *and* ϵ_{rel}^n *when solving* (6.2) *if* $\mathcal{T}_h^{n-1} = \mathcal{T}_h^n$ *and* (6.3) *if* $\mathcal{T}_h^{n-1} \neq \mathcal{T}_h^n$. *The time error estimator is given by* (4.18) *and* $TOL = 0.125$.

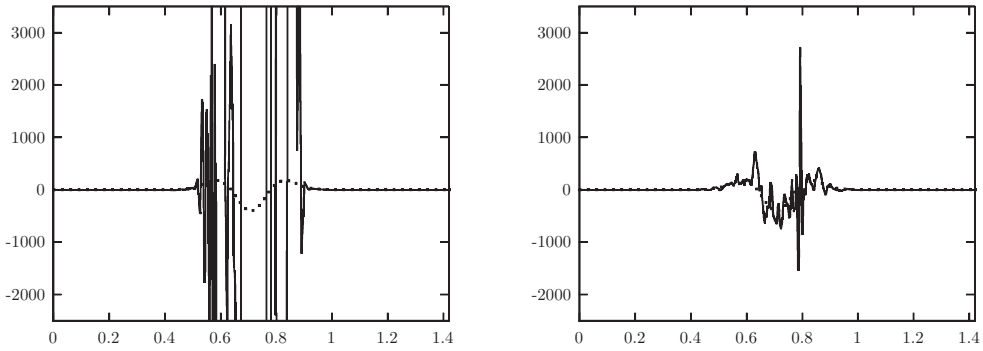


FIG. 6.7. *Example* 6.1. w_h^n *(solid line) compared to* $\partial^2 u/\partial t^2$ *(dotted line) at time* $t = 0.5$ *and when* $TOL = 0.125$. *Left: We solve* (2.4) *if* $\mathcal{T}_h^{n-1} = \mathcal{T}_h^n$ *and* (5.4) *if* $\mathcal{T}_h^{n-1} \neq \mathcal{T}_h^n$. *Right: We solve* (6.2) *if* $\mathcal{T}_h^{n-1} = \mathcal{T}_h^n$ *and* (6.3) *if* $\mathcal{T}_h^{n-1} \neq \mathcal{T}_h^n$.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

We now consider another test case with more anisotropic finite elements. This example illustrates the situation where the solution varies in only one direction and thus leads to very stretched meshes. We do all the computations by replacing $\nabla(u_h^{n-1})$ and $\nabla r_h^n(u_h^{n-1})$ in (2.4) and in (5.4), , respectively, by its Zienkiewicz–Zhu recovery, respectively, $\Pi_h^n\left(\nabla(u_h^{n-1})\right)$ and $\Pi_h^n\left(\nabla r_h^n(u_h^{n-1})\right)$.

Example 6.2. Set $\Omega = (0,1)\times(0,1)$, the unit square, $T = 1.5$, and choose u^0 and f so that the solution u of (2.1) is given by

$$u(x,y,t) = 0.5 + 0.5\tanh\left(\frac{x - 0.2 - 0.3\ \beta_2(t)}{0.05}\right), \tag{6.5}$$

where

$$\beta_2(t) = \begin{cases} 0.5 + 0.5\tanh\left(\dfrac{t-0.4}{0.05}\right) & \text{if } t \le 0.75, \\ 1.5 + 0.5\tanh\left(\dfrac{t-1.1}{0.05}\right) & \text{if } t > 0.75. \end{cases}$$

Note that we consider here a problem with mixed Dirichlet–Neumann boundary condition. Indeed, we impose $u = 0$ along the left side of Ω, $u = 1$ along the right side, and impose homogeneous Neumann boundary conditions along the top and bottom sides. Thus, the solution u exhibits an internal layer moving with a velocity $0.3\beta_2'(t)$ having sharp peaks centered at times $t = 0.4$ and $t = 1.1$. We apply the adaptive algorithm with the three-point time error estimator $\widetilde{\eta}^T$. Adapted meshes are shown in Figure 6.8 at time $t = 0, 0.05, 1$, and 1.5 and with a tolerance $TOL = 0.03125$. We plot in Figure 6.9 (left) the corresponding time step size evolution. We see that the time step size follows the velocity profile of the solution. Thus, for high velocity the adaptive algorithm chooses to use small time step and vice-versa. In Table 6.4 (top) we report numerical experiments with several values of the tolerance TOL. We observe that η^{ZZ} is asymptotically exact and that the space anisotropic error estimator η^A and the three-point time error estimator $\widetilde{\eta}^T$ are equivalent to the true error. The number of time steps is approximatively multiplied by $\sqrt{2}$ when TOL is divided by two so that the optimal order is recovered.

We now use (4.18) instead of (4.19) in the adaptive algorithm. Since the Dirichlet boundary conditions are prescribed only on the part of the boundary in this example, the two-point time error estimator, namely, the definition of w_h^n (4.4), should be changed. We require here w_h^n to vanish only along the left and right sides of Ω, and we impose homogeneous Neumann boundary conditions along the top and bottom sides. The evolution of the time step size is reported in Figure 6.9 (right) for the same simulation as in Figure 6.8. A chaotic evolution is observed. Indeed, we see many perturbations, of which the most important of them appear during the two peaks of acceleration and deceleration. We can also note that between the two peaks, when the velocity is very small, the time step size stays approximatively constant around a value of 0.02 whereas we expect an increase. For solutions exhibiting high aspect ratio, the behavior of the two-point time error estimator seems then not to be a really good representation of the true error. In Table 6.4 (bottom) we report some numerical results with several values of tolerance. We see that the two-point time error estimator is not of optimal order as the number of time iterations grows when the tolerance is divided by two.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

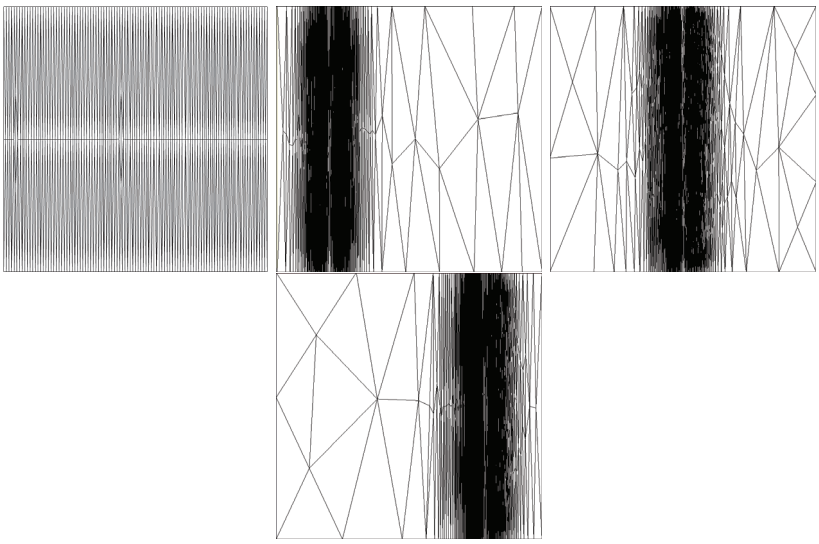


FIG. 6.8. *Example* 6.2. *Adapted meshes obtained with the time error estimator* (4.19) *and the tolerance* $TOL = 0.03125$. *From left to right: time* $t = 0$, 0.05, 1, *and* 1.5 *(*320, 1042, 1015, *and* 892 *nodes, respectively).*

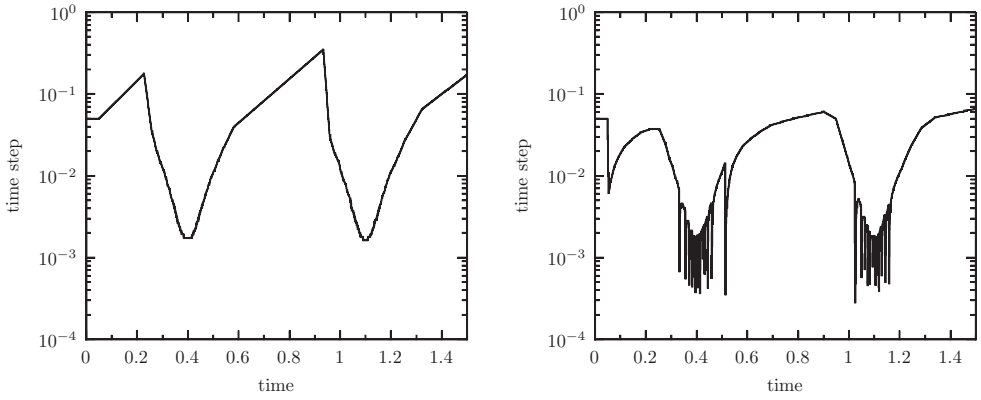


FIG. 6.9. *Example* 6.2. *Time step size with respect to time* t *with* $TOL = 0.03125$. *Left: The time error estimator is given by* (4.19). *Right: The time error estimator is given by* (4.18).

TABLE 6.4

Example 6.2. *True error and effectivity indices of the adapted solution at final time* $T = 1.5$. *Top: The time error estimator* (4.19) *is used. Bottom: The time error estimator* (4.18) *is used.*

TOL	ϵ_{rel}	ϵ_{abs}	ei^{ZZ}	ei^{A}	$\widetilde{ei}^{T}$	nb_n	nb_τ	nb_m	$\overline{ar}$	ar
0.125	0.03	0.095	0.99	2.87	2.68	155	142	84	63	231
0.0625	0.015	0.046	0.99	2.89	2.91	348	201	52	108	519
0.03125	0.007	0.022	0.99	2.96	2.99	892	285	52	165	705
0.015625	0.004	0.012	1.00	2.88	2.71	4408	401	40	118	847
0.125	0.03	0.09	1.01	2.93	2.85	163	187	81	69	305
0.0625	0.015	0.046	1.00	2.98	2.83	345	301	57	109	479
0.03125	0.0075	0.023	1.00	2.95	2.96	892	909	52	169	948
0.015625	0.0037	0.011	1.00	2.89	2.95	6054	10147	38	86	904

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

7. Conclusion. Numerical experiments show that $\widetilde{\eta}^T$ is a good representation of the true error even with solutions exhibiting high aspect ratio. Indeed, the expected second order of convergence with respect to the time step has been recovered experimentally for several test cases. However, for the time error estimator $\widehat{\eta}^T$, we did not manage to recover experimentally the optimal second order of convergence. The difficulty in approximating ∇w_h^n involved in $\widehat{\eta}^T$ seems to be the major problem of this approach. We have extended these results to other problems such as the time-dependent (convection dominated) convection diffusion [PP08] and simulation of electroosmotic flows in microfluidic devices [PPG09].

REFERENCES

[AL96] T. Apel and G. Lube, *Anisotropic mesh refinement in stabilized Galerkin methods*, Numer. Math., 74 (1996), pp. 261–282.

[AMN06] G. Akrivis, C. Makridakis, and R. H. Nochetto, *A posteriori error estimates for the Crank-Nicolson method for parabolic equations*, Math. Comp., 75 (2006), pp. 511–531.

[AMN07] G. Akrivis, C. Makridakis, and R. H. Nochetto, *Optimal order a posteriori error estimates for a class of runge-kutta and galerkin methods*, in preparation, 2007.

[AN98] T. Apel and S. Nicaise, *The finite element method with anisotropic mesh grading for elliptic problems in domains with corners and edges*, Math. Methods Appl. Sci., 21 (1998), pp. 519–549.

[AO93] M. Ainsworth and J. T. Oden, *A unified approach to a posteriori error estimation using element residual methods*, Numer. Math., 65 (1993), pp. 23–50.

[AO00] M. Ainsworth and J. T. Oden, *A Posteriori Error Estimation in Finite Element Analysis*, Pure and Applied Mathematics, Wiley-Interscience, John Wiley & Sons, New York, 2000.

[Ape99] T. Apel, *Anisotropic Finite Elements: Local Estimates and Applications*, Adv. Numer. Math., B. G. Teubner, Stuttgart, 1999.

[BBM05] A. Bergam, C. Bernardi, and Z. Mghazli, *A posteriori analysis of the finite element discretization of some parabolic equations*, Math. Comp., 74 (2005), pp. 1117–1138.

[BJP04] E. Burman, A. Jacot, and M. Picasso, *Adaptive finite elements with high aspect ratio for the computation of coalescence using a phase-field model*, J. Comput. Phys., 195 (2004), pp. 153–174.

[BL96] H. Borouchaki and G. Laug, *The BL2D mesh generator: Beginner's guide, user's and programmer's manual*, Technical Report RT-0194, Institut National de Recherche en Informatique et Automatique (INRIA), Rocquencourt, 78153 Le Chesnay, France, 1996.

[BR03] W. Bangerth and R. Rannacher, *Adaptive Finite Element Methods for Differential Equations*, Lectures in Mathematics ETH Zürich, Birkhäuser Verlag, Basel, 2003.

[BS01] I. Babuška and T. Strouboulis, *The Finite Element Method and Its Reliability*, Numerical Mathematics and Scientific Computation, The Clarendon Press Oxford University Press, New York, 2001.

[BV00] C. Bernardi and R. Verfürth, *Adaptive finite element methods for elliptic equations with non-smooth coefficients*, Numer. Math., 85 (2000), pp. 579–608.

[Car04] C. Carstensen, *All first-order averaging techniques for a posteriori finite element error control on unstructured grids are efficient and reliable*, Math. Comp., 73 (2004), pp. 1153–1165.

[CF04] Z. Chen and J. Feng, *An adaptive finite element algorithm with reliable and efficient error control for linear parabolic problems*, Math. Comp., 73 (2004), pp. 1167–1193.

[Clé75] Ph. Clément, *Approximation by finite element functions using local regularization*, Rev. Française Automat. Informat. Recherche Opérationnelle Sér. RAIRO Analyse Numérique, 9 (1975), pp. 77–84.

[DL92] R. Dautray and J.-L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology, Vol.* 5, Springer-Verlag, Berlin, 1992. Evolution problems. I, with the collaboration of Michel Artola, Michel Cessenat, and Hélène Lanchon, and translated from the French by Alan Craig.

[EJ91] K. Eriksson and C. Johnson, *Adaptive finite element methods for parabolic problems. I. A linear model problem*, SIAM J. Numer. Anal., 28 (1991), pp. 43–77.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

[EJ95] K. Eriksson and C. Johnson, *Adaptive finite element methods for parabolic problems. IV. Nonlinear problems*, SIAM J. Numer. Anal., 32 (1995), pp. 1729–1749.

[FP01] L. Formaggia and S. Perotto, *New anisotropic a priori error estimates*, Numer. Math., 89 (2001), pp. 641–667.

[FP03] L. Formaggia and S. Perotto, *Anisotropic error estimates for elliptic problems*, Numer. Math., 94 (2003), pp. 67–92.

[Han05] W. Han, *A Posteriori Error Analysis via Duality Theory*, volume 8 of Advances in Mechanics and Mathematics, Springer-Verlag, New York, 2005. With applications in modeling and numerical approximations.

[KN03] G. Kunert and S. Nicaise, *Zienkiewicz-Zhu error estimators on anisotropic tetrahedral and triangular finite element meshes*, M2AN Math. Model. Numer. Anal., 37 (2003), pp. 1013–1043.

[Kun00] G. Kunert, *An a posteriori residual error estimator for the finite element method on anisotropic tetrahedral meshes*, Numer. Math., 86 (2000), pp. 471–490.

[Kun02] G. Kunert, *Toward anisotropic mesh construction and error estimation in the finite element method*, Numer. Methods Partial Differential Equations, 18 (2002), pp. 625–648.

[KV00] G. Kunert and R. Verfürth, *Edge residuals dominate a posteriori error estimates for linear finite element methods on anisotropic triangular and tetrahedral meshes*, Numer. Math., 86 (2000), pp. 283–303.

[NZ04] A. Naga and Z. Zhang, *A posteriori error estimates based on the polynomial preserving recovery*, SIAM J. Numer. Anal., 42 (2004), pp. 1780–1800.

[Pic98] M. Picasso, *Adaptive finite elements for a linear parabolic problem*, Comput. Methods Appl. Mech. Engrg., 167 (1998), pp. 223–237.

[Pic03] M. Picasso, *An anisotropic error indicator based on Zienkiewicz-Zhu error estimator: Application to elliptic and parabolic problems*, SIAM J. Sci. Comput., 24 (2003), pp. 1328–1355.

[Pic06] M. Picasso, *Adaptive finite elements with large aspect ratio based on an anisotropic error estimator involving first order derivatives*, Comput. Methods Appl. Mech. Engrg., 196 (2006), pp. 14–23.

[PP08] M. Picasso and V. Prachittham, *An adaptive algorithm for the Crank-Nicolson scheme applied to a time-dependent convection-diffusion problem*, J. Comput. Appl. Math., 2008, submitted.

[PPG09] V. Prachittham, M. Picasso, and M. Gijs, *Adaptive finite elements with large aspect ratio for electroosmosis and pressure-driven microflows*, Int. J. Numer. Methods Fluids, 2009, to appear.

[Rod94] R. Rodríguez, *Some remarks on Zienkiewicz-Zhu estimator*, Numer. Methods Partial Differential Equations, 10 (1994), pp. 625–635.

[Sie96] K. G. Siebert, *An a posteriori error estimator for anisotropic refinement*, Numer. Math., 73 (1996), pp. 373–398.

[Ver94] R. Verfürth, *A posteriori error estimates for nonlinear problems. Finite element discretizations of elliptic equations*, Math. Comp., 62 (1994), pp. 445–475.

[Ver96] R. Verfürth, *A Review of A Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques*, Wiley-Teubner, Leipzig, 1996.

[Ver03] R. Verfürth, *A posteriori error estimates for finite element discretizations of the heat equation*, Calcolo, 40 (2003), pp. 195–212.

[XZ04] J. Xu and Z. Zhang, *Analysis of recovery type a posteriori error estimators for mildly structured grids*, Math. Comp., 73 (2004), pp. 1139–1152.

[ZN05] Z. Zhang and A. Naga, *A new finite element gradient recovery method: Superconvergence property*, SIAM J. Sci. Comput., 26 (2005), pp. 1192–1213.

[ZZ87] O. C. Zienkiewicz and J. Z. Zhu, *A simple error estimator and adaptive procedure for practical engineering analysis*, Internat. J. Numer. Methods Engrg., 24 (1987), pp. 337–357.

[ZZ92] O. C. Zienkiewicz and J. Z. Zhu, *The superconvergent patch recovery and a posteriori error estimates. I. The recovery technique*, Internat. J. Numer. Methods Engrg., 33 (1992), pp. 1331–1364.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

SPARSE TENSOR-PRODUCT FOKKER-PLANCK-BASED METHODS FOR NONLINEAR BEAD-SPRING CHAIN MODELS OF DILUTE POLYMER SOLUTIONS

par

PASCAL DELAUNAY, ALEXEI LOZINSKI, AND ROBERT G. OWENS

paru dans

CRM Proceedings and Lecture Notes

Centre de Recherches Mathématiques
CRM Proceedings and Lecture Notes

Sparse tensor-product Fokker-Planck-based methods for nonlinear bead-spring chain models of dilute polymer solutions

Pascal Delaunay, Alexei Lozinski, and Robert G. Owens

Abstract. In this paper we present a spectral method for the solution of a high-dimensional Fokker-Planck equation describing the evolution of the configuration probability density function for a dilute solution of idealized polymer chains. The numerical method draws on ideas found in the literature on sparse grids and allows solutions to be obtained at much lower cost, for comparable accuracy, than using a full tensor-product of one-dimensional basis functions. Numerical results for homogeneous shear flow serve to illustrate the attractive features of the new method.

Keywords: *Fokker-Planck equation, Brownian dynamics, bead-spring chains, sparse tensor product.*

1. Introduction

The dimensionless equation of conservation of linear momentum and of mass for an incompressible polymeric solution may be written as

$$Re\frac{D\boldsymbol{v}}{Dt} - \beta\nabla^2\boldsymbol{v} + \nabla p = \nabla\cdot\boldsymbol{\tau}, \tag{1}$$

$$\nabla\cdot\boldsymbol{v} = 0, \tag{2}$$

where $\boldsymbol{v}$ denotes the velocity, p the pressure, Re a Reynolds number and β is a solvent to total viscosity ratio. The tensor $\boldsymbol{\tau}$ accounts for the contribution of the dissolved polymer to the Cauchy stress tensor and has traditionally been determined as the solution to a closed-form differential or integral equation. In acknowledgement of the unphysical (and, therefore, undesirable) assumptions that may be implied in modelling the elastic stress tensor of a complex fluid as the solution to a closed-form equation a more recent trend, in the so-called "micro-macro" approach [13], has been the computation of $\boldsymbol{\tau}$ from a coarse-grained kinetic theory model for the polymer conformations. Although most papers in the micro-macro rheological literature have been concerned with the solution of stochastic differential equations (see, for example, [12, 14, 19]), there is growing representation by advocates of numerical methods specially adapted to the solution of an equivalent Fokker-Planck equation [5, 6, 7, 8, 9, 17, 23, 27].

Our own recent work has been directed towards developing Fokker-Planck-based high-order methods for both dilute [4, 15] and concentrated [16] polymer solutions that are at the same time cheaper and more accurate than stochastic techniques.

©0000 (copyright holder)

The number of configuration space dimensions in all our papers to date has been low (≤ 3), however, and the cost advantage of Fokker-Planck-based methods is quickly lost once high-dimensional simulations are attempted. This is because of the so-called "curse of dimension": standard full tensor-product bases for the probability density function (pdf) ψ lead to computational effort that can grow exponentially with total dimension. Higher-order configuration spaces are encountered in bead-spring chain models of polymer solutions, for example.

Our aim in this paper is to present a suitably inexpensive numerical method for the determination of the joint probability density function ψ for the internal configurations of a bead-spring chain that Fokker-Planck-based numerical methods will become competitive with stochastic techniques for (at least some) configuration dimensions exceeding 3. Recently, Venkiteswaran and Junk [24, 25] have employed a quasi-Monte Carlo method [18] for the simulation of the Fokker-Planck equation for a dilute solution of polymers modelled by a bead-spring chain and undergoing simple shear flow. Monte Carlo methods are a well-known approach for overcoming the curse of dimension in high dimensional systems since the error estimate in terms of the number of nodes M (say) behaves like $\sqrt{M}$, independently of the dimension. Although the provable error estimate of Venkiteswaran and Junk [25] was no better than for a standard Monte Carlo method, their new method manifested less variance in the results and, for fixed accuracy, the authors achieved an improvement in the computational time over the simple Monte Carlo method. Another way of attempting to overcome the curse of dimension in the present context is by using sparse tensor-product spaces [22, 28] in order to represent ψ. In this class of methods a high-dimensional basis is derived from a one-dimensional basis by forming tensor products of univariate formulae whose indices lie in a simplex. An extensive and recent review of sparse grid methods may be found in the survey paper of Bungartz and Griebel [3]. The advantages of using a sparse grid approach are seen in the cost savings made when compared with the use of traditional full tensor-product bases for the dependent variables, without compromising on the order of accuracy. For example, whereas the number of degrees of freedom on a full tensor product finite element mesh of mesh size h in d dimensions grows like $O(h^{-d})$, that associated with a sparse tensor product only grows like $O(h^{-1}|\log h|^{d-1})$. At the same time, the order of accuracy in the H^1 norm of both methods is $O(h^p)$ for finite elements of degree $p \geq 1$ and sufficiently smooth approximated functions [26].

In Section 2 we begin by explaining how the multi-dimensional Fokker-Planck equation is derived from the bead-spring micro-mechanical model. The eigenfunctions of the Fokker-Planck operator for a single dumbbell are computed using a Galerkin spectral method in configuration space and we then employ both full and sparse tensor products of these eigenfunctions to represent the configuration pdf ψ for the full chain. Exactly how this is done is described in Section 3. An implicit stochastic algorithm is detailed in Section 4, and in the same section we explain how the elastic stress contribution to the total Cauchy stress is calculated on the basis of the solutions to the stochastic and equivalent Fokker-Planck equations. The advantages of using a sparse tensor product basis for ψ rather than the usual full tensor product for such functions is showcased in Section 5 by comparing the costs of the two approaches against that of the Brownian dynamics simulation for simple shear flow at various shear rates.

2. Micromechanical model and governing equations

The micromechanical model used to represent a single linear polymer chain in this paper consists of $(d+1)$ beads, each having the same mass m and joined consecutively by d nonlinear massless springs, as shown in Fig. 1. The equations of motion for the bead having position vector $\boldsymbol{r}_j$ $(j = 2, \ldots, d)$ may be written

$$m\frac{d^2\boldsymbol{r}_j}{dt^2} = \zeta\left(\boldsymbol{v}(\boldsymbol{r}_j) - \frac{d\boldsymbol{r}_j}{dt}\right) + \boldsymbol{F}(\boldsymbol{r}_{j+1} - \boldsymbol{r}_j) - \boldsymbol{F}(\boldsymbol{r}_j - \boldsymbol{r}_{j-1}) + \boldsymbol{B}_j, \tag{3}$$

whereas for the beads at the ends the chain, (3) simplifies to

$$m\frac{d^2\boldsymbol{r}_1}{dt^2} = \zeta\left(\boldsymbol{v}(\boldsymbol{r}_1) - \frac{d\boldsymbol{r}_1}{dt}\right) + \boldsymbol{F}(\boldsymbol{r}_2 - \boldsymbol{r}_1) + \boldsymbol{B}_1, \tag{4}$$

$$m\frac{d^2\boldsymbol{r}_{d+1}}{dt^2} = \zeta\left(\boldsymbol{v}(\boldsymbol{r}_{d+1}) - \frac{d\boldsymbol{r}_{d+1}}{dt}\right) - \boldsymbol{F}(\boldsymbol{r}_{d+1} - \boldsymbol{r}_d) + \boldsymbol{B}_{d+1}. \tag{5}$$

In Eqns. (3)-(5) ζ is a friction coefficient, $\boldsymbol{v}(\boldsymbol{r}_j)$ is the velocity of the solvent at the point with position vector $\boldsymbol{r}_j$, $\boldsymbol{F}(\boldsymbol{r}_{j+1} - \boldsymbol{r}_j) = Hf(|\ \boldsymbol{r}_{j+1} - \boldsymbol{r}_j\ |)(\boldsymbol{r}_{j+1} - \boldsymbol{r}_j)$ (for some spring constant H and spring law f) is the force acting on the jth bead due to tension in the spring of the jth segment and $\boldsymbol{B}_j$ denotes the Brownian forces acting on the jth bead. The acceleration terms in Eqns. (3)-(5) may be neglected (see [10]

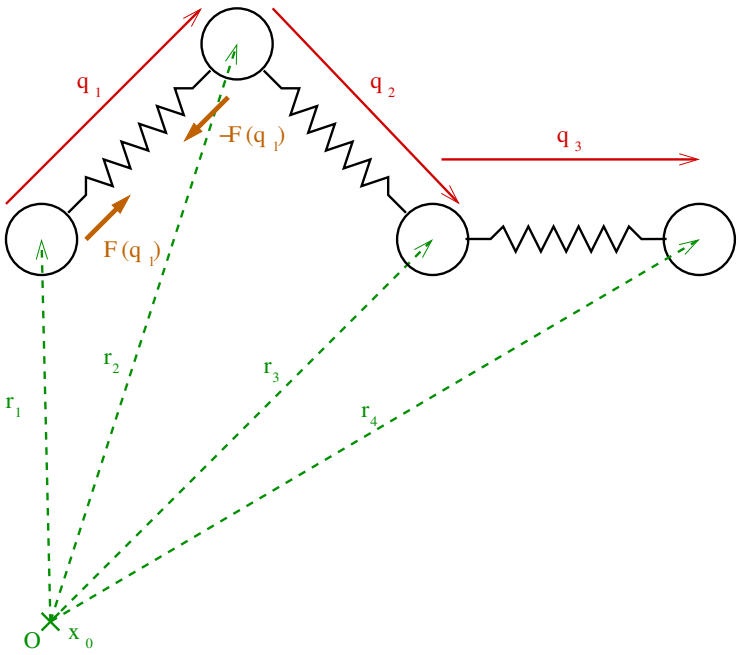


FIGURE 1. Schematic of a bead-spring chain of $d = 3$ segments

for a justification) and no account is taken of any segment-segment or bead-bead interaction apart from that which exists through the tension in the segment springs. In the present paper for simplicity, the chains are supposed to live on a plane and the vectors $\boldsymbol{r}_j$, $\boldsymbol{B}_j$ etc. are hence two-dimensional. Note that simulations of bead-spring chains evolving in the 3-dimensional physical space should be feasible by the approach analogous to that described below (cf. [4] for the case of a single dumbbell model).

The Brownian forces are approximated by a Wiener process:

$$\boldsymbol{B}_j dt = \sqrt{2k_B T\zeta} d\boldsymbol{W}_j, \tag{6}$$

where k_B is Boltzmann's constant, T is the absolute temperature and $\boldsymbol{W}_j(t) = (W_{j,1}(t), W_{j,2}(t))$ is a Gaussian process in 2 dimensions with

$$\langle W_{j,k}(t)\rangle = 0, \qquad \langle W_{j,k}(t) W_{j',k'}(t')\rangle = \delta_{jj'}\delta_{kk'} \min(t,t'). \tag{7}$$

Introducing the jth connector vector

$$\boldsymbol{q}_j = \boldsymbol{r}_{j+1} - \boldsymbol{r}_j, \qquad j = 1, \ldots, d, \tag{8}$$

and the position vector of the centre of mass

$$\boldsymbol{r}_c = \frac{1}{d+1}\sum_{j=1}^{d+1} \boldsymbol{r}_j, \tag{9}$$

we see that the equations of motion (3)-(5) may be rewritten in the form

$$d\boldsymbol{q}_j(t) = \left[\boldsymbol{v}(\boldsymbol{r}_{j+1}) - \boldsymbol{v}(\boldsymbol{r}_j) - \frac{1}{\zeta}\sum_{k=1}^{d} A_{jk}\boldsymbol{F}(\boldsymbol{q}_k)\right] dt + \sqrt{\frac{4k_B T}{\zeta}} d\boldsymbol{W}_j^q(t), \tag{10}$$

$$d\boldsymbol{r}_c(t) = \frac{1}{d+1}\sum_{j=1}^{d+1} \boldsymbol{v}(\boldsymbol{r}_j) dt + \sqrt{\frac{2k_B T}{\zeta(d+1)}} d\boldsymbol{W}^c(t), \tag{11}$$

where $\boldsymbol{W}_j^q(t) = \frac{\boldsymbol{W}_{j+1} - \boldsymbol{W}_j}{\sqrt{2}}$ and $\boldsymbol{W}^c(t) = \frac{1}{\sqrt{d+1}}\sum_{j=1}^{d+1} \boldsymbol{W}_j(t)$ both have the same distribution as $\boldsymbol{W}_j$, and $\boldsymbol{A}$ is the Rouse matrix

$$\boldsymbol{A} = \begin{pmatrix} 2 & -1 & & (0) \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ (0) & & -1 & 2 \end{pmatrix}.$$

Suppose that the flow is homogeneous (i.e., $\nabla\boldsymbol{v}$ is a constant matrix so that we are permitted to write $\boldsymbol{v}(\boldsymbol{r}_{j+1}) - \boldsymbol{v}(\boldsymbol{r}_j) = \nabla\boldsymbol{v}\cdot\boldsymbol{q}_j$). Then, after scaling $\boldsymbol{q}_k$ with $l_0 = \sqrt{\frac{k_B T}{H}}$ it may be shown (see Theorem 3.30 in Öttinger [19], for example) that the Fokker-Planck equation corresponding to (10)-(11) and describing the evolution of the configuration pdf ψ is

$$\frac{\partial\psi}{\partial t} = \mathcal{L}_{FP}^d(\psi) = \sum_{k=1}^{d} L^k\psi - \frac{1}{4\lambda}\sum_{k=2}^{d}(N_x^{k-1}M_x^k + N_y^{k-1}M_y^k)\psi - \frac{1}{4\lambda}\sum_{k=1}^{d-1}(N_x^{k+1}M_x^k + N_y^{k+1}M_y^k)\psi. \tag{12}$$

where

$$L^k = \nabla_k\cdot(-\nabla\boldsymbol{v}\cdot\boldsymbol{q}_k + \frac{1}{2\lambda}\boldsymbol{F}(\boldsymbol{q}_k) + \frac{1}{2\lambda}\nabla_k),\ M_x^k = \frac{\partial}{\partial q_{k,x}},\ M_y^k = \frac{\partial}{\partial q_{k,y}},$$

$$N_x^k = F_x(\boldsymbol{q}_k) + \frac{\partial}{\partial q_{k,x}} \text{ and } N_y^k = F_y(\boldsymbol{q}_k) + \frac{\partial}{\partial q_{k,y}}. \tag{13}$$

In (13) $\lambda = \zeta/4H$ is a configuration relaxation time,

$$\boldsymbol{q}_k = \begin{pmatrix} q_{k,x} \\ q_{k,y} \end{pmatrix} \text{ and } \nabla_k = \begin{pmatrix} \frac{\partial}{\partial q_{k,x}} \\ \frac{\partial}{\partial q_{k,y}} \end{pmatrix}. \tag{14}$$

In the present paper the spring force law employed for each of the segments is that of the FENE model [2]

$$\boldsymbol{F}(\boldsymbol{q}) = \frac{\boldsymbol{q}}{1 - \frac{|\boldsymbol{q}|^2}{b}}, \tag{15}$$

where $\sqrt{b} = q_{max}/l_0$ and q_{max} is the maximum extensibility of the segment spring.

3. A high-order sparse grid Fokker-Planck-based method

In this section we describe how we generate numerical approximations to the leading eigenfunctions of the Fokker-Planck operator $\mathcal{L}^1_{FP}$ and then employ a sparse tensor product of these eigenfunctions to represent the configuration pdf ψ for the full chain.

3.1. Construction of an eigenbasis. The connector vector $\boldsymbol{q}$ for a single FENE dumbbell ($d = 1$) lives in a disk of radius $\sqrt{b}$ and we may write the two cartesian components (q_x, q_y) of $\boldsymbol{q}$ in terms of planar polar coordinates

$$q_x = \rho\cos\theta,\ q_y = \rho\sin\theta, \tag{16}$$

with $\rho \in [0, \sqrt{b}]$ and $\theta \in [0, 2\pi]$. The first step that we take towards solving the Fokker-Planck equation (12) numerically is the construction of an eigenbasis consisting of solutions $\widetilde{\psi}$ to the eigenproblem[1]

$$\mathcal{L}^1_{FP}\widetilde{\psi} = \lambda_i\widetilde{\psi}. \tag{17}$$

We compute the leading N (say) eigenfunctions corresponding to the eigenvalues $\lambda_1, \ldots, \lambda_N$ having real parts that are smallest in absolute value: $0 \geq Re(\lambda_1) \geq Re(\lambda_2) \geq \ldots \geq Re(\lambda_N) \geq \ldots$ This is done using a Fourier-Legendre spectral Galerkin method in a manner to be described below.

We denote the even and odd[2] parts of an eigenfunction $\widetilde{\psi}$ satisfying (17) by $\widetilde{\psi}_0$ and $\widetilde{\psi}_1$. Then

$$\widetilde{\psi} = \widetilde{\psi}_0 + \widetilde{\psi}_1, \tag{18}$$

where $\widetilde{\psi}_0$ and $\widetilde{\psi}_1$ are unique and defined by

$$\begin{aligned} \widetilde{\psi}_0(\rho,\theta) &= \frac{\widetilde{\psi}(\rho,\theta) + \widetilde{\psi}(\rho,\theta+\pi)}{2}, \\ \widetilde{\psi}_1(\rho,\theta) &= \frac{\widetilde{\psi}(\rho,\theta) - \widetilde{\psi}(\rho,\theta+\pi)}{2}. \end{aligned} \tag{19}$$

We note that since the spaces of even and odd functions are invariant under the Fokker-Planck operator $\mathcal{L}^1_{FP}$ at equilibrium, both $\widetilde{\psi}_0$ and $\widetilde{\psi}_1$ are themselves eigenfunctions of this operator, belonging to the same eigenspace as $\widetilde{\psi}$.

Let us introduce the change of dependent variables $\widetilde{\psi} \longrightarrow \widetilde{\alpha}$ for the eigenproblem (17) where

$$\widetilde{\psi}_0 = \left(1 - \frac{\rho^2}{b}\right)^s \widetilde{\alpha}_0,\ \widetilde{\psi}_1 = \rho\left(1 - \frac{\rho^2}{b}\right)^s \widetilde{\alpha}_1, \tag{20}$$

[1]The eigenvalues λ_i are not to be confused with the relaxation time λ in (12).

[2]A function $f(t,\rho,\theta)$ will be called *even* if $f(t,\rho,\theta) = f(t,\rho,\theta+\pi)$ and *odd* if $f(t,\rho,\theta) = -f(t,\rho,\theta+\pi)$

$\widetilde{\alpha}_0$ and $\widetilde{\alpha}_1$ are, respectively, the even and odd parts of $\widetilde{\alpha}$, and s is a real parameter greater than unity. The motivation for the change of variables from $\widetilde{\psi}$ to $\widetilde{\alpha}$ is firstly to circumvent the numerical problems created by the unboundedness of $\boldsymbol{F}$ as $\rho^2 \longrightarrow b$ in the term $\boldsymbol{F}(\boldsymbol{q})\psi$ of $\mathcal{L}^{(1)}_{FP}$ (see Eqn. (12) when $d = 1$). Secondly, the factors multiplying $\widetilde{\alpha}_0$ and $\widetilde{\alpha}_1$ in (20) ensure that $\widetilde{\psi} = 0$ on $\rho = \sqrt{b}$. Finally, if $\widetilde{\psi}$ is even we see from (20) that $\partial\widetilde{\psi}/\partial\rho = 0$ identically at $\rho = 0$ and, if odd, that $\widetilde{\psi} = 0$ at $\rho = 0$.

Using the shorthand notation

$$\psi^0_a(\rho) = \rho^a \left(1 - \frac{\rho^2}{b}\right)^s, \qquad a = 0, 1, \tag{21}$$

we see that we may write

$$\widetilde{\psi} = \psi^0_0\widetilde{\alpha}_0 + \psi^0_1\widetilde{\alpha}_1, \tag{22}$$

and that therefore there exists an invertible operator $\mathcal{R}^{(1)}$ (say) such that $\widetilde{\alpha} = \mathcal{R}^{(1)}\widetilde{\psi}$ and defined for any function $\widetilde{\psi}$ of $(t, \boldsymbol{q})$ having even and odd parts $\widetilde{\psi}_0$ and $\widetilde{\psi}_1$ as

$$\mathcal{R}^{(1)}\widetilde{\psi} = \left(\psi^0_0\right)^{-1}\widetilde{\psi}_0 + \left(\psi^0_1\right)^{-1}\widetilde{\psi}_1. \tag{23}$$

The original eigenproblem (17) may now be written in terms of $\widetilde{\alpha}$ as

$$\widehat{\mathcal{L}}^1_{FP}\widetilde{\alpha} = \lambda_i\widetilde{\alpha}, \tag{24}$$

where

$$\widehat{\mathcal{L}}^1_{FP} = \mathcal{R}^{(1)}\mathcal{L}^1_{FP}\left(\mathcal{R}^{(1)}\right)^{-1}, \tag{25}$$

or, in terms of the action of $\widehat{\mathcal{L}}^1_{FP}$ on the even and odd parts of $\widetilde{\alpha}$,

$$\widehat{\mathcal{L}}^1_{FP}\widetilde{\alpha}_a = \left(\psi^0_a\right)^{-1}\mathcal{L}^1_{FP}\left(\psi^0_a\widetilde{\alpha}_a\right), \qquad a = 0, 1. \tag{26}$$

3.1.1. *A spectral Galerkin method.* In this paper $\widetilde{\alpha}_0$ and $\widetilde{\alpha}_1$ will be represented in terms of even and odd Fourier-Legendre series in $[-1, 1] \times [0, 2\pi]$ after performing the change of variable ρ to η, where

$$\rho^2 = b\left(\frac{1+\eta}{2}\right).$$

This yields

$$\widetilde{\alpha} = \sum_{i=0}^{1}\sum_{k=0}^{N_R}\sum_{l=i}^{2N_F}\alpha_{ikl}h_k(\eta)\Theta_{il}(\theta), \tag{27}$$

where

$$\Theta_{il}(\theta) = (1-i)\cos l\theta + i\sin l\theta, \tag{28}$$

and $\{h_k(\eta)\}_{0\le k\le N_R}$ are Lagrange interpolating polynomials based on the $(N_R + 1)$ Gauss-Legendre points in $(-1, 1)$. From the representation (27) of $\widetilde{\alpha}$ it may be seen that the total number of Fourier modes used is $1 + 2N_F + 2N_F = 4N_F + 1$.

An algebraic system of equations corresponding to (24) is obtained using a Galerkin spectral method in which test functions of the form $h_m(\eta)$, $h_m(\eta)\Theta^{odd}_{il}$ or $h_m(\eta)\Theta^{even}_{il}$ $(0 \le m \le N_R,\ 1 \le i \le 2,\ 1 \le l \le N_F)$ are employed. Integrals with respect to η are evaluated using the standard Gauss quadrature rule based on the Gauss-Legendre points, and integrals with respect to θ are evaluated analytically. The eigenvalues and associated eigenvectors of the resulting $(N_R + 1)(4N_F + 1) \times (N_R + 1)(4N_F + 1)$ matrix are computed by first reducing the matrix to upper Hessenberg form and then using a hybrid double-shifted LR-QR algorithm (see Haag and Watkins [11]). All of

this is done by a call to the standard Intel library routine EVLRG. We shall denote the eigenfunctions of the single segment operator $\widehat{\mathcal{L}}^1_{FP}$ that are computed in this way by $\widetilde{\varphi}_k$ $(k = 1, \ldots, (N_R + 1)(4N_F + 1))$ in the sequel.

3.2. A sparse tensor product spectral method for $\psi(t, \boldsymbol{q}_1, \ldots, \boldsymbol{q}_d)$. As a generalisation of the (unique) decomposition (18) of the function $\widetilde{\psi}$ into a sum of its even and odd parts it may be easily shown that a unique representation exists of any function ψ of the independent variables t and $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_d$ in terms of 2^d new functions $\psi^{(d)}_{\boldsymbol{p}}$

$$\psi(t, \boldsymbol{q}_1, \ldots, \boldsymbol{q}_d) = \sum_{p_1,\ldots,p_d=0}^{1} \psi^{(d)}_{\boldsymbol{p}}(t, \rho_1, \theta_1, \ldots, \rho_d, \theta_d), \tag{29}$$

where $\boldsymbol{p} = (p_1, \ldots, p_d)$ is a parity vector, the component p_i being equal to zero if $\psi^{(d)}_{\boldsymbol{p}}$ is even in the variable $\boldsymbol{q}_i$ and p_i equalling 1 if $\psi^{(d)}_{\boldsymbol{p}}$ is odd in $\boldsymbol{q}_i$, i.e. for $k = 1, \ldots, d$

$$\psi^{(d)}_{\boldsymbol{p}}(t, \rho_1, \theta_1, \ldots, \rho_k, \theta_k + \pi, \ldots, \rho_d, \theta_d) = (-1)^{p_k} \psi^{(d)}_{\boldsymbol{p}}(t, \rho_1, \theta_1, \ldots, \rho_k, \theta_k, \ldots, \rho_d, \theta_d).$$

As in the change of variable (22) we now write down the solution ψ to the full Fokker-Planck equation (12) in terms of new functions $\alpha^{(d)}_{\boldsymbol{p}}$:

$$\begin{aligned} \psi(t, \boldsymbol{q}_1, \ldots, \boldsymbol{q}_d) &= \sum_{p_1,\ldots,p_d=0}^{1} \left(\prod_{j=1}^{d} \rho_j^{p_j} (1 - \frac{\rho^2}{b})^s \right) \alpha^{(d)}_{\boldsymbol{p}}(t, \rho_1, \theta_1, \ldots, \rho_d, \theta_d), \\ &= \sum_{p_1,\ldots,p_d=0}^{1} \psi^0_{\boldsymbol{p}} \alpha^{(d)}_{\boldsymbol{p}}(t, \rho_1, \theta_1, \ldots, \rho_d, \theta_d) \end{aligned} \tag{30}$$

where

$$\psi^0_{\boldsymbol{p}}(\rho_1, \ldots, \rho_d) = \prod_{j=1}^{d} \rho_j^{p_j} \left(1 - \frac{\rho_j^2}{b}\right)^s = \prod_{j=1}^{d} \psi^0_{p_j}(\rho_j), \tag{31}$$

$\psi^0_{p_j}(\rho_j)$ being as defined in (21).

In just the same way as we introduced the change of variables $\widetilde{\alpha} = \widetilde{\alpha}_0 + \widetilde{\alpha}_1$ in Section 3.1 we now write down the new variable

$$\alpha = \sum_{p_1,\ldots,p_d=0}^{1} \alpha^{(d)}_{\boldsymbol{p}}(t, \rho_1, \theta_1, \ldots, \rho_d, \theta_d), \tag{32}$$

in terms of which we will attempt to express the Fokker-Planck equation (12). This is easily done by observing that there exists an invertible operator $\mathcal{R}^{(d)}$ such that

$$\alpha = \mathcal{R}^{(d)} \psi, \tag{33}$$

where, given any function ψ of t and $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_d$ written (uniquely) in the form (29)

$$\mathcal{R}^{(d)} \psi = \sum_{p_1,\ldots,p_d=0}^{1} \left(\psi^0_{\boldsymbol{p}}\right)^{-1} \psi^{(d)}_{\boldsymbol{p}}(t, \rho_1, \theta_1, \ldots, \rho_d, \theta_d). \tag{34}$$

Eqn. (12) therefore becomes

$$\frac{\partial \alpha}{\partial t} = \widehat{\mathcal{L}}^d_{FP}(\alpha) \tag{35}$$

where

$$\widehat{\mathcal{L}}^d_{FP} = \mathcal{R}^{(d)} \mathcal{L}^d_{FP} \left(\mathcal{R}^{(d)}\right)^{-1}. \tag{36}$$

The transformed Fokker-Planck equation may be written in full as

$$\frac{\partial \alpha}{\partial t} = \widehat{\mathcal{L}}^d_{FP}\alpha = \sum_{k=1}^{d} \widehat{L}^k\alpha - \frac{1}{4\lambda}\sum_{k=2}^{d}(\widehat{N}_x^{k-1}\widehat{M}_x^k + \widehat{N}_y^{k-1}\widehat{M}_y^k)\alpha - \frac{1}{4\lambda}\sum_{k=1}^{d-1}(\widehat{N}_x^{k+1}\widehat{M}_x^k + \widehat{N}_y^{k+1}\widehat{M}_y^k)\alpha, \tag{37}$$

where

$$\widehat{L}^k = \mathcal{R}^{(d)} L^k \left(\mathcal{R}^{(d)}\right)^{-1},\ \widehat{N}_x^{k-1}\widehat{M}_x^k = \mathcal{R}^{(d)} N_x^{k-1} M_x^k \left(\mathcal{R}^{(d)}\right)^{-1}, \text{etc.} \tag{38}$$

Equation (37) is solved subject to an initial condition consisting of the equilibrium solution. In terms of the pdf ψ this is

$$\psi = \psi^{(eq)} = \left(\frac{b+2}{2\pi b}\right)^d \prod_{k=1}^{d}\left(1 - \frac{\rho_k^2}{b}\right)^{b/2}, \tag{39}$$

which is equivalent (see (32)) to

$$\alpha = \alpha^{(eq)} = \alpha_{\mathbf{0}}^{(d)} = \left(\frac{b+2}{2\pi b}\right)^d \prod_{k=1}^{d}\left(1 - \frac{\rho_k^2}{b}\right)^{b/2-s}. \tag{40}$$

3.2.1. *A sparse grid spectral Galerkin method.* We will now utilise the eigenfunctions $\{\widetilde{\varphi}_k\}$ described in Section 3.1 to generate a basis for the variable α (see Eqn. (32)). Define a sequence of spaces $V_0, V_1, V_2, \ldots$ such that

$$V_0 = \emptyset, \tag{41}$$

and

$$V_l = span\left\{\widetilde{\varphi}_1, \ldots, \widetilde{\varphi}_{N_l}\right\}, \tag{42}$$

where $N_l = \min(2^l - 1, (N_R+1)(4N_F+1))$, $l = 1, \ldots, L$ and L is the integer defined such that, given N_R and N_F, $2^{L-1} - 1 < (N_R+1)(4N_F+1) \leq 2^L - 1$.

Then the spaces $V_0, V_1, V_2, \ldots$ are nested:

$$V_0 \subset V_1 \subset V_2 \subset \ldots$$

and we may define difference spaces $W_1, W_2, W_3, \ldots$ expressing the hierarchical excess by $W_l = V_l \setminus V_{l-1}$.

Letting $\boldsymbol{i} = (i_1, \ldots, i_d)$ be a $d-$dimensional multi-index we introduce the product function

$$\Phi_{\boldsymbol{i}}(\eta_1, \theta_1, \ldots, \eta_d, \theta_d) = \prod_{j=1}^{d} \widetilde{\varphi}_{i_j}(\eta_j, \theta_j). \tag{43}$$

From the manner in which the eigenproblem (24) is solved numerically, each eigenfunction $\widetilde{\varphi}_{i_j}$ turns out to be either odd or even (and even if they weren't we could construct such a set). We seek to approximate α at level l in one of two ways. The first is a full tensor product (FTP)

$$\alpha \approx \sum_{\boldsymbol{i}:1\leq\|\boldsymbol{i}\|_\infty\leq N_l} \alpha_{\boldsymbol{i}}(t)\Phi_{\boldsymbol{i}}(\eta_1, \theta_1, \ldots, \eta_d, \theta_d) \in V_l \bigotimes \cdots \bigotimes V_l, \tag{44}$$

and the second a sparse tensor product (STP) [22, 28]

(45)
$$\alpha \approx \sum_{\boldsymbol{k}:d\le\|\boldsymbol{k}\|_1\le d+l-1}\ \sum_{\boldsymbol{i}\in\mathcal{I}_{\boldsymbol{k},d}} \alpha_{\boldsymbol{i}}(t)\Phi_{\boldsymbol{i}}(\eta_1,\theta_1,\ldots,\eta_d,\theta_d) \in \bigoplus_{\boldsymbol{k}:d\le\|\boldsymbol{k}\|_1\le d+l-1} W_{k_1}\bigotimes\cdots\bigotimes W_{k_d},$$

where the index set $\mathcal{I}_{\boldsymbol{k},d}$ in (45) is defined to consist of those $\boldsymbol{i}$ such that $\widetilde{\varphi}_{i_j}\in W_{k_j}$:

(46)
$$\mathcal{I}_{\boldsymbol{k},d} = \{\boldsymbol{i}\in\mathbb{R}^d : 2^{k_j-1}\le i_j\le \min(2^{k_j}-1,(N_R+1)(4N_F+1))\}.$$

As an example of the formulae (44) and (45) we take $d=2$ and $l=2$. In this case, the FTP representation (44) becomes

$$\alpha \approx \sum_{i_1=1}^{3}\sum_{i_2=1}^{3}\alpha_{i_1,i_2}(t)\widetilde{\varphi}_{i_1}\widetilde{\varphi}_{i_2},$$

whereas (45) would now read

$$\alpha \approx \alpha_{1,1}\widetilde{\varphi}_1\widetilde{\varphi}_1+\alpha_{1,2}\widetilde{\varphi}_1\widetilde{\varphi}_2+\alpha_{2,1}\widetilde{\varphi}_2\widetilde{\varphi}_1+\alpha_{1,3}\widetilde{\varphi}_1\widetilde{\varphi}_3+\alpha_{3,1}\widetilde{\varphi}_3\widetilde{\varphi}_1.$$

Let us introduce a dual space V_l^* given by

$$V_l^* = span\{f_1,\ldots,f_{N_l}\},$$

where each function f_i $(i=1,2,\ldots,N_l)$ in this dual basis is defined as

(47)
$$f_i(v) = \langle\widetilde{\varphi}_i^*, v\rangle_1, \quad \forall v\in V_l.$$

In (47) $\langle\cdot,\cdot\rangle_1$ is an inner product on V_l defined by

$$\begin{aligned}\langle f,g\rangle_1 &= \int_{\theta=0}^{2\pi}\int_{\rho=0}^{\sqrt{b}} f(\rho,\theta)g(\rho,\theta)\ \rho\ d\rho d\theta,\\ &= \frac{b}{4}\int_{\theta=0}^{2\pi}\int_{\eta=-1}^{1} f(\rho(\eta),\theta)g(\rho(\eta),\theta)d\eta d\theta,\end{aligned} \tag{48}$$

and the functions $\{\widetilde{\varphi}_i^*\}_{i=0}^{N_l}\in V_l$, chosen so that $f_i(\widetilde{\varphi}_j)=\delta_{ij}$. For this to be the case, we express $\widetilde{\varphi}_i^*$ as a linear combination of the basis functions $\{\widetilde{\varphi}_j\}$:

$$\widetilde{\varphi}_i^* = \sum_{j=1}^{N_l} a_{ij}\widetilde{\varphi}_j, \tag{49}$$

and see that

$$\langle\widetilde{\varphi}_i^*,\widetilde{\varphi}_k\rangle_1 = \sum_{j=1}^{N_l} a_{ij}\langle\widetilde{\varphi}_j,\widetilde{\varphi}_k\rangle_1, \tag{50}$$

so that the $N_l\times N_l$ matrix $\boldsymbol{A}=(a_{ij})$ should be chosen as the inverse of the Gram matrix $\boldsymbol{G}$ having (i,j)th component

$$g_{ij} = \langle\widetilde{\varphi}_i,\widetilde{\varphi}_j\rangle_1. \tag{51}$$

An algebraic set of discrete equations corresponding to (37) is found by using a Galerkin method as follows: The $d-$dimensional operator

$$\frac{\partial}{\partial t} - \widehat{\mathcal{L}}_{FP}^d,$$

is applied to either the full (44) or sparse (45) tensor product representation of α. Then the product is taken with test functions of the form

$$\Phi_{\boldsymbol{j}}^* = \Pi_{l=1}^d\widetilde{\varphi}_{j_l}^*, \tag{52}$$

(with $\boldsymbol{j}$ drawn from the same index set as in (44) or (45)) to get, for all suitable $\boldsymbol{j}$,

$$\langle \frac{\partial \alpha}{\partial t} - \widehat{\mathcal{L}}^d_{FP}\alpha, \Phi^*_{\boldsymbol{j}}\rangle_d = 0, \tag{53}$$

where $\langle\cdot,\cdot\rangle_d$ generalises the inner product introduced in (48) to an integral computed over the domain in $\mathbb{R}^{2d}$: $\{|\boldsymbol{q}_j| < \sqrt{b},\ j = 1,\ldots,d\}$.

Writing α in its form (44) or (45), (53) may be expressed as

$$\sum_{\boldsymbol{i}} \frac{d\alpha_{\boldsymbol{i}}(t)}{dt}\langle \Phi_{\boldsymbol{i}}, \Phi^*_{\boldsymbol{j}}\rangle_d - \alpha_{\boldsymbol{i}}(t)\langle \widehat{\mathcal{L}}^d_{FP}\Phi_{\boldsymbol{i}}, \Phi^*_{\boldsymbol{j}}\rangle_d = 0, \tag{54}$$

where the index set for $\boldsymbol{i}$ and $\boldsymbol{j}$ depends on whether an STP or FTP representation for α is used. Using the orthogonality properties of the bases $\{\widetilde{\varphi}^*_i\}$ and $\{\widetilde{\varphi}_j\}$ (that is, by construction, $f_i(\widetilde{\varphi}_j) = \delta_{ij}$) we then arrive at

$$\begin{aligned}
\frac{d\alpha_{\boldsymbol{j}}(t)}{dt} &- \sum_{k=1}^{d}\sum_{i_k=1}^{N_l} \alpha_{j_1\ldots j_{k-1}i_kj_{k+1}\ldots j_d}\langle \widehat{L}^k\widetilde{\varphi}_{i_k}, \widetilde{\varphi}^*_{j_k}\rangle_1 \\
&- \frac{1}{4\lambda}\sum_{k=2}^{d}\sum_{i_{k-1}=1}^{N_l}\sum_{i_k=1}^{N_l} \alpha_{j_1\ldots j_{k-2}i_{k-1}i_kj_{k+1}\ldots j_d} \\
&\qquad \times \Big[\langle \widehat{N}^{k-1}_x\widetilde{\varphi}_{i_{k-1}}, \widetilde{\varphi}^*_{j_{k-1}}\rangle_1\langle \widehat{M}^k_x\widetilde{\varphi}_{i_k}, \widetilde{\varphi}^*_{j_k}\rangle_1 \\
&\qquad\quad + \langle \widehat{N}^{k-1}_y\widetilde{\varphi}_{i_{k-1}}, \widetilde{\varphi}^*_{j_{k-1}}\rangle_1\langle \widehat{M}^k_y\widetilde{\varphi}_{i_k}, \widetilde{\varphi}^*_{j_k}\rangle_1\Big] \\
&- \frac{1}{4\lambda}\sum_{k=1}^{d-1}\sum_{i_k=1}^{N_l}\sum_{i_{k+1}=1}^{N_l} \alpha_{j_1\ldots j_{k-1}i_ki_{k+1}j_{k+2}\ldots j_d} \\
&\qquad \times \Big[\langle \widehat{M}^k_x\widetilde{\varphi}_{i_k}, \widetilde{\varphi}^*_{j_k}\rangle_1\langle \widehat{N}^{k+1}_x\widetilde{\varphi}_{i_{k+1}}, \widetilde{\varphi}^*_{j_{k+1}}\rangle_1 \\
&\qquad\quad + \langle \widehat{M}^k_y\widetilde{\varphi}_{i_k}, \widetilde{\varphi}^*_{j_k}\rangle_1\langle \widehat{N}^{k+1}_y\widetilde{\varphi}_{i_{k+1}}, \widetilde{\varphi}^*_{j_{k+1}}\rangle_1\Big] = 0.
\end{aligned} \tag{55}$$

Discretization in time is effected via an implicit scheme where, denoting the vector of unknowns by $\boldsymbol{\alpha}$, the partial time derivative is approximated by a simple forward difference

$$\frac{\partial \boldsymbol{\alpha}}{\partial t} \approx \frac{\boldsymbol{\alpha}^{n+1} - \boldsymbol{\alpha}^n}{\Delta t}, \tag{56}$$

and $\boldsymbol{\alpha}$ is evaluated at time step $(n+1)$ in the terms in (37) corresponding to $\sum_k \widehat{L}^k\alpha$ and at time level n elsewhere. All of this leads to the fully discrete scheme

$$\begin{aligned}
\boldsymbol{\alpha}^{n+1} - \Delta t\sum_{k=1}^{d}\bar{L}^k\boldsymbol{\alpha}^{n+1} =& \boldsymbol{\alpha}^n - \frac{\Delta t}{4\lambda}\sum_{k=2}^{d}(\bar{N}^{k-1}_x\bar{M}^k_x + \bar{N}^{k-1}_y\bar{M}^k_y)\boldsymbol{\alpha}^n \\
&- \frac{\Delta t}{4\lambda}\sum_{k=1}^{d-1}(\bar{N}^{k+1}_x\bar{M}^k_x + \bar{N}^{k+1}_y\bar{M}^k_y)\boldsymbol{\alpha}^n,
\end{aligned} \tag{57}$$

where an overbar on a quantity indicates that it is a matrix. The coefficient matrix of $\boldsymbol{\alpha}^{n+1}$ in the scheme (57) is approximated (to $O(\Delta t)^2$) in our code by the matrix equivalent of the operator $\bigotimes_{k=1}^{d}\left(\boldsymbol{I} - \Delta t\widehat{L}^k\right)$.

4. Stochastic algorithm and stress calculator

In homogeneous polymeric flows the probability density function ψ will be independent of the centre of mass of a chain $\boldsymbol{r}_c$. We therefore neglect the stochastic equation (11) for the centre of mass of a polymer chain (it just moves with the local velocity) and, scaling the position vector $\boldsymbol{q}_j$ of the jth connector vector with $l_0 = \sqrt{\frac{k_B T}{H}}$, see that its equation of motion (10) becomes

$$d\boldsymbol{q}_j = \left[\nabla \boldsymbol{v} \cdot \boldsymbol{q}_j - \frac{1}{2\lambda}\frac{\boldsymbol{q}_j}{1-q_j^2/b}\right] dt + \left[\frac{1}{4\lambda}\boldsymbol{F}(\boldsymbol{q}_{j+1}) + \frac{1}{4\lambda}\boldsymbol{F}(\boldsymbol{q}_{j-1})\right] dt + \sqrt{\frac{1}{\lambda}} d\boldsymbol{W}_j^q. \tag{58}$$

Eqn. (58) is solved subject to an initial condition consisting of a random sample $\{\boldsymbol{q}_1, \cdots, \boldsymbol{q}_d\}$ drawn from the equilibrium distribution (39). Since all the equations (58) ($j = 1, \ldots, d$) are coupled we advance the calculation through one time step Δt by performing one sweep from $\boldsymbol{q}_1$ to $\boldsymbol{q}_d$ of an iterative method whereby, for a given index 'j' the vectors $\{\boldsymbol{q}_1, \ldots, \boldsymbol{q}_{j-1}\}$ are assumed known at $t = (n+1)\Delta t$ and $\{\boldsymbol{q}_{j+1}, \ldots, \boldsymbol{q}_d\}$ at time $t = n\Delta t$. This leads to the scheme

$$\begin{aligned} q_{j,x}^{n+1} + \frac{\triangle t}{2\lambda}\frac{q_{j,x}^{n+1}}{1-q^2/b} &= q_{j,x}^n + \left(\nabla \boldsymbol{v} \cdot \boldsymbol{q}_j^n\right)_x \Delta t \\ &+ \left[F_x(\boldsymbol{q}_{j+1}^n) + F_x(\boldsymbol{q}_{j-1}^{n+1})\right] \Delta t + \sqrt{\frac{\Delta t}{\lambda}}\widetilde{W}_{j,x} = r_{j,x}^n, \end{aligned} \tag{59}$$

$$\begin{aligned} q_{j,y}^{n+1} + \frac{\Delta t}{2\lambda}\frac{q_{j,y}^{n+1}}{1-q^2/b} &= q_y^n + \left(\nabla \boldsymbol{v} \cdot \boldsymbol{q}_j^n\right)_y \Delta t \\ &+ \left[F_y(\boldsymbol{q}_{j+1}^n) + F_y(\boldsymbol{q}_{j-1}^{n+1})\right] \Delta t + \sqrt{\frac{\Delta t}{\lambda}}\widetilde{W}_{j,y} = r_{j,y}^n, \end{aligned} \tag{60}$$

where

$$q^2 = (q_{j,x}^{n+1})^2 + (q_{j,y}^{n+1})^2$$

and

$$\widetilde{\boldsymbol{W}_j} = \frac{1}{\sqrt{\Delta t}}(\boldsymbol{W}_j^q(t+\Delta t) - \boldsymbol{W}_j^q(t)), \tag{61}$$

which implies that each component $\widetilde{W}_j \sim N(0,1)$. >From the previous equations

$$q_{j,x}^{n+1} = \frac{r_{j,x}^n}{1 + \frac{\triangle t}{2\lambda}\frac{1}{1-q^2/b}} \text{ and } q_{j,y}^{n+1} = \frac{r_{j,y}^n}{1 + \frac{\triangle t}{2\lambda}\frac{1}{1-q^2/b}}, \tag{62}$$

and therefore,

$$\boldsymbol{q}_j^{n+1}\left(1 + \frac{\triangle t}{2\lambda}\frac{1}{1-q^2/b}\right) = \boldsymbol{r}_j^n.$$

Taking absolute values in the above and setting $r = \sqrt{\left(r_{j,x}^n\right)^2 + \left(r_{j,y}^n\right)^2}$ we get the cubic equation

$$q^3 - rq^2 - bq\left(1 + \frac{\Delta t}{2\lambda}\right) + br = 0.$$

This equation has one real root q between 0 and $\sqrt{b}$. Once this has been computed $q_{j,x}$ and $q_{j,y}$ may be found from Eqn. (62).

The elastic extra stress tensor $\boldsymbol{\tau}$ appearing in (1) may be written in the form

$$\boldsymbol{\tau} = \sum_{j=1}^{d} \boldsymbol{\tau}_j, \tag{63}$$

where $\boldsymbol{\tau}_j$ is the contribution of the jth segment in all the polymer chains of the solution to the total elastic extra stress. In the case of a homogeneous flow the stress depends only upon time. The Kramers expression for $\boldsymbol{\tau}_j$ at time $n\Delta t$ now reads

$$\boldsymbol{\tau}_j(n\Delta t) \approx \frac{(1-\beta)}{We}\frac{(b+4)}{b}\left[\boldsymbol{I} + \langle \boldsymbol{q}_j^n \bigotimes \boldsymbol{F}(\boldsymbol{q}_j^n)\rangle\right], \tag{64}$$

where $\langle\cdot\rangle$ denotes an average taken over all M chains in the solution and We denotes a Weissenberg number, defined as $We = \lambda\dot{\gamma}$; a dimensionless shear rate. In terms of the joint probability density function ψ, $\boldsymbol{\tau}_j(n\Delta t)$ is written as

$$\begin{aligned}\boldsymbol{\tau}_j(n\Delta t) = \frac{(1-\beta)}{We}\frac{(b+4)}{b}\Bigg[\boldsymbol{I} + \int_{|\boldsymbol{q}_1|<\sqrt{b}} \cdots \int_{|\boldsymbol{q}_d|<\sqrt{b}} \boldsymbol{q}_j \bigotimes \boldsymbol{F}(\boldsymbol{q}_j) \\ \times\psi(n\Delta t, \boldsymbol{q}_1, \ldots, \boldsymbol{q}_d) d\boldsymbol{q}_1 \ldots d\boldsymbol{q}_d\Bigg],\end{aligned} \tag{65}$$

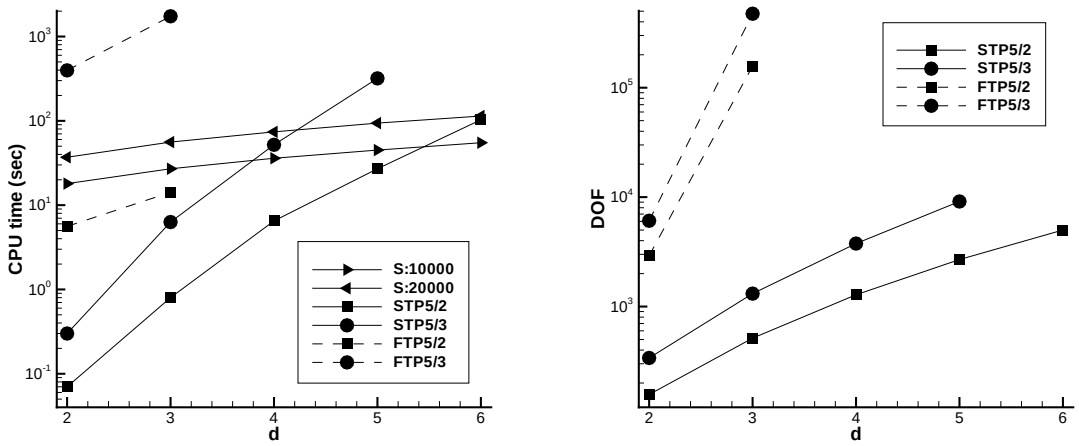
and the convergence of (64) as $M \longrightarrow \infty$ to the evaluation of (65) with the exact ψ is assured by the strong law of large numbers.

5. Example: Steady simple shear flow

In this section we present results for a homogeneous shear flow $\boldsymbol{v} = (v_x, v_y) = (y, 0)$ using both the FTP (44) and STP (45) representations for the transformed distribution function $\widetilde{\alpha}$. The stochastic method (59)-(64) was used with $M = 10000$ and 20000 chains and the tensile elastic stress τ_{xx} computed with $M = 20000$ considered to be a reference solution, against which the Fokker-Planck-based calculations of τ_{xx} could be compared. To this end, for the Weissenberg numbers considered, a time step $\Delta t = 0.01$ was used and an average value for τ_{xx} over the time interval $[7, 10]$ computed. Averages of τ_{xx} computed over later time intervals showed no significant variation and τ_{xx} was considered, as a consequence, to have reached steady state.

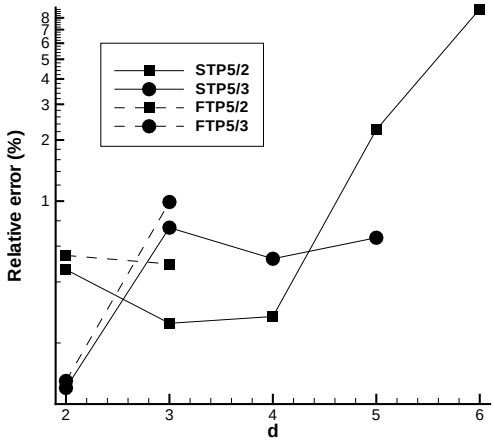
The CPU time in seconds on a Pentium IV machine, as well as the relative error for τ_{xx} committed by the STP and FTP methods for differing values of N_R and N_F and at two different Weissenberg numbers ($We = 1$ and 3) are shown in Figs. 2 and 3. Unless otherwise stated, the level of approximation l in the STP method was taken equal to the maximum value L. At both Weissenberg numbers the STP method is considerably cheaper than an FTP method computed with the same level of discretisation for the eigenbasis. At the same time, at the lower Weissenberg number, the relative error with the STP method is lower than with the corresponding FTP method and the STP method may be seen to be competitive with the stochastic approach for longer chains than its FTP counterpart. At the larger Weissenberg number, the FTP calculations are more accurate than those performed with the STP representation and the deterioration of the accuracy advantage of the STP method for small chains ($d \leq 3$) may be attributable to the use of a Galerkin formulation. FTP calculations were not considered feasible for $d > 3$.

A point of interest in Fig. 3 is that by taking a finer mesh for the eigenfunction calculation ($N_R = 12, N_F = 8$) but choosing the level of approximation $l = 7 < L$ we can get a significant improvement in accuracy over the FTP results from a coarser mesh ($N_R = 5, N_F = 3$) at comparable CPU cost.



(a)

(b)



(c)

FIGURE 2. $We = 1$. (a) CPU time (s), (b) number of degrees of freedom, and (c) percentage relative error for τ_{xx} computed using (65). FTP i/j: full tensor product (44), $N_R = i, N_F = j$. STP i/j: sparse tensor product (45), $N_R = i, N_F = j$. Also shown are the CPU time consumed by the stochastic method (59)-(64) with $M = 10,000$ and $20,000$ realizations.

6. Conclusions

In this paper we have presented a new spectral Galerkin method for the solution of a high-dimensional Fokker-Planck equation arising in the modelling of dilute solutions of long idealized polymers. The method is inspired by ideas in the sparse grid literature and at modest levels of fluid memory and shear rate ($We \approx 1$) is demonstrably superior in accuracy and cost to a traditional full tensor product representation for the polymer configuration distribution function. The method extends the attractiveness of Fokker-Planck-based methods (as a competitive alternative to stochastic

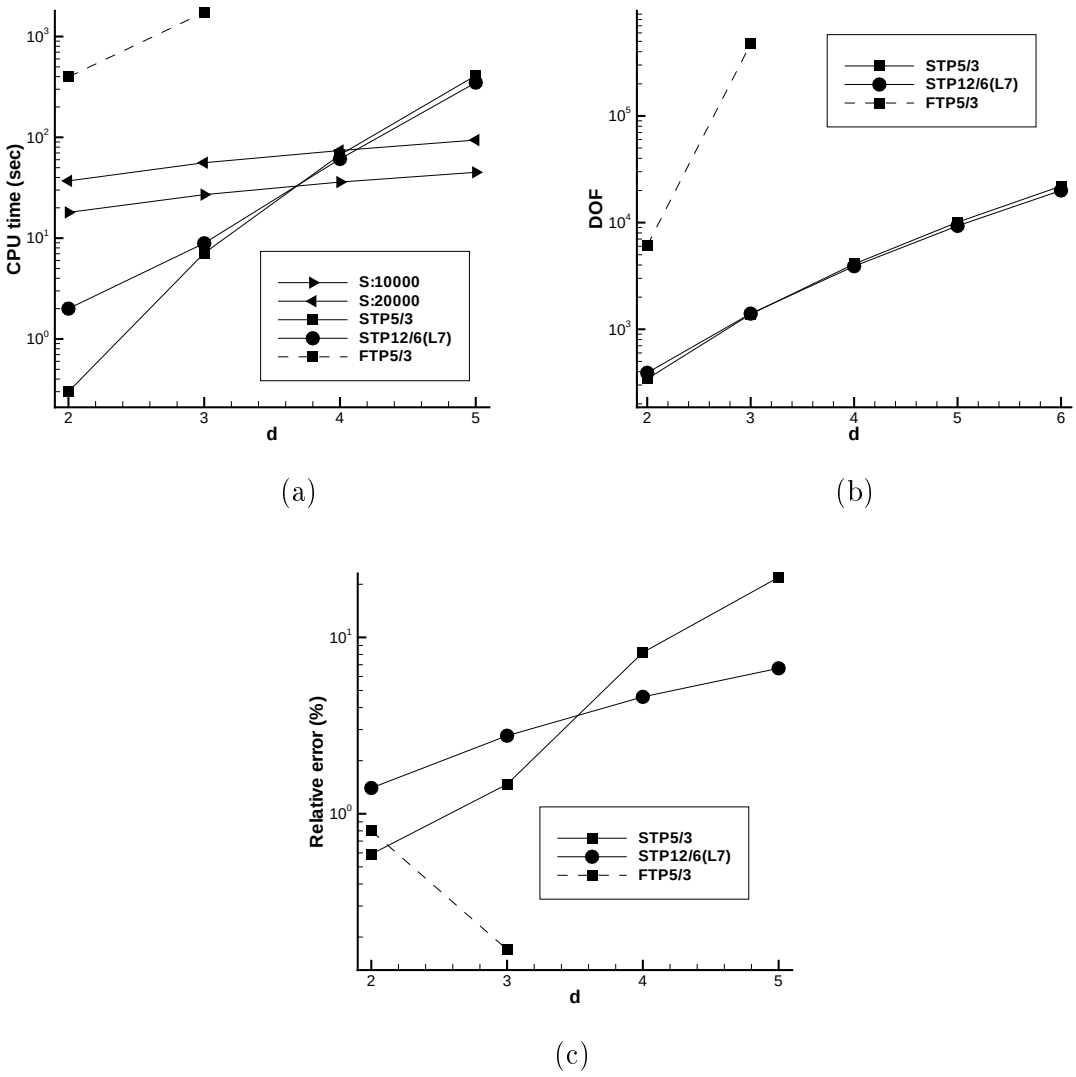


FIGURE 3. $We = 3$. (a) CPU time (s), (b) number of degrees of freedom, and (c) percentage relative error for τ_{xx} computed using (65). FTP 5/3: full tensor product (44), $N_R = 5, N_F = 3$. STP 5/3: sparse tensor product (45), $N_R = 5, N_F = 3$. STP 12/8($L7$): sparse tensor product (45), $N_R = 12, N_F = 8, l = 7$. Also shown are the CPU time consumed by the stochastic method (59)-(64) with $M = 10,000$ and $20,000$ realizations.

simulations) to longer polymer chains than was previously possible, although ultimately in shear flows at sufficiently high Weissenberg numbers (strong flows, highly elastic liquids) it would seem probable that well constructed stochastic methods will have the advantage in terms of cost and accuracy.

Acknowledgments

The second author was supported for the duration of this project by CTI project 6437.1 IWS-IW

References

[1] S. Achatz, High order sparse grid methods for elliptic partial differential equations with variable coefficients, Computing **71** (2003) 1–15.
[2] R. B. Bird, C. F. Curtiss, R. C. Armstrong and O. Hassager, Dynamics of Polymeric Liquids, Volume 2, Kinetic Theory, John Wiley & Sons, New York (1987).
[3] H-J. Bungartz and M. Griebel, Sparse grids, Acta Numerica **13** (2004) 147–269.
[4] C. Chauvière and A. Lozinski, Simulation of complex viscoelastic flows using Fokker-Planck equation: 3D FENE model, J. Non-Newtonian Fluid Mech., **122** (2004) 201–214.
[5] X. Fan, Molecular model and flow calculation: I. The numerical solutions to multibead-rod models in homogeneous flows, Acta Mech. Sin., **5** (1989) 49–59.
[6] X. Fan, Molecular models and flow calculations: II. Simulation of steady planar flow, Acta Mech. Sin., **5** (1989) 216–226.
[7] X.-J. Fan, Viscosity, first normal-stress coefficient, and molecular stretching in dilute polymer solutions, J. Non-Newtonian Fluid Mech., **17** (1985) 125–144.
[8] X.-J. Fan, Viscosity, first normal-stress coefficients and molecular stretching in concentrated solutions and melts, J. Non-Newtonian Fluid Mech., **17** (1985) 251-265.
[9] X.-J. Fan, A numerical investigation of the hydrodynamic interaction for Hookean dumbbell suspensions in steady state shear flow, J. Chem. Phys., **85** (1985) 6237–6238.
[10] M. Fixman, Dynamics of polymer chains, J. Chem. Phys. **42** (1965) 3831–3837.
[11] J. B. Haag and D. S. Watkins, QR-like algorithms for the nonsymmetric eigenvalue problem, ACM Trans. Math. Software **19** (1993) 407–418.
[12] M. A. Hulsen, A. P. G. van Heel and B. H. A. A. van den Brule, Simulation of viscoelastic flows using Brownian configuration fields, J. Non-Newtonian Fluid Mech., **70** (1997) 79–101.
[13] R. Keunings, Micro-macro methods for the multiscale simulation of viscoelastic flow using molecular models of kinetic theory, In: Rheology Reviews 2004, D. M. Binding and K. Walters (Eds.), British Society of Rheology (2004) 67–98.
[14] M. Laso and H.-C. Öttinger, Calculation of viscoelastic flow using molecular models: the CONNFFESSIT approach, J. Non-Newtonian Fluid Mech., **47** (1993) 1–20.
[15] A. Lozinski and C. Chauvière, A fast solver for Fokker-Planck equation applied to viscoelastic flows calculations: 2D FENE model, J. Comput. Phys. **189** (2003) 607–625.
[16] A. Lozinski, C. Chauvière, J. Fang and R. G. Owens, Fokker-Planck simulations of fast flows of melts and concentrated polymer solutions in complex geometries, J. Rheol. **47** (2003) 535–561.
[17] R. Nayak, Molecular simulation of liquid crystal polymer flow: a wavelet-finite element analysis, PhD thesis, MIT, Cambridge, MA. (1998).
[18] H. Niederreiter, Random Number Generation and Quasi-Monte Carlo Methods, Society for Industrial and Applied Mathematics, Philadelphia (1992).
[19] H. C. Öttinger, Stochastic Processes in Polymeric Fluids, Springer-Verlag, Berlin, Heidelberg (1996).
[20] C. Schwab and R. A. Todor, Sparse finite elements for stochastic elliptic problems - higher order moments, Computing **71** (2003) 43–63.
[21] C. Schwab and R. A. Todor, Sparse finite elements for elliptic problems with stochastic loading, Numer. Math. **95** (2003) 707–734.
[22] S. A. Smolyak, Quadrature and interpolation formulas for tensor products of certain classes of functions, Dokl. Akad. Nauk SSSR **4** (1963) 240–243.
[23] J. K. C. Suen, Y. L. Joo and R. C. Armstrong, Molecular orientation effects in viscoelasticity, Annu. Rev. Fluid Mech., **34** (2002) 417–444.
[24] G. Venkiteswaran and M. Junk, Quasi-Monte Carlo algorithms for diffusion equations in high dimensions, Math. Comput. Simul. **68** (2004) 23–41.
[25] G. Venkiteswaran and M. Junk, A QMC approach for high dimensional Fokker-Planck equations modelling polymeric liquids, Math. Comput. Simul. **68** (2004) 43–56.

[26] T. von Petersdorff and C. Schwab, Numerical solution of parabolic equations in high dimensions, ESAIM: Math. Model. Numer. Anal. **38** (2004) 93–127.
[27] H. R. Warner, Kinetic theory and rheology of dilute suspensions of finitely extendible dumbbells, Ind. Eng. Chem. Fundam., **11** (1972) 379–387.
[28] C. Zenger, Sparse grids, In: Parallel Algorithms for Partial Differential Equations: Proceedings of the Sixth GAMM-Seminar Kiel, W. Hackbusch (Ed.), Notes on Numerical Fluid Mechanics, **31** (1991) 241–251.

CUST, Génie Mathématique et Modélisation, Université Blaise Pascal II, Clermont-Ferrand, France.

Chaire d'Analyse et Simulations Numériques, Section de Mathématiques, Ecole Polytechnique Fédérale de Lausanne, CH 1015 Lausanne, Switzerland
Current address: Department of Mathematics, University of Houston, Houston, TX 77204-3008, USA

Département de Mathematiques et de Statistique, Université de Montréal, CP 6128, succursale centre-ville, Montréal QC H3C 3J7, Canada