Molecular and multiscale methods for the numerical simulation of materials

Frédéric Legoll

THÈSE

présentée pour l'obtention du titre de

Docteur de l'Université Pierre et Marie Curie

Spécialité : Mathématiques Appliquées

par

Frédéric LEGOLL

Sujet : *Méthodes moléculaires et multi-échelles pour la simulation numérique des matériaux*

Soutenue le 31 août 2004 devant le jury composé de :

Président :	Evariste SANCHEZ-PALENCIA
Rapporteurs :	Grégoire ALLAIRE Michel CROUZEIX
Examinateurs :	Pascal FREY Karam SAB
Directeurs de thèse :	Claude LE BRIS Yvon MADAY

À Nadine,

À mes parents et à mon frère,

À ma famille.

Remerciements

Je souhaite vivement remercier Claude Le Bris et Yvon Maday, qui ont dirigé mes travaux avec beaucoup de conviction et de qualité pédagogique. Leur enthousiasme et leurs qualités humaines m'ont beaucoup apporté. J'ai pu apprécier le dynamisme et l'exigence de rigueur de Claude au cours des très nombreuses discussions que j'ai eues avec lui. Je voudrais aussi le remercier pour sa disponibilité et ses conseils. Collaborer avec Yvon a été très enrichissant, et je lui suis très reconnaissant pour ses nombreuses suggestions.

Mes remerciements vont également à Eric Cancès, pour toutes les discussions que nous avons eues, sa disponibilité et ses constants encouragements. Merci également de m'avoir donné la possibilité d'enseigner à l'ENPC.

Participer à l'Action de Recherche Coopérative PRESTISSIMO a été un très grand plaisir, et une partie de cette thèse n'aurait pas pu aboutir sans les discussions que nous avons eues. Je tiens donc à chaleureusement remercier tous les membres de l'ARC, Eric Cancès, François Castella, Philippe Chartier, Erwan Faou, Claude Le Bris et Gabriel Turinici.

Je voudrais remercier Xavier Blanc pour sa patience et son sens de l'écoute, son souci de la rigueur et ses encouragements.

Merci également à Régis Monneau, dont j'ai pu apprécier l'inventivité, pour notre collaboration fructueuse.

Je suis très reconnaissant à Grégoire Allaire et à Michel Crouzeix pour avoir bien voulu rapporter sur ce travail, ainsi qu'à Pascal Frey, Karam Sab et Evariste Sanchez-Palencia qui ont accepté de faire partie du jury. Merci pour leurs nombreuses questions !

J'ai pu bénéficier, au cours de cette thèse, d'un fort environnement industriel, et je souhaite remercier EDF R & D, et en particulier Jean-Louis Vaudescal et Olivier Dubois, de m'avoir accepté au sein de leur équipe pendant ces trois années. Merci à Véronique Duwig pour ce que nous avons fait ensemble, et à Guy Bencteux, Alain Léger et Eric Lorentz pour leur intérêt dans mon travail. Merci aussi à Renaud Masson (EDF R & D) et à Georges Cailletaud (Centre des Matériaux, ENSMP) pour m'avoir introduit aux polycristaux.

Ma gratitude va également à tous les chercheurs du CEA avec lesquels j'ai eu plaisir à échanger, Mireille Defranceschi, Christophe Denoual, Jean-Bernard Maillet, Yves-Patrick Pellegrini et Gilles Zérah.

Au cours de cette thèse, j'ai eu la chance et le plaisir d'encadrer plusieurs stagiaires, Yann Revalor, Antonin Orriols et Mohamed El Makrini, ce qui fut très enrichissant.

Le CERMICS est un laboratoire idéal pour faire une thèse, du fait de la richesse et de la diversité des compétences de ses chercheurs. Merci donc à Alexandre Ern pour nos discussions sur les éléments finis, à Jean-François Delmas et à Bernard Lapeyre pour nos discussions sur les méthodes numériques en probabilité, et à Gabriel Turinici pour son aide en calcul scientifique et en anglais. Jacques Daniel a toujours été d'une aide précieuse pour les problèmes informatiques, et Jean-Philippe Chancelier a parfaitement su le seconder. Je remercie également les secrétaires du CERMICS, Sylvie Berte et Khadija Elouali, pour leur aide efficace et leur extrême gentillesse.

La bonne humeur qui règne au CERMICS doit aussi à l'ensemble des thésards. J'ai une pensée particulière pour Maxime, qui m'a supporté au CERMICS et à EDF ! Merci à ceux avec qui j'ai partagé des moments, Laetitia, Bouhari, Maxime, Adel, Adrien, Linda, Yousra, Antoine, Tony, Monsieur François, Nicola, Gabriel Stoltz. Par ailleurs, les échanges scientifiques que j'ai eus avec Maxime Barrault, Antoine Gloria, Tony Lelièvre et Gabriel Stoltz, ainsi qu'avec Barbara Di Ventura et Patrice Hauret, ont été très stimulants.

Merci aux autres membres du CERMICS pour leur disponibilité et leur contribution à la bonne ambiance : Geoffray Adde, Aurélien Alfonsi, Anne Auger, Marc Barton-Smith, Kengy Barty, Marie-Pierre Bavouzet, Héloïse Beaugendre, Gilbert Caplain, Elisabetta Carlini, Maureen Clerc, Guy Cohen, Michel De Lara, Claude Dion, Hervé Galicher, Mohamed Amin Ghorbel, Thérèse Guilbaud, Julien Guyon, Benjamin Jourdain, Olivier Juan, Renaud Keriven, René Lalement, Fabien Le Jeune, Mazyar Mirrahim, Antonin Orriols, Jean-François Pommaret, Jennifer Proft, Thierry Salset, Pierre Tardif d'Hamonville, Emmanuel Temam, ainsi qu'à Olivier Alvarez et à Jean-Luc Guermond.

Je voudrais aussi remercier le Corps des Ponts qui permet aux jeunes ingénieurs qui le désirent de faire de la recherche, ainsi que l'ensemble du personnel de l'Ecole Nationale des Ponts et Chaussées. A l'heure où je termine ma thèse, une nouvelle page s'ouvre au LAMI, ce dont je suis très reconnaissant à Karam Sab et à Denis Duhamel.

Enfin, je ne serais pas ce que je suis sans mon entourage, et je souhaite donc ici exprimer toute ma gratitude envers Nadine, mes parents et mon frère.

Sujet : Méthodes moléculaires et multi-échelles pour la simulation numérique des matériaux.

Résumé : Le travail de cette thèse a porté sur l'étude de modèles moléculaires et de méthodes multi-échelles pour la simulation numérique des matériaux.

Dans une première partie (les chapitres 2, 3 et 4), on s'intéresse à une modélisation à l'échelle atomistique. La physique statistique montre alors que les grandeurs macroscopiques pertinentes sont des moyennes dans l'espace des phases du système étudié. La dynamique moléculaire est une approche pour calculer ces moyennes. L'évolution en temps du système est simulée (par exemple suivant les équations de Newton), ce qui permet de calculer des moyennes temporelles le long des trajectoires du système. Sous l'hypothèse d'ergodicité, ces moyennes convergent en temps long vers la moyenne dans l'espace des phases. Nous nous intéressons ici au rythme de convergence des moyennes temporelles, et faisons en particulier l'analyse de quelques schémas numériques.

Dans un deuxième temps, nous nous intéressons à des approches multi-échelles. Le chapitre 6 est consacré à l'analyse numérique d'une méthode couplant un modèle atomistique avec un modèle de continuum : le domaine de calcul est partitionné en deux sous-domaines, l'un décrit par un modèle de continuum, l'autre par un modèle atomistique. Nous étudions en particulier le critère permettant de choisir en chaque point du matériau le modèle (discret ou continu) qui est utilisé.

Enfin, le chapitre 7 est consacré à l'homogénéisation numérique de modèles de polycristaux, décrivant le comportement de la matière à l'échelle du micromètre.

Title : Molecular and multiscale methods for the numerical simulation of materials.

Abstract : We investigate in this thesis some molecular models and some multiscale methods for the numerical simulation of materials.

The first part (chapters 2, 3 and 4) is devoted to an atomistic modelling. Statistical physics shows that the relevant quantities at the macroscopic scale are phase space averages. Molecular dynamics can be used to compute these averages. The time evolution of the system is simulated, that allows one to compute time averages along the trajectories of the system. Under the ergodic assumption, these averages converge in the long time limit to the phase space averages. We study here the convergence rate of the time averages, and provide a numerical analysis of several schemes.

In a second part, we study some multiscale approaches. The chapter 6 is devoted to the numerical analysis of a method that couples an atomistic model with a continuum model : the computational domain is split into two subdomains, one described by a continuum model, the other one described by an atomistic model. In particular, we study the criterion that governs the choice, at each material point, of the model (discrete or continuous).

In the chapter 7, we study the numerical homogenization of some polycrystal models, that describe matter at the micrometric scale.

Sommaire

Introduction générale

Le travail de cette thèse a porté sur l'analyse numérique de méthodes pour la dynamique moléculaire, et sur l'étude de méthodes multi-échelles pour la simulation des matériaux.

Un matériau peut être décrit à de nombreuses échelles d'espace : à l'échelle atomique, c'est un ensemble de particules discrètes. A l'échelle macroscopique, la matière forme un continuum, c'est le domaine de la mécanique : la déformation de la matière et les contraintes sont décrites par des champs. De nombreuses échelles intermédiaires peuvent être identifiées.

Dans ce travail, nous nous intéressons à l'étude de plusieurs modèles, correspondant à des échelles d'espace différentes, et nous nous intéressons aussi à leur couplage : un modèle à une échelle fine peut être utilisé pour calculer les paramètres d'un modèle à une échelle plus grossière (c'est un *couplage séquentiel*), ou bien les deux modèles peuvent être utilisés simultanément dans le même calcul (c'est un *couplage en parallèle*).

Les différents modèles abordés sont présentés dans le premier chapitre. Décrivons-les ici rapidement, en commençant par l'échelle la plus fine.

A l'échelle atomistique, un matériau est un ensemble de particules qui interagissent entre elles via des potentiels interatomiques. Nous ne tenons pas compte ici de la nature quantique des particules, si bien que l'état du système à l'échelle atomistique (microscopique) est complètement décrit par les positions et les vitesses de toutes les particules qui le composent. Lorsqu'on considère un système comportant un grand nombre de particules, la physique statistique indique qu'on peut décrire le système à l'échelle macroscopique (globale) par un petit nombre de champs (la température, les déformations, les contraintes, ...) et que la valeur de ces champs est définie comme la moyenne de certaines fonctions sur l'ensemble des configurations microscopiques du système (ou encore *espace des phases*). Par exemple, la température est proportionnelle à la moyenne de l'énergie cinétique du système, qui est une fonction des vitesses des particules. Les quantités pertinentes ne sont donc pas les positions et les vitesses de toutes les particules, mais des moyennes dans l'espace des phases, et ce sont ces quantités que l'on cherche à connaître.

La dynamique moléculaire est une approche pour calculer ces moyennes. Dans ce modèle, l'évolution en temps, par exemple suivant les équations de Newton, de la position et de la vitesse de toutes les particules est simulée. L'utilité de la dynamique

moléculaire repose sur l'équivalence entre moyennes dans l'espace des phases et moyennes temporelles le long des trajectoires en temps ainsi générées, dans la limite d'une trajectoire de longueur infinie. Il est possible de calculer numériquement (ou au moins d'approcher) ces moyennes temporelles, ce qui permet donc d'accéder aux moyennes dans l'espace des phases. D'un point de vue mathématique, il s'agit d'étudier le comportement en temps long de solutions de systèmes dynamiques. Dans le cas le plus simple, ces systèmes sont Hamiltoniens, et nous nous intéressons dans les chapitres 2 et 3 à l'analyse numérique de plusieurs schémas pour le calcul de moyennes temporelles.

Dans d'autres cas, ces systèmes ne sont pas Hamiltoniens, mais ils conservent néanmoins une mesure. Le chapitre 4 est consacré à la construction de schémas numériques d'intégration qui conservent cette mesure, pour des équations d'évolution qui intéressent la dynamique moléculaire.

A l'autre extrémité du spectre des échelles, la mécanique décrit la matière comme un continuum. Cette approche devient discutable lorsqu'on cherche à prendre en compte des phénomènes localisés dans le matériau, et dont les dimensions caractéristiques sont proches des dimensions atomiques. La description de la propagation de fractures dans des matériaux cristallins est un exemple de tels phénomènes : loin de la fracture, le réseau atomique constituant le matériau peut être supposé parfait. Au niveau de la pointe de la fracture, des liaisons atomiques se cassent, et sur les lèvres de la fracture, le réseau atomique est fortement distordu. Bien d'autres exemples existent pour lesquels des singularités apparaissent dans le matériau. La description précise de ces singularités nécessite l'utilisation d'un modèle à une échelle atomistique, mais d'autre part, pour éviter des effets de bord, il est nécessaire de considérer des matériaux suffisamment grands.

Pour traiter ce genre de problèmes, une approche consiste à partitionner le domaine de calcul, et à utiliser un modèle atomistique dans les zones qui le nécessitent, tandis qu'un modèle de continuum est utilisé dans le reste du matériau. Un grand nombre de méthodes multi-échelles procédant de ce principe ont été proposées récemment. Ces méthodes reposent souvent sur un critère empirique permettant de choisir en chaque point du matériau le modèle (discret ou continu) qui est utilisé. Nous nous intéressons au chapitre 6 à l'analyse numérique d'un exemple très simple d'une telle méthode, en mettant l'accent sur le critère d'adaptivité.

Concernant les thèmes mentionnés ci-dessus, les contributions originales de cette thèse sont :

- l'analyse numérique des méthodes classiques de dynamique moléculaire pour le calcul de moyennes temporelles en temps long, et la proposition et l'analyse de schémas permettant une convergence plus rapide des moyennes temporelles vers leur limite en temps infini ;
- la construction, pour certains systèmes dynamiques utilisés en dynamique moléculaire, et qui conservent une mesure, d'algorithmes d'intégration qui conservent cette mesure ;

– l'analyse numérique, dans un cas simple, d'une méthode multi-échelle couplant un modèle de continuum avec un modèle atomistique.

Enfin, une partie du travail de cette thèse est consacrée à l'étude de modèles décrivant le comportement de la matière à une échelle mésoscopique. La plupart des métaux ne sont pas constitués d'un réseau atomique parfait : ils sont souvent constitués d'un agrégat de grains. Dans chaque grain, le réseau peut être considéré en première approximation comme parfait, mais son orientation varie d'un grain à un autre. Pour certains métaux, la taille caractéristique d'un grain est de l'ordre du micromètre, si bien que chaque grain peut être décrit par un modèle de continuum. La loi de comportement, qui relie les contraintes aux déformations, est homogène à l'intérieur d'un grain, mais ses caractéristiques changent d'un grain à un autre. Dans le chapitre 7, nous nous intéressons à l'homogénéisation de matériaux formés par un ensemble suffisamment grand de grains.

Cette étude est donc reliée à une approche multi-échelle, mais il s'agit ici plutôt d'un couplage séquentiel : une loi effective est déterminée à partir de la loi à l'échelle fine, qui n'est plus prise en compte par la suite.

J'ai pu bénéficier au cours de cette thèse d'un environnement à la fois académique et industriel, puisque j'ai partagé mon temps entre le CERMICS, laboratoire de mathématiques appliquées de l'ENPC, et EDF R & D. Les techniques de dynamique moléculaire sont couramment utilisées dans de nombreux projets d'EDF, tandis que les méthodes multi-échelles constituent un enjeu fort pour l'avenir. Cette collaboration industrielle m'a aussi permis de travailler sur des problèmes de mécanique moléculaire, pour lesquels il s'agit de minimiser l'énergie d'un système moléculaire par rapport aux positions des atomes le constituant. Ce travail est présenté dans le chapitre 5.

Publications dans des revues à comité de lecture

[P1] E. Cancès, F. Castella, Ph. Chartier, E. Faou, C. Le Bris, F. Legoll, G. Turinici, *High-order averaging schemes with error bounds for thermodynamical properties calculations by molecular dynamics simulations*, Journal of Chemical Physics, à paraître (2004).

[P2] E. Cancès, F. Castella, Ph. Chartier, E. Faou, C. Le Bris, F. Legoll, G. Turinici, *Long time averaging for integrable Hamiltonian dynamics*, Numerische Mathematik, soumis (2003).

[P3] F. Legoll, R. Monneau, *Designing reversible measure invariant algorithms with applications to molecular dynamics*, Journal of Chemical Physics **117**, 23 (2002), pp. 10452-10464.

[P4] X. Blanc, C. Le Bris, F. Legoll, *Analysis of a prototypical multiscale method coupling atomistic and continuum mechanics*, Mathematical Modelling and Numerical Analysis, soumis (2004).

[P5] F. Legoll, *Numerical homogenization of nonlinear viscoplastic two-dimensional polycrystals*, Computational and Applied Mathematics, à paraître (2004).

Chapitre 1

Introduction à la dynamique moléculaire et à quelques méthodes multi-échelles pour les matériaux

Le but de ce chapitre est de remettre en perspective le travail effectué au cours de cette thèse, et de le situer dans son contexte général. Les résultats originaux sont présentés dans les sections 1.1.3, 1.1.4 et dans la fin de la section 1.1.5.1 pour ce qui concerne le travail relié à la dynamique moléculaire, et dans la section 1.2.1.1 et la fin de la section 1.2.2 pour ce qui concerne les méthodes multi-échelles.

A plusieurs reprises, un problème bien connu de l'industrie nucléaire, celui de la modification des propriétés mécaniques des cuves des centrales nucléaires, est cité. C'est en effet un problème qui permet d'illustrer plusieurs des thèmes abordés au cours de cette thèse, puisqu'il fait à la fois intervenir des questions reliées à la dynamique moléculaire et des questions reliées aux approches multi-échelles. Une autre raison pour le mentionner est bien sûr ma collaboration avec EDF, qui m'a permis de découvrir ce problème et de comprendre certains de ses enjeux !

1.1 Présentation succincte de la dynamique moléculaire

Dans cette section, nous donnons une brève description de la dynamique moléculaire. On s'intéresse d'abord aux échelles des systèmes physiques décrits et aux problématiques qu'on aborde. Les outils mathématiques et les méthodes numériques sont ensuite exposés, ainsi que les difficultés rencontrées, à la fois théoriques et pratiques. Cet exposé se veut donc avant tout méthodologique. Une présentation plus physique peut être lue dans [2, 7, 16], tandis que beaucoup de questions d'implémentation sont discutées dans [2, 13].

1.1.1 Les modèles physiques et les objectifs

La dynamique moléculaire s'intéresse à des systèmes atomiques ou moléculaires, décrits dans l'approximation classique[1]. Les atomes constituant le système sont donc considérés comme des particules ponctuelles de position $q_i \in \mathbb{R}^3$ et d'impulsion $p_i \in \mathbb{R}^3$, $i = 1, \ldots, N$, où N est le nombre d'atomes du système. On s'intéresse surtout à la description de systèmes en phase liquide ou solide : l'échelle d'espace est donc la distance caractéristique entre deux atomes dans un système en phase condensée, soit de l'ordre de l'Angström (10^{-10} m). Le nombre d'atomes considérés est couramment de l'ordre de 10^5, et les équipes disposant des moyens de calcul les plus performants simulent aujourd'hui des systèmes comportant de l'ordre de 10^9 atomes.

Par définition, se donner un *état microscopique* du système, c'est se donner une valeur pour l'ensemble des variables microscopiques

$$(q, p) = (q_1, \ldots, q_N, p_1, \ldots, p_N) \in \mathbb{R}^{3N} \times \mathbb{R}^{3N}.$$

L'énergie d'un tel état est

$$H(q, p) = \sum_{i=1}^{N} \frac{p_i^2}{2m_i} + V(q_1, q_2, \ldots, q_N), \tag{1.1}$$

où le premier terme est l'énergie cinétique du système (m_i est la masse des atomes) et le second terme son énergie potentielle. Toute la physique du système est contenue dans l'expression de cette fonction V en fonction des positions q des particules. La fonction $H(q, p)$ est aussi appelée Hamiltonien du système.

L'utilisation de la dynamique moléculaire est motivée par le constat suivant : lorsqu'on s'intéresse à des systèmes à l'échelle atomique, le modèle fondamental est le modèle quantique, qui s'appuie sur l'équation de Schrödinger. Cependant, la taille des systèmes pour lesquels des calculs numériques peuvent être menés à l'aide de ce modèle est très petite : quelques centaines d'atomes si on ne s'intéresse qu'à des propriétés statiques, moins encore si on s'intéresse à la dynamique du système. Dans ce dernier cas, la durée de la trajectoire qu'on peut simuler est de l'ordre de la picoseconde.

Or, dans un grand nombre de domaines, on s'intéresse à des systèmes comportant plusieurs milliers d'atomes ou plus (typiquement plus de 100 000 pour certains systèmes biologiques), et sur des temps allant jusqu'à la microseconde. Il est donc nécessaire de faire des approximations et de changer de modèle. En plus d'une approximation déjà faite dans les modèles de chimie quantique, qui consiste à décrire les noyaux comme des particules classiques ponctuelles, on suppose que l'effet des

[1] Il s'agit donc de dynamique moléculaire *classique*, mais cet adjectif est souvent omis ; au contraire, on parle de dynamique moléculaire *ab initio* lorsqu'une description strictement quantique des électrons est présente dans le modèle.

électrons sur le système peut être pris en compte à travers un terme d'énergie potentielle qui ne dépend que de la position des noyaux. L'énergie potentielle $V(q)$ qui apparaît dans (1.1) rend donc compte des interactions entre noyaux, mais aussi des interactions entre les noyaux et l'ensemble des électrons, qui ne sont pas explicitement décrits dans un modèle de dynamique moléculaire (classique).

Les systèmes qu'on considère sont constitués de N atomes de même espèce chimique, ou bien d'un grand nombre de molécules, qui sont toutes du même type (c'est le cas si on simule de l'eau, cf. la figure 1.1) ou bien de plusieurs types différents. Certaines équipes en biochimie [41, 42, 56] simulent par exemple une portion de la membrane d'une cellule humaine, avec l'eau à l'intérieur et à l'extérieur de la cellule.

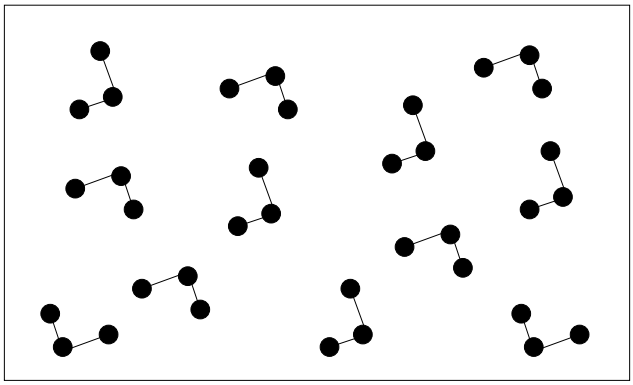

FIG. 1.1 – Exemple schématique de la simulation d'un système composé de molécules tri-atomiques.

Des exemples d'énergie potentielle V sont donnés dans [7, 16]. De façon générale, l'expression de V est de la forme

$$V(q) = \sum_{i=1}^{N} V_1(q_i) + \sum_{i=1}^{N} \sum_{j>i} V_2(q_i, q_j) + \sum_{i=1}^{N} \sum_{j>i} \sum_{k>j} V_3(q_i, q_j, q_k) + \dots \qquad (1.2)$$

Dans cette expression, le terme V_1 modélise l'interaction des particules avec un champ (électrostatique, magnétique, ...) extérieur, tandis que le terme V_2 est un terme d'interaction de paire, qui ne dépend (pour des raisons d'invariance galiléenne) que de la distance entre les deux particules i et j, soit $V_2(q_i, q_j) = V_2(|q_i - q_j|)$. Ce terme modélise soit des interactions entre des particules d'une même molécule, soit entre des particules de molécules distinctes. Sans simplification supplémentaire, son coût calcul est proportionnel à N^2.

La prise en compte d'un terme comme $\sum_{i=1}^{N} \sum_{j>i} \sum_{k>j} V_3(q_i, q_j, q_k)$ dans l'expression (1.2) rend le coût calcul proportionnel à N^3, si tous les triplets d'atomes (i, j, k) participent à la somme. Un tel coût n'est pas acceptable du point de vue du calcul numérique. C'est pourquoi la somme sur tous les triplets d'atomes n'est que très rarement prise en compte, et cette approximation semble satisfaisante du point

de vue physique (cf. [2], p. 9). Les termes à trois corps les plus fréquemment pris en compte modélisent une énergie intra-moléculaire. Ainsi, l'énergie potentielle d'une molécule composée d'au moins trois atomes peut dépendre de l'angle $\theta_{A,B,C}$ entre les deux liaisons $A-B$ et $B-C$ (cf. la figure 1.2), et cette dépendance est prise en compte à travers un terme à trois corps du type $V_3(q_A, q_B, q_C) = V_3(\theta_{A,B,C})$. Le coût d'évaluation des interactions à trois corps est donc proportionnel au nombre de molécules simulées, car on ne considère plus que des triplets d'atomes consécutifs.

Pour les termes à 4 corps, la situation est évidemment la même : seules des énergies intra-moléculaires sont prises en compte. Ainsi, lorsqu'une molécule comporte 4 atomes C, D, E et F formant une chaîne linéaire, on définit l'angle dihédral $\phi_{C,D,E,F}$ par l'angle formé entre le plan $C-D-E$ et le plan $D-E-F$ (cf. la figure 1.2). L'énergie d'une telle molécule dépend en général de l'angle dihédral, et cette dépendance est prise en compte via un terme à 4 corps du type $V_4(q_C, q_D, q_E, q_F) = V_4(\phi_{C,D,E,F})$.

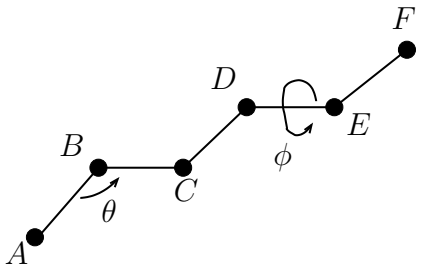

FIG. 1.2 – Sur une molécule linéaire, définition de l'angle $\theta_{A,B,C}$ entre deux liaisons et de l'angle dihédral $\phi_{C,D,E,F}$.

La dynamique moléculaire est un outil utilisé dans de nombreux domaines (en chimie, en sciences des matériaux [211, 212], en biologie [53–55]), et dont le but est essentiellement de répondre à deux questions, le calcul de l'évolution en temps d'un système moléculaire, et surtout le calcul de moyennes thermodynamiques[2]. Nous décrivons maintenant ces deux problématiques.

1.1.1.1 Calcul de la dynamique d'un système

La dynamique d'un système classique isolé est donnée par les équations de Newton

$$\frac{d\,q_i(t)}{d\,t} = \frac{p_i(t)}{m_i}, \quad \frac{d\,p_i(t)}{d\,t} = F_i(q(t)) = -\frac{\partial V}{\partial q_i}(q(t)), \tag{1.3}$$

où les forces F_i dérivent du potentiel V. Les équations (1.3) s'écrivent aussi sous la forme du système dynamique Hamiltonien

$$\frac{d\,q(t)}{d\,t} = \frac{\partial H}{\partial p}(q(t), p(t)), \quad \frac{d\,p(t)}{d\,t} = -\frac{\partial H}{\partial q}(q(t), p(t)), \tag{1.4}$$

[2]Cette seconde question est beaucoup plus abordée que la première. Nous commençons néanmoins par la description du calcul de l'évolution en temps d'un système moléculaire, car ce calcul sera nécessaire pour calculer des moyennes thermodynamiques.

où H est la fonction définie par (1.1). Le problème qu'on se pose est le calcul de l'évolution en temps du système à partir d'une configuration initiale $(q^0, p^0) \in \mathbb{R}^{3N} \times \mathbb{R}^{3N}$ donnée.

Une particularité du problème (par rapport à des calculs de trajectoires en astronautique ou en mécanique céleste) est le grand nombre N de particules à considérer, ainsi que le coût d'évaluation des forces $F(q)$, étant donnée une configuration q du système (le coût d'évaluation des interactions de paire est proportionnel à N^2). Ces deux spécificités imposent des choix sur les méthodes numériques utilisables, comme nous le verrons ci-dessous.

Une autre spécificité du problème réside dans les différentes échelles de temps qui apparaissent dans la dynamique (1.4). Les vibrations des liaisons atomiques (par exemple, la liaison $H - O$ dans la molécule d'eau) ont une période de l'ordre de la femtoseconde (10^{-15} s), tandis que certains phénomènes intéressants du point de vue des applications ont des temps caractéristiques de l'ordre de la microseconde (10^{-6} s). Cette variété des échelles de temps pose des problèmes qui ne sont pas encore aujourd'hui complètement résolus.

Donnons ici quelques exemples d'applications et de calculs qui intéressent la communauté des physiciens et des chimistes.

Un problème connu de l'industrie nucléaire est la modification des propriétés mécaniques des aciers des cuves des centrales, sous l'effet du bombardement neutronique. De façon très schématique, un neutron issu du cœur du réacteur percute un atome de l'acier de cuve, qui acquiert brutalement beaucoup d'énergie et va à son tour percuter les atomes environnants (comme dans un billard). Il s'ensuit une sorte de réaction en chaîne (on parle de cascade), jusqu'à ce que le réseau retrouve un état d'équilibre. Certaines zones sont alors devenues amorphes : les positions des atomes ne sont plus sur un réseau ordonné (cf. la figure 1.3). Ces zones constituent autant de points de faiblesse de l'acier. La dynamique moléculaire est un outil qui permet, à partir de la position et de la vitesse de l'atome excité, de calculer la configuration du réseau atomique une fois l'équilibre revenu[3].

Dans le domaine biologique, un problème non résolu à l'heure actuelle est celui du lien entre la structure 3D d'une protéine et la séquence des acides aminés qui la composent. Plus précisément, une protéine est constituée d'une chaîne linéaire d'acides aminés (il existe une vingtaine d'acides aminés différents, deux protéines sont différentes si la séquence des acides aminés dans les deux chaînes est différente). *In vivo*, cette chaîne d'acides aminés se replie sur elle-même pour former des structures 3D complexes. La séquence de nombreuses protéines est connue expérimentalement, de même que leur structure 3D, par contre le lien entre ces deux informations est encore

[3]Les processus physiques qui conduisent de la formation de ces zones amorphes à la modification macroscopique des propriétés mécaniques sont multiples et complexes. Entre le choc initial et le retour à l'équilibre, c'est-à-dire entre les deux états représentés sur la figure 1.3, il s'écoule environ 20 picosecondes, et les dimensions de la zone qui devient amorphe sont de l'ordre de 20 nanomètres. L'évolution des propriétés mécaniques met en jeu des phénomènes sur des échelles de temps et d'espace beaucoup plus grandes.

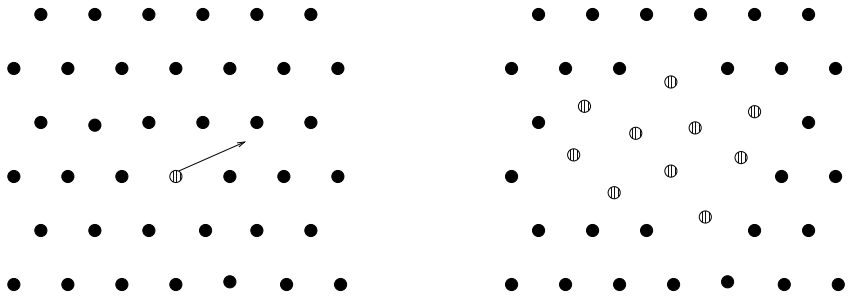

FIG. 1.3 – Cascade dans une cuve de réacteur nucléaire (représentation schématique) : à gauche, réseau parfait avant évènement, l'atome avec la flèche est celui percuté par un neutron ; à droite, réseau déformé en fin d'évènement (les atomes hachurés au centre forment une zone amorphe).

mal compris. Ce lien est important car c'est la structure 3D de la protéine qui est responsable de ses propriétés biologiques. De nombreuses équipes ont donc cherché à calculer le repliement d'une protéine, en partant par exemple d'une configuration linéaire.

Un autre problème est celui de la propagation de fractures au sein de nanotubes [206]. La fracture avance car des liaisons atomiques se cassent. L'idée est donc de simuler un nanotube (un système d'un millier d'atomes), en imposant des conditions de type Dirichlet à ses extrémités, et d'étudier à partir de quelle déformation une fracture se propage (cf. la figure. 1.4).

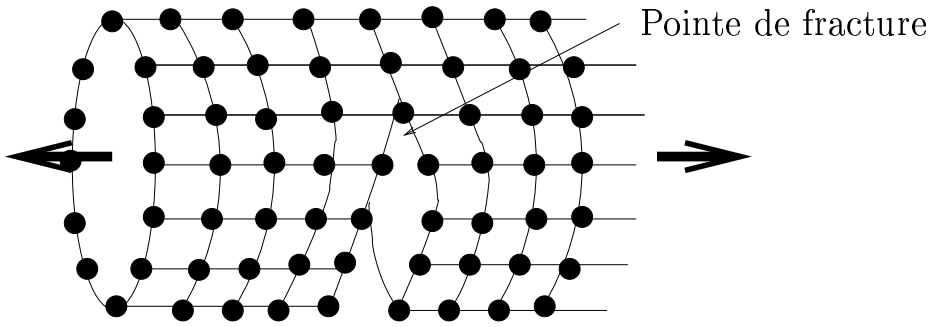

FIG. 1.4 – Nanotube mis en extension dans la direction horizontale : on impose un déplacement aux atomes du premier plan vertical à gauche et aux atomes du dernier plan vertical à droite. Une fracture se propage dans la direction transverse (comme une fermeture éclair qui s'ouvre). Les lignes fines représentent des liaisons atomiques (pour simplifier le dessin, la maille du réseau atomique est supposée carrée).

1.1.1.2 Calcul de moyennes thermodynamiques

La problématique du calcul de moyennes thermodynamiques est reliée à la physique statistique. Lorsqu'on considère des systèmes comportant un grand nombre de

particules, il convient de distinguer les variables à l'échelle microscopique de celles à l'échelle macroscopique. Les premières sont celles relatives aux particules qui forment le système (leur position, leur vitesse), tandis que les secondes sont relatives au système dans son ensemble (masse volumique, pression, ...). Comme précisé ci-dessus, se donner un état du système, c'est spécifier la valeur de toutes les variables microscopiques. Cependant, seules les informations macroscopiques ont une pertinence physique, car ce sont les seules qui sont en pratique mesurables. Sous l'hypothèse de l'équilibre thermodynamique local, la physique statistique permet de montrer que :

1. on peut décrire le système à l'échelle macroscopique par un petit nombre de champs (la masse volumique, la vitesse, la température, l'énergie interne, la pression ou le tenseur des contraintes, ...), qui sont des variables macroscopiques ;
2. la valeur de ces champs est définie comme une moyenne d'une certaine fonction (dans la terminologie physique, on parle d'*observable*) sur un ensemble d'états (de configurations) microscopiques du système ;
3. dans certains cas, il est possible de relier ces grandeurs macroscopiques entre elles pour obtenir une *loi constitutive* à l'échelle macroscopique (la loi de Mariotte, qui relie la pression, le volume occupé et la température d'un gaz parfait, en est un exemple).

L'ensemble des configurations qu'on prend en compte pour calculer la moyenne mentionnée au point 2 est relié aux conditions physiques dans lesquelles le système est étudié. Se donner un *ensemble thermodynamique*, c'est par définition se donner un ensemble d'états microscopiques, et les probabilités relatives des états pour le calcul de la moyenne.

L'exemple le plus simple est celui de l'ensemble thermodynamique NVE, pour lequel on suppose que le nombre de particules N du système, le volume V accessible aux particules et l'énergie E du système sont fixés et connus. Seules les configurations microscopiques telles que l'énergie du système soit égale à E interviennent dans la moyenne, et elles sont équiprobables.

Formalisons maintenant les différentes notions évoquées. Soit $A(q,p)$ une observable, c'est-à-dire une fonction de (q,p). La *moyenne thermodynamique* (dite aussi *moyenne statistique* ou encore *moyenne d'ensemble* dans la communauté physique) de A est définie par

$$\langle A \rangle = \frac{\displaystyle\int_\Omega A(q,p)\, d\mu}{\displaystyle\int_\Omega d\mu}, \tag{1.5}$$

où la mesure $d\mu$ et le domaine $\Omega \subset \mathbb{R}^{3N} \times \mathbb{R}^{3N}$ dépendent de l'ensemble thermodynamique dans lequel on travaille. Le domaine Ω, appelé *espace des phases* du système, est l'espace des configurations (q,p) qui lui sont accessibles[4]. Travailler

[4]Dans la communauté mathématique, on désigne ainsi (1.5) sous le terme de moyenne dans l'espace des phases.

dans l'ensemble NVE correspond à faire le choix

$$\Omega = \Omega_{NVE} = \left\{(q,p) \in V \times \mathbb{R}^{3N};\ H(q,p) = E\right\} \tag{1.6}$$

et

$$d\mu = d\mu_{NVE} = \frac{d\sigma}{\|\nabla H\|_2}, \tag{1.7}$$

où $d\sigma$ est la mesure induite sur la variété $H^{-1}(E)$ par la mesure euclidienne $dq\,dp$ de $\mathbb{R}^{3N} \times \mathbb{R}^{3N}$ et où $\|\cdot\|_2$ est la norme euclidienne. Pour la distinguer de la valeur calculée dans d'autres ensembles, nous noterons $\langle A \rangle_{NVE}$ la quantité (1.5) avec les choix (1.6) et (1.7).

Donnons maintenant quelques exemples d'observable A. Pour un système tel qu'un polymère ou un alcane, une grandeur importante est la distance entre le premier atome de la chaîne et le dernier (cf. la figure 1.5) : l'observable s'écrit donc $A(q,p) = |q_1 - q_N|$.

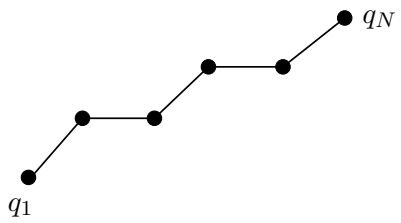

FIG. 1.5 – Représentation schématique d'un alcane ou d'un polymère : chaque particule représente un groupe CH_2 dans le premier cas et un monomère dans le second cas.

Dans un liquide, la pression est définie par (1.5) avec

$$A(q,p) = \frac{1}{3V} \sum_{i=1}^{N} \left(\frac{p_i^2}{m_i} + q_i \cdot F_i(q) \right),$$

où $F_i(q)$ sont les forces.

Dans un solide, lorsque les déformations sont suffisamment petites, on peut faire l'approximation de l'élasticité linéaire : le champ de contrainte σ (qui est une variable macroscopique qu'on peut calculer sous la forme (1.5) pour un certain A) est relié au champ de gradient de déformation ε (lui aussi macroscopique et calculable par (1.5)) par une relation du type

$$\sigma = \Lambda : \varepsilon.$$

Le tenseur d'élasticité Λ (tenseur 9×9) peut être calculé par dynamique moléculaire [51].

Remarque 1.1.1 *Il est possible de travailler dans d'autres ensembles thermodynamiques. Travailler dans l'ensemble NVT signifie que le nombre de particules N, le volume V qui leur est accessible et la température T du système sont connus, mais*

que l'énergie du système n'est pas connue. La moyenne d'une observable est alors définie par (1.5), avec les choix

$$\Omega = \Omega_{NVT} = V \times \mathbb{R}^{3N} \tag{1.8}$$

et

$$d\mu = d\mu_{NVT} = \exp\left(-\frac{H(q,p)}{k_B T}\right)\, dq\, dp, \tag{1.9}$$

où k_B est la constante de Boltzmann, et nous noterons cette moyenne $\langle A \rangle_{NVT}$.

Les raisons qui motivent les choix de Ω et de $d\mu$, la description d'autres ensembles thermodynamiques et l'influence du choix de l'ensemble thermodynamique de travail sont détaillées dans [4]. Mentionnons simplement ici qu'à la limite d'un nombre de particules N infini, tous les ensembles sont équivalents (la moyenne d'une observable devient indépendante du choix de l'ensemble thermodynamique dans lequel elle est calculée), mais que pour les valeurs de N aujourd'hui considérées dans la pratique, il peut y avoir des différences non négligeables entre deux moyennes calculées dans deux ensembles thermodynamiques différents.

Remarque 1.1.2 *Certaines grandeurs importantes, comme le coefficient de diffusion dans un liquide, ne s'expriment pas sous la forme (1.5), mais sous une forme un peu différente. Soit*

$$C(t) = \frac{\displaystyle\int_\Omega B\left(\Phi_t(q,p),(q,p)\right)\, d\mu}{\displaystyle\int_\Omega d\mu}, \tag{1.10}$$

où Ω et $d\mu$ ont la même signification que dans l'expression (1.5), où Φ_t est le flot associé à la dynamique du système (c'est l'état du système à l'instant t sachant qu'il est en (q,p) à l'instant initial), et où B est une fonction définie sur $\mathbb{R}^{6N} \times \mathbb{R}^{6N}$. Dans l'ensemble NVE, l'espace des phases Ω et la mesure $d\mu$ sont donnés par (1.6) et (1.7) et Φ_t est le flot du système (1.4). La quantité $C(t)$ est appelée coefficient d'auto-corrélation (on regarde les corrélations de deux états du système séparés par un intervalle de temps t). Par exemple, le coefficient d'autocorrélation en vitesse $C_v(t)$ est défini par (1.10) avec

$$B((q',p'),(q,p)) = \frac{1}{N}\sum_{i=1}^{N} \frac{p_i' \cdot p_i}{m_i^2}.$$

Le coefficient de diffusion d'un liquide, qui est un paramètre macroscopique, est donné par

$$D = \frac{1}{3}\int_0^\infty C_v(t)\, dt.$$

Le calcul de coefficients d'auto-corrélation est un problème important en pratique, nous ne l'abordons pas plus avant ici, en préférant nous concentrer sur le cas simple qui est le calcul d'une moyenne thermodynamique définie par (1.5).

Le problème est donc de calculer l'intégrale (1.5), ce qui n'est pas simple à cause de la grande dimension ($6N$) dans laquelle on travaille. Pour cette raison, il n'est pas envisageable d'utiliser les méthodes classiques comme celles des points de Gauss. Deux approches sont utilisées pour le calcul de (1.5) :

1. l'approche Monte Carlo d'une part,
2. l'approche dynamique moléculaire d'autre part.

L'approche Monte Carlo consiste à générer aléatoirement une suite de points $(q_n, p_n)_{n\in\mathbb{N}}$ qui échantillonnent l'espace des phases Ω suivant la densité de probabilité $C\, d\mu$, où C est une constante de normalisation. Par construction, les points $(q_n, p_n)_{n\in\mathbb{N}}$ sont décorrélés les uns des autres. Historiquement, c'est cette méthode qui a été la première mise en œuvre pour le calcul de (1.5) (cf. [50]). Depuis, elle fait l'objet d'une importante littérature, à la fois dans le domaine probabiliste [10] et dans les domaines de chimie ou de physique.

Passons maintenant à l'approche dynamique moléculaire, qui est celle qui a été étudiée dans le cadre de cette thèse. Le cas le plus simple est celui du calcul d'une moyenne dans l'ensemble NVE (le cas de l'ensemble NVT est traité dans la section 1.1.5). Cette approche repose sur le caractère supposé ergodique du système Hamiltonien (1.4) (nous revenons sur ce point ci-dessous). Sous cette hypothèse d'ergodicité, on a, pour presque toute condition initiale $(q_0, p_0) \in \Omega_{NVE}$,

$$\langle A\rangle_{NVE} = \lim_{T\to\infty} \frac{1}{T} \int_0^T A(q(t), p(t))\, dt, \tag{1.11}$$

où $(q(t), p(t))$ est la solution de (1.4) de condition initiale (q_0, p_0). La condition initiale est choisie telle que $H(q_0, p_0) = E$, où E est l'énergie à laquelle on souhaite travailler, et qui apparaît dans la définition de $\langle A\rangle_{NVE}$ (cf. (1.5), (1.6) et (1.7)). Comme l'énergie $H(q, p)$ est un invariant de (1.4), on voit que la trajectoire $(q(t), p(t))$ appartient à l'ensemble Ω_{NVE}. L'équivalence (1.11) entre moyenne d'ensemble et moyenne temporelle signifie que la trajectoire $t \mapsto (q(t), p(t))$ "remplit" l'espace des phases Ω_{NVE} avec la densité de probabilité $d\mu_{NVE}$. Le calcul de $\langle A\rangle_{NVE}$, qui est la quantité qui nous intéresse *in fine*, se fait donc par le calcul d'une trajectoire du système et d'une moyenne temporelle sur cette trajectoire. Ces deux étapes sont réalisables en pratique.

Les premiers systèmes à avoir été étudiés par l'approche dynamique moléculaire sont des systèmes avec interaction de paire (l'énergie potentielle V définie en (1.2) ne comporte que le terme V_2), tout d'abord dans le cas des sphères dures [39], puis dans le cas d'un potentiel de Lennard-Jones [45, 46, 58, 59]. Dans le premier cas, le potentiel est défini par

$$V_2(r) = 0 \text{ si } r \le r_c, \quad V_2(r) = +\infty \text{ sinon},$$

où r_c est un rayon de coupure représentant le rayon des particules qui interagissent. Dans le second cas, plus réaliste du point de vue physique, le potentiel est donné

par

$$V_2(r) = 4\varepsilon \left(\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6} \right), \tag{1.12}$$

où ε et σ sont des paramètres du modèle. Une question classique consiste à étudier les courbes de changement de phase d'un tel système (courbe de coexistence liquide-vapeur, température au point triple, ...).

Revenons sur l'hypothèse d'ergodicité. A notre connaissance, les seuls systèmes pour lesquels elle est démontrée sont des systèmes de sphères dures (le billard de Sinaï [34], cf. aussi [18,26,32,33]). Donc, pour la majorité des systèmes moléculaires étudiés, le caractère ergodique n'est ni démontré, ni infirmé. Les tests numériques semblent montrer que cette hypothèse est vérifiée, au sens où les moyennes temporelles convergent vers une limite indépendante de la condition initiale. Remarquons qu'une condition nécessaire pour que l'égalité (1.11) soit vraie est que la dynamique ne préserve qu'un seul invariant, l'énergie. S'il existe d'autres invariants (l'impulsion totale, le moment cinétique, ...), alors la trajectoire reste sur la surface d'isovaleur de ces invariants. Dans (1.11), la moyenne d'ensemble sur Ω_{NVE} doit alors être remplacée par la moyenne sur la surface d'isovaleur de tous les invariants.

Les deux objectifs de la dynamique moléculaire que nous venons de décrire, le calcul de l'évolution en temps d'un système et le calcul de moyennes thermodynamiques, font donc appel au calcul d'une solution du système dynamique (1.4). La motivation est cependant différente suivant l'objectif poursuivi. Dans le premier cas, détaillé dans la section 1.1.1.1, l'objet d'étude est bien la dynamique du système : on cherche à comprendre les mouvements collectifs d'atomes, les mécanismes dynamiques de repliement de protéines, ... Lorsque l'objectif est le calcul d'une moyenne d'ensemble, suivre la trajectoire du système n'est qu'un moyen parmi d'autres pour échantillonner l'espace des phases suivant la densité de probabilité $d\mu$. L'analyse numérique des deux cas n'est pas la même, car on peut faire une erreur sur la trajectoire qui disparaît lorsqu'on considère une moyenne sur cette trajectoire[5].

1.1.2 Intégration en temps d'un système Hamiltonien

1.1.2.1 Algorithmes symplectiques

Dans cette partie, nous nous intéressons aux méthodes numériques pour calculer une solution approchée de (1.4). Ce calcul intervient en effet dans les deux objectifs détaillés ci-dessus. Le système (1.4) d'équations différentielles ordinaires (EDO) possède la propriété fondamentale d'être un système Hamiltonien, et les méthodes numériques spécifiques à ce cas font l'objet d'une importante littérature (on pourra en particulier consulter [5, 9]). La notion centrale est celle de symplecticité.

[5] Cette remarque ne s'applique pas au calcul de coefficients d'auto-corrélation (1.10), dans la définition desquels la trajectoire du système intervient explicitement.

Définition 1.1.1 *Soit $\chi(x)$ une application de $\mathbb{R}^{2d}$ dans $\mathbb{R}^{2d}$, de classe C^1, et soit $jac\,\chi$ sa matrice jacobienne. On dit que l'application χ est symplectique si*

$$\forall x \in \mathbb{R}^{2d}, \quad (jac\,\chi(x))^T \; J \; (jac\,\chi(x)) \; = \; J,$$

où la matrice J est définie par

$$J = \begin{pmatrix} 0 & I_d \\ -I_d & 0 \end{pmatrix},$$

où I_d est la matrice identité de taille $d \times d$.

On peut montrer qu'un système dynamique est Hamiltonien si et seulement si son flot est symplectique. La symplecticité est donc une caractéristique des flots des systèmes Hamiltoniens, et il est naturel de chercher à construire des schémas numériques dont le flot est lui aussi symplectique[6]. La définition d'un schéma numérique symplectique est la suivante :

Définition 1.1.2 *Soit Γ une fonction de $\mathbb{R}^{2d}$ dans $\mathbb{R}^{2d}$, on considère la dynamique*

$$\frac{d\,x(t)}{d\,t} = \Gamma(x(t)), \quad x(0) = x_0. \tag{1.13}$$

On se donne un schéma d'intégration et un pas de temps Δt. Soit $\Psi_{\Delta t}$ le flot numérique*, c'est-à-dire la fonction telle que le schéma numérique s'écrive*

$$x_{n+1} = \Psi_{\Delta t}(x_n),$$

où x_n est une approximation de $x(n\Delta t)$. Si la fonction $\Psi_{\Delta t}$ est symplectique (pour tout Δt) dès que la dynamique (1.13) est Hamiltonienne, alors la méthode numérique est dite symplectique.

Avant de donner des exemples de schémas numériques symplectiques, expliquons maintenant leurs qualités. L'énergie $H(q,p)$ est un invariant du système (1.4), et on s'intéresse à la préservation, par le schéma numérique, de cet invariant. Lorsqu'on utilise un schéma non symplectique (schéma d'Euler, ou bien schéma de Runge-Kutta, même d'ordre élévé), on constate que l'énergie n'est pas bien conservée : la quantité $H(q_n,p_n)$, où (q_n,p_n) est l'approximation fournie par le schéma numérique de $(q(n\Delta t),p(n\Delta t))$, s'éloigne de l'énergie initiale $H_0 = H(q_0,p_0)$ à une vitesse exponentielle en la longueur $n\Delta t$ de la trajectoire simulée. Au contraire, lorsque le schéma utilisé est symplectique, on constate une très bonne conservation de l'énergie : la quantité $H(q_n,p_n)$ fluctue autour de l'énergie initiale H_0, sans s'en éloigner, et ceci pendant un temps très long. Plus précisément, on a le théorème suivant (cf. [9], théorème 8.1 p. 312) :

[6] Cet intérêt pour les schémas symplectiques est justifié par leurs excellentes propriétés numériques qui sont décrites ci-dessous.

Théorème 1.1.1 *On suppose que $H : D \subset \mathbb{R}^{3N} \times \mathbb{R}^{3N} \to \mathbb{R}$ est analytique, et que le schéma numérique utilisé pour intégrer (1.4) est un schéma symplectique de pas Δt et d'ordre r_0. On note $(q_n, p_n)_{n\in\mathbb{N}}$ la trajectoire numérique fournie par le schéma. Si la solution numérique reste dans un compact $K \subset D$, alors il existe Δt_0 et C_0 tels que, pour tout $\Delta t \leq \Delta t_0$, on a*

$$\forall n \text{ t.q. } n\Delta t \leq \exp\left(\frac{\Delta t_0}{2\Delta t}\right), \ |H(q_n, p_n) - H(q_0, p_0)| \leq C_0 \Delta t^{r_0}.$$

La preuve de ce théorème repose sur l'analyse rétrograde, dont nous expliquons ici brièvement l'idée. Etant donné le système dynamique (1.13), qui à ce stade là n'est pas nécessairement Hamiltonien, on note $\Psi_{\Delta t}$ le flot numérique fourni par un schéma de pas Δt. On cherche maintenant une fonction $\Gamma_{\Delta t}(x)$ telle que le flot exact, noté $\Phi_{\Delta t}(x_0, t)$, de l'équation

$$\frac{d\,x(t)}{d\,t} = \Gamma_{\Delta t}(x(t)) \tag{1.14}$$

vérifie

$$\forall x_0, \quad \Phi_{\Delta t}(x_0, \Delta t) = \Psi_{\Delta t}(x_0).$$

Il est en général impossible de trouver une telle fonction $\Gamma_{\Delta t}(x)$, par contre on peut trouver une fonction $\Gamma_{\Delta t}(x)$ telle que, étant donné un compact K, le flot exact de *l'équation modifiée* (1.14) vérifie

$$\forall x_0 \in K, \quad \Phi_{\Delta t}(x_0, \Delta t) = \Psi_{\Delta t}(x_0) + O\left(\Delta t\, e^{-c/\Delta t}\right) \tag{1.15}$$

pour une certaine constante $c > 0$ indépendante de Δt. Donc, à des termes exponentiellement petits en Δt près, le flot numérique $\Psi_{\Delta t}$ est égal au flot exact de l'équation modifiée. Si de plus le système initial (1.13) est Hamiltonien (de Hamiltonien H) et que la méthode numérique utilisée pour l'intégrer est symplectique et d'ordre r_0, alors le système dynamique (1.14) est lui aussi Hamiltonien, de Hamiltonien $H_{\Delta t}$ tel que

$$H_{\Delta t}(q, p) = H(q, p) + O\left(\Delta t^{r_0}\right), \tag{1.16}$$

uniformément en (q, p) sur tout compact. On étudie maintenant la conservation de l'énergie. On voit que

$$\begin{aligned}
H_{\Delta t}(q_n, p_n) - H_{\Delta t}(q_0, p_0) &= \sum_{i=0}^{n-1} H_{\Delta t}(q_{i+1}, p_{i+1}) - H_{\Delta t}(q_i, p_i) \\
&= \sum_{i=0}^{n-1} H_{\Delta t}(\Psi_{\Delta t}(q_i, p_i)) - H_{\Delta t}(q_i, p_i) \\
&= \sum_{i=0}^{n-1} H_{\Delta t}(\Phi_{\Delta t}(q_i, p_i, \Delta t)) - H_{\Delta t}(q_i, p_i) \\
&\quad + O\left(e^{-c/\Delta t}\right) \\
&= O\left(e^{-c/\Delta t}\right),
\end{aligned}$$

où on a utilisé à la dernière ligne le fait que le flot $\Phi_{\Delta t}$ conserve $H_{\Delta t}$. En utilisant (1.16), on voit donc que l'énergie H est conservée par le schéma numérique à une erreur $O\left(\Delta t^{r_0}\right)$ près.

Les critères à considérer lors du choix d'un schéma numérique pour l'intégration de (1.4) sont donc les suivants :

- la symplecticité du schéma, qui permet de préserver sur le long terme l'énergie du système ;
- le caractère explicite du schéma ; ce critère est motivé par le grand nombre de particules qu'il faut simuler, qui rend les schémas implicites difficiles d'emploi. La contrepartie est une contrainte de stabilité forte sur le pas de temps.
- le nombre d'évaluation des forces $F(q)$ nécessaire à chaque pas. Ce terme est en effet très cher à calculer. C'est aussi pour cette raison que les schémas dits "multi-derivatives", qui font appel à la dérivée du membre de droite du système dynamique (dans le cas présent, la dérivée des forces par rapport aux positions), ne sont pas utilisables.

Dans la plupart des simulations de dynamique moléculaire, c'est le schéma velocity Verlet [58] qui est utilisé. On introduit la matrice diagonale

$$M = \mathrm{diag}(m_1, m_1, m_1, m_2, m_2, m_2, \ldots, m_N, m_N, m_N),$$

qui est de dimension $3N$. Notant Δt le pas de temps, le schéma velocity Verlet, appliqué aux équations (1.3), s'écrit

$$\begin{aligned} p_{n+1/2} &= p_n + \frac{\Delta t}{2} F(q_n), \\ q_{n+1} &= q_n + \Delta t\, M^{-1}\, p_{n+1/2}, \\ p_{n+1} &= p_{n+1/2} + \frac{\Delta t}{2} F(q_{n+1}). \end{aligned}$$

Ce schéma est très populaire car il satisfait toutes les contraintes exposées ci-dessus. De plus, il est facile à implémenter, assez précis (c'est un schéma d'ordre 2) et, enfin, il préserve une autre propriété qualitative des équations (1.3), qui est leur réversibilité en temps.

La construction de ce schéma repose sur un argument de splitting d'opérateur (cf. [9], pp. 43) entre d'une part, l'opérateur $p\frac{\partial}{\partial q}$, et d'autre part l'opérateur $F(q)\frac{\partial}{\partial p}$. Le splitting considéré est symétrique, ce qui assure la réversibilité en temps du schéma numérique. Le schéma velocity Verlet est aussi une méthode de Runge-Kutta partitionnée sur le Hamiltonien séparable (1.1) (cf. [9] pp. 34).

L'analyse numérique des schémas d'intégration des systèmes Hamiltoniens est aujourd'hui bien comprise (cf. [9]). Nous décrivons maintenant deux difficultés, à la fois théorique et numérique, qu'on rencontre dans le domaine de la dynamique moléculaire.

1.1.2.2 Evaluation des forces à longue portée

L'étape la plus coûteuse du calcul de la trajectoire d'un système en dynamique moléculaire est l'évaluation des forces provenant du terme

$$\sum_{i=1}^{N} \sum_{j>i} V_2(|q_i - q_j|) \tag{1.17}$$

dans l'expression (1.2). Sans plus de simplifications, l'évaluation de ce terme a un coût proportionnel à N^2. Il faut ici distinguer deux types de potentiel.

On définit souvent les potentiels de paire $V_2(r)$ à courte portée comme ceux, en 3D, qui décroissent plus vite que $1/r^3$ quand r tend vers $+\infty$ (le potentiel de Lennard-Jones (1.12) en est un exemple). Pour ces potentiels, il est possible d'utiliser un rayon de coupure r_c : on ne considère dans la somme (1.17) que les particules i et j telles que $|q_i - q_j| < r_c$. Comme le potentiel décroît rapidement, l'erreur commise est considérée comme faible (l'influence de la troncature est discutée dans [2], pp. 24-29). Avec cette simplification, et en tenant à jour la liste[7] des particules i et j proches au sens ci-dessus, on réduit le coût calcul de $O(N^2)$ à $O(N)$.

Le problème est plus difficile pour des potentiels à longue portée, comme les potentiels électrostatiques, pour lesquels il faut calculer une somme du type

$$\sum_{i=1}^{N} \sum_{j>i} \frac{Z_i Z_j}{|q_i - q_j|}, \tag{1.18}$$

où Z_i est la charge électrostatique de l'atome i. L'expérience numérique montre qu'on commet une erreur significative en utilisant un rayon de coupure. Deux méthodes (au moins) ont été développées pour traiter ce cas, et sont couramment utilisées en dynamique moléculaire : la méthode Fast Multipole Method (FMM) (cf. [78, 79]) ou une méthode fondée sur les sommes d'Ewald (cf. [80] pour les aspects mathématiques et [74, 75, 81, 87, 89] pour des exemples de mise en œuvre).

La méthode FMM ramène le coût calcul de $O(N^2)$ à $O(N \ln N)$. Cependant, le potentiel (1.18), qui est une fonction continue des positions q, est alors approché par un potentiel discontinu. En effet, si deux noyaux sont éloignés l'un de l'autre, leur interaction est calculée de façon approchée, alors que s'ils sont proches, on conserve l'expression exacte. Or, l'analyse rétrograde (cf. le théorème 1.1.1 ci-dessus) comme la théorie KAM[8] dont on a besoin ci-dessous (cf. la section 1.1.3.1) s'appuient sur l'analyticité du Hamiltonien. Ces deux théories ne s'appliquent donc pas dans le cadre d'un potentiel calculé par la méthode FMM.

[7] On parle des listes de Verlet [58], du nom de la première personne à avoir proposé et implémenté cette idée.

[8] du nom de ses auteurs, Kolmogorov, Arnold et Moser ; cf. par exemple [5], et [9] pp. 327.

1.1.2.3 Présence de plusieurs échelles de temps

Pour intégrer (1.4), comme on utilise des méthodes explicites, le choix du pas de temps doit respecter une contrainte de stabilité, qui fait intervenir la plus haute fréquence présente dans le système.

Dans les systèmes moléculaires, la plus haute fréquence présente est celle associée à la vibration de la longueur des liaisons atomiques : la période correspondante est de l'ordre de la femtoseconde, et l'amplitude de vibration est très petite. Du point de vue de la modélisation, deux approches existent :

- soit l'énergie potentielle inclut un terme du type

$$\frac{1}{2}\omega^2 \left(|q_i - q_j| - r_{eq}\right)^2,$$

  où i et j sont deux atomes reliés par une liaison dont la longueur d'équilibre est r_{eq}. Alors la vibration de cette liaison est explicitement décrite, ce qui nécessite de choisir un pas de temps de l'ordre de la femtoseconde.
- soit la liaison est modélisée comme une contrainte holonome, ce qui permet de prendre des pas de temps plus grands mais transforme le système (1.4), qui est un système d'EDO, en un système algébro-différentiel, plus compliqué à résoudre numériquement.

Les deux approches existent : dans le second cas, l'algorithme fréquemment utilisé est l'algorithme Rattle, qui est symplectique [44] (cf. aussi [40, 72]).

Même si cette fréquence très élevée est supprimée, les systèmes considérés en pratique font tout de même intervenir une multitude d'échelles de temps. De plus, en pratique, le cas est fréquent où les forces qui varient lentement sont celles qui sont chères à évaluer. Plusieurs méthodes ont été développées pour traiter ce problème, qui est fondamentalement un problème d'EDO avec des oscillations rapides. Les méthodes de Gautschi et de Deuflhard (revisitées par Hairer, Hochbruck et Lubich, cf. [9] pp. 417 pour une analyse globale) s'appuient sur la connaissance exacte de la fréquence élevée.

Nous décrivons maintenant brièvement la méthode "Impulse", proposée en 1992, et qui a été très employée dans la communauté chimiste (cf. par exemple [82, 105] pour un emploi de cette méthode en combinaison avec la méthode FMM et la méthode des sommes d'Ewald ; un autre exemple d'application est décrit dans [90]), jusqu'à ce que ses limitations soient observées puis comprises [73, 76, 77, 93]. Supposons que les forces $F(q)$ soient décomposées en un terme rapide $F_r(q)$ et un terme lent $F_l(q)$, soit $F(q) = F_r(q) + F_l(q)$. En s'appuyant sur les idées de splitting d'opérateur, Tuckerman *et al.* ont proposé le schéma suivant [100] :

$$(q_n^a, p_n^a) = \exp\left(\frac{\Delta t}{2} F_l(q) \frac{\partial}{\partial p}\right)(q_n, p_n), \tag{1.19}$$

$$(q_n^b, p_n^b) = \left(\Psi^r_{\Delta t/m}\right)^m (q_n^a, p_n^a), \tag{1.20}$$

$$(q_{n+1}, p_{n+1}) = \exp\left(\frac{\Delta t}{2} F_l(q) \frac{\partial}{\partial p}\right)(q_n^b, p_n^b), \tag{1.21}$$

où $\Psi^r_{\Delta t/m}$ est n'importe quel schéma, de petit pas de temps $\Delta t/m$, qui intègre les équations de Newton dans lesquelles seules les forces rapides $F_r(q)$ sont prises en compte (en pratique, on choisit ici encore l'algorithme de velocity Verlet).

Le problème avec cette méthode est que des résonances apparaissent. On comprend très bien pourquoi en supposant que la solution exacte du problème s'écrit

$$q(t) = q_m(t) + \frac{1}{\Omega}\cos(\Omega t),$$

où $q_m(t)$ est une fonction dont les temps caractéristiques de variation sont très supérieurs à $2\pi/\Omega$, qui est la période du mouvement rapide. La normalisation devant le coefficient de haute fréquence assure que l'énergie reste bornée même si Ω est grand. La variable q oscille très rapidement (avec une pulsation Ω) autour d'une position moyenne qui est $q_m(t)$, qui elle-même évolue beaucoup plus lentement. Si le pas de temps Δt est égal à une période $2\pi/\Omega$ du mouvement rapide, alors les forces $F_l(q)$ sont évaluées aux temps $k2\pi/\Omega$, $k \in \mathbb{N}$. A ces instants là, la valeur de $q(t)$ vaut $q_m\left(k\frac{2\pi}{\Omega}\right) + \frac{1}{\Omega}$, alors que la valeur moyenne de q, une fois les oscillations de forte fréquence intégrées, vaut $q_m\left(k\frac{2\pi}{\Omega}\right)$. Les forces F_l sont donc systématiquement calculées en dehors de la trajectoire moyenne, et cette erreur systématique est la cause des instabilitées observées.

S'inspirant de la méthode Impulse, la méthode "Mollified Impulse Method" a ensuite été proposée [83] (cf. aussi [85]). Elle repose sur l'idée de ne pas calculer les forces lentes en la valeur de la position trouvée après l'étape (1.20), mais sur une position moyenne.

1.1.3 Calcul de moyennes d'ensemble : analyse numérique dans un cadre simple

On passe maintenant à la deuxième utilisation standard de la dynamique moléculaire, c'est-à-dire le calcul de moyennes d'ensemble. Le problème a été exposé dans la section 1.1.1.2 et la formule fondamentale sur laquelle on s'appuie est la formule (1.11). La méthode standard pour calculer $\langle A\rangle_{NVE}$ consiste à choisir un temps de simulation T, puis à approcher la moyenne temporelle

$$\langle A\rangle(T) = \frac{1}{T}\int_0^T A(q(t),p(t))\,dt$$

par la somme de Riemann

$$\langle A\rangle^{Rie}_{num}(T) = \frac{1}{N_m}\sum_{n=0}^{N_m-1} A(q_n,p_n), \tag{1.22}$$

où (q_n,p_n) est la trajectoire numérique fournie par un algorithme de pas $\Delta t = T/N_m$ appliqué aux équations (1.3). La moyenne $\langle A\rangle_{NVE}$ est donc approchée par (1.22).

1.1.3.1 La méthode standard

Un des résultats de cette thèse (cf. [P1, P2]) concerne l'analyse numérique de la méthode décrite ci-dessus. Il a été établi en collaboration avec Eric Cancès, François Castella, Philippe Chartier, Erwan Faou, Claude Le Bris et Gabriel Turinici, au sein de PRESTISSIMO, qui est une Action de Recherche Coopérative de l'INRIA. Nous nous sommes placés dans le cadre de systèmes Hamiltoniens *complètement intégrables* (nous revenons dans la section 1.1.3.3 sur la validité de cette hypothèse en pratique). Par définition[9], un tel système dynamique sur les variables $(q,p) \in \mathbb{R}^{3N} \times \mathbb{R}^{3N}$ admet $3N$ invariants, qu'on note $I_1(q,p), \ldots, I_{3N}(q,p)$, et qui vérifient

$$\forall j_1, j_2, \quad \nabla_q I_{j_1} \cdot \nabla_p I_{j_2} = \nabla_p I_{j_1} \cdot \nabla_q I_{j_2}.$$

On pose

$$S(q,p) = \left\{ (x,y) \in \mathbb{R}^{3N} \times \mathbb{R}^{3N} \text{ t.q. } \forall j \in [1,3N],\ I_j(x,y) = I_j(q,p) \right\}.$$

Soit (q_0,p_0) une condition initiale : par définition, la trajectoire $(q(t),p(t))$ du système Hamiltonien (1.4) issue de cette condition initiale reste sur la variété $S(q_0,p_0)$, et la moyenne d'ensemble qu'il faut considérer est la moyenne sur $S(q_0,p_0)$. Posons donc $I_j^0 = I_j(q_0,p_0)$, si bien que

$$S(q_0,p_0) = \left\{ (x,y) \in \mathbb{R}^{3N} \times \mathbb{R}^{3N} \text{ t.q. } \forall j \in [1,3N],\ I_j(x,y) = I_j^0 \right\}. \tag{1.23}$$

La moyenne d'ensemble qu'on cherche à calculer est

$$\langle A \rangle_{I_1^0,\ldots,I_{3N}^0} = \frac{\displaystyle\int_{S(q_0,p_0)} A(q,p)\, d\mu_{I_1^0,\ldots,I_{3N}^0}}{\displaystyle\int_{S(q_0,p_0)} d\mu_{I_1^0,\ldots,I_{3N}^0}}, \tag{1.24}$$

où $d\mu_{I_1^0,\ldots,I_{3N}^0}$ est la mesure invariante sur $S(q_0,p_0)$ (cette mesure invariante généralise (1.7)). Sous une hypothèse de non-résonance, et supposant que le Hamiltonien $H(q,p)$ et l'observable A sont analytiques, on peut montrer (cf. [3] p. 287) que

$$\langle A \rangle_{I_1^0,\ldots,I_{3N}^0} = \frac{1}{T} \int_0^T A(q(t),p(t))\, dt + O\left(\frac{1}{T}\right), \tag{1.25}$$

où $(q(t),p(t))$ est la solution de (1.4) de condition initiale (q_0,p_0). On rappelle ici brièvement les étapes de la preuve, ce qui nous sera utile dans la suite :

1. Dans le cas d'un oscillateur harmonique en une dimension, le Hamiltonien s'écrit $H(q,p) = \dfrac{p^2}{2m} + \dfrac{1}{2}\omega^2 q^2$, et la trajectoire s'écrit

$$q(t) = C_0 \cos(\omega t + \phi), \quad p(t) = -C_0\, \omega \sin(\omega t + \phi).$$

[9] Une définition précise d'un système Hamiltonien complètement intégrable est donnée dans [3, 5, 9]. Nous retenons ici les hypothèses les plus fortes.

Comme la fonction $\theta \mapsto A(C_0 \cos\theta, -C_0\,\omega \sin\theta)$ est périodique, on peut la décomposer en série de Fourier, soit

$$A(C_0 \cos\theta, -C_0\,\omega \sin\theta) = \sum_{k\in\mathbb{Z}} a_k \exp(ik\theta). \tag{1.26}$$

En insérant ce développement dans la moyenne temporelle de A, on voit[10] que

$$\begin{aligned}\frac{1}{T}\int_0^T A(q(t),p(t))\,dt &= \frac{1}{T}\int_0^T \sum_{k\in\mathbb{Z}} a_k \exp(ik\phi)\exp(ik\omega t)\,dt \\ &= a_0 + \sum_{k\neq 0} a_k \exp(ik\phi)\frac{1}{T}\frac{e^{ik\omega T}-1}{ik\omega}.\end{aligned}$$

Le terme a_0 est la moyenne d'ensemble qu'on souhaite calculer, et le second terme de la somme est donc l'erreur, qui décroît[11] comme $1/T$, ce qui donne donc l'estimation (1.25).

2. On traite maintenant le cas d'un système Hamiltonien intégrable. On note $\mathbb{T} = \mathbb{R}/2\pi$ le tore $[0, 2\pi]$. Puisque le système (1.4) est intégrable, il existe un changement de variables symplectique local ψ qui permet de passer des variables initiales $(q,p) \in \mathbb{R}^{3N}\times\mathbb{R}^{3N}$ à des variables $(a,\theta) \in \mathbb{R}^{3N}\times\mathbb{T}^{3N}$ dites variables action-angle telles que, si $(q(t),p(t))$ est solution de (1.4), alors

$$(a(t),\theta(t)) = \psi^{-1}(q(t),p(t))$$

vérifie l'équation

$$\frac{d\,a(t)}{d\,t} = 0, \quad \frac{d\,\theta(t)}{d\,t} = \omega(a(t)).$$

Par construction, ψ est périodique en θ. La dynamique dans les nouvelles variables peut être intégrée, et donc

$$A(q(t),p(t)) = A\circ\psi\left(a(0),\theta(0)+t\omega(a(0))\right).$$

Comme ψ est périodique de sa seconde variable, l'argument utilisé à l'étape 1 peut être réutilisé, ce qui montre (1.25).

En pratique, on ne calcule qu'une approximation de la moyenne temporelle qui apparaît au membre de droite de (1.25), en choisissant un pas de temps $\Delta t > 0$ d'intégration numérique. Sous des hypothèses mathématiques fortes (dont la complète intégrabilité du système Hamiltonien, l'analyticité des fonctions H et A, et la symplecticité du schéma numérique), on peut montrer [P2] l'estimation d'erreur suivante :

$$\left|\langle A\rangle_{num}^{Rie}(T) - \langle A\rangle_{I_1^0,\ldots,I_{3N}^0}\right| \leq C(r_0,H,A)\left(\frac{1}{T}+\Delta t^{r_0}\right), \tag{1.27}$$

[10] L'hypothèse d'analyticité sur A permet de permuter l'intégrale en temps et la somme sur $k\in\mathbb{Z}$.

[11] En dimension $d>1$, le dénominateur $ik\omega$ devient $ik\cdot\omega$, où $k\in\mathbb{Z}^d\setminus\{0\}$ et $\omega\in\mathbb{R}^d$. L'hypothèse de non-résonance permet de minorer $|k\cdot\omega|$.

où r_0 est l'ordre du schéma numérique, $\langle A \rangle_{num}^{Rie}(T)$ est l'expression algébrique (1.22), et $C(r_0, H, A)$ est une constante indépendante de T et de Δt (mais qui dépend de r_0 et des propriétés de A et de H).

La preuve de ce résultat est basée sur les arguments suivants :

1. le système Hamiltonien (1.4) possède $3N$ tores invariants, qui sont définis par $I_j(q,p) = \text{Constante}$.
2. puisque le schéma numérique est symplectique, le flot numérique est quasiment égal au flot exact d'un Hamiltonien modifié $H_{\Delta t}$ qui est égal à H à Δt^{r_0} près (cf. la preuve du théorème 1.1.1 et les relations (1.15) et (1.16)).
3. la théorie KAM montre que ce Hamiltonien $H_{\Delta t}$ possède lui aussi $3N$ tores quasi-invariants, et que ces tores sont Δt^{r_0}-proches des tores invariants de H. Autrement dit, il existe $3N$ fonctions $I_{\Delta t,j}(q,p)$, $j = 1, \ldots, 3N$, telles que

$$I_{\Delta t,j}(q,p) = I_j(q,p) + O(\Delta t^{r_0}), \tag{1.28}$$

   et telles que $I_{\Delta t,j}(q,p)$ soient des quasi-invariants[12] pour le système Hamiltonien associé à $H_{\Delta t}$.
4. grace à la relation (1.28), la moyenne de A sur la variété $S(q_0,p_0)$ définie par (1.23) est égale, à un terme d'erreur en Δt^{r_0} près, à la moyenne de A sur la variété $S_{\Delta t}(q_0,p_0)$ définie par

$$S_{\Delta t}(q_0,p_0) = \left\{(x,y) \text{ t.q. } \forall j \in [1,3N],\ I_{\Delta t,j}(x,y) = I_{\Delta t,j}(q_0,p_0)\right\}.$$

5. Il reste maintenant à estimer l'erreur entre la moyenne de A sur la variété $S_{\Delta t}(q_0,p_0)$ et la moyenne temporelle fournie par (1.22). Cette fois-ci, les points (q_n,p_n) de la trajectoire numérique qui sont pris en compte dans la moyenne temporelle sont (quasiment) sur la variété $S_{\Delta t}(q_0,p_0)$ sur laquelle on calcule la moyenne d'ensemble. On peut montrer que cette erreur décroît comme $1/T$ (c'est essentiellement la même preuve que pour démontrer (1.25)).

1.1.3.2 Accélération de la convergence

Le rythme de convergence en fonction du temps de simulation T peut être amélioré. Revenons une nouvelle fois à l'oscillateur harmonique de Hamiltonien $H(q,p) = \dfrac{p^2}{2m} + \dfrac{1}{2}\omega^2 q^2$, et définissons la moyenne temporelle par l'expression

$$\langle A \rangle^{(2)}(T) = \left(\frac{2}{T}\right)^2 \int_0^{T/2} \int_0^{T/2} A(q(t_1+t_2), p(t_1+t_2))\, dt_1\, dt_2. \tag{1.29}$$

[12]Les fonctions $I_{\Delta t,j}(q,p)$ sont constantes sur la trajectoire du système Hamiltonien associé à $H_{\Delta t}$ à un terme exponentiellement petit en le pas de temps Δt près.

En utilisant le développement en série de Fourier (1.26), on obtient

$$\begin{aligned}\langle A\rangle^{(2)}(T) &= \left(\frac{2}{T}\right)^2 \int_0^{T/2}\int_0^{T/2} \sum_{k\in\mathbb{Z}} a_k e^{ik\phi} e^{ik\omega t_1}\, e^{ik\omega t_2}\, dt_1\, dt_2 \\ &= a_0 + \sum_{k\neq 0} a_k e^{ik\phi} \left(\frac{2}{T}\int_0^{T/2} e^{ik\omega t}\, dt\right)^2,\end{aligned}$$

et le second terme du membre de droite, qui est l'erreur, décroît comme $1/T^2$. La moyenne temporelle (1.29) converge donc vers la moyenne d'ensemble au rythme $1/T^2$. Il est bien sûr possible de généraliser la formule (1.29) pour obtenir un rythme de convergence en $1/T^k$, pour tout entier k.

La formule (1.29) peut s'interpréter de deux manières :

- le calcul d'une moyenne temporelle se fait à partir d'une trajectoire, il faut donc choisir une condition initiale, et l'expression (1.29) est une façon de moyenner les résultats par rapports aux conditions initiales, puisqu'on moyenne la quantité $\dfrac{2}{T}\displaystyle\int_{t_1}^{t_1+T/2} A(q(t_2),p(t_2))\, dt_2$ par rapport à t_1.
- la formule (1.29) peut se récrire comme

$$\int_0^T A(q(t),p(t))\ f(t)\, dt$$

  pour une certaine fonction f, ce qui montre qu'il s'agit d'une moyenne temporelle *filtrée* du signal $t \mapsto A(q(t),p(t))$.

Par des arguments similaires à ceux utilisés pour prouver l'estimation (1.27), on peut montrer le résultat suivant : pour tout $k \in \mathbb{N}^*$, il existe des poids w_n^{k,N_m} tels que la moyenne filtrée

$$\langle A\rangle_{num}^{(k)}(T) = \sum_{n=0}^{N_m-1} w_n^{k,N_m}\, A(q_n,p_n), \tag{1.30}$$

où (q_n,p_n) est encore le flot numérique fourni par un algorithme symplectique de pas $\Delta t = T/N_m$ appliqué aux équations (1.3), vérifie l'estimation

$$\left|\langle A\rangle_{num}^{(k)}(T) - \langle A\rangle_{I_1^0,\ldots,I_{3N}^0}\right| \leq C(r_0,k,H,A)\left(\frac{1}{T^k} + \Delta t^{r_0}\right). \tag{1.31}$$

La formule avec poids (1.30) est une généralisation de la formule (1.22) : pour $k=1$, on a $w_n^{1,N_m} = 1/N_m$ pour tout n. Nous avons vérifié numériquement que l'estimation (1.31) est optimale.

Il est enfin possible [P1,P2] d'implémenter la formule avec poids de telle façon que le coût calcul de la moyenne soit négligeable devant le coût calcul de la trajectoire (comme c'est le cas avec l'expression (1.22)).

1.1.3.3 L'hypothèse de complète intégrabilité

Pour démontrer les estimations (1.27) et (1.31), nous avons supposé que le système Hamiltonien est complètement intégrable, ce qui nous a permis d'utiliser la théorie KAM. Remarquons que des résultats similaires peuvent être obtenus dans le cas de systèmes presque-intégrables[13], pour lesquels on peut encore utiliser la théorie KAM.

Il est bien connu que la plupart des systèmes considérés dans les applications ne sont pas intégrables. Il est en effet très rare qu'un système de N particules possède $3N$ invariants. C'est pourquoi, après avoir testé la formule (1.30) sur des systèmes intégrables, nous l'avons testé sur des systèmes non intégrables [P1], dans certaines conditions physiques (système de particules en phase solide interagissant via le potentiel de Lennard-Jones, ...). Nous constatons sur ces exemples que l'estimation (1.31) est encore vérifiée.

Donc, bien que nous ne sachions pas en faire la preuve, l'efficacité du schéma (1.30) semble dépasser le cadre des systèmes complètement intégrables.

1.1.4 Le cas de systèmes explorant plusieurs bassins d'énergie potentielle

Nous commençons par préciser ce qu'est un bassin d'énergie potentielle. Chaque minimum local de l'énergie potentielle correspond à un état métastable. Pour que le système passe d'un état métastable à un autre, il faut qu'il franchisse une barrière de potentiel (cf. la figure 1.6). A chaque minimum local $q^i \in \mathbb{R}^{3N}$ de l'énergie potentielle, on peut associer un domaine de $\mathbb{R}^{3N}$ qui est l'ensemble des positions $q^0 \in \mathbb{R}^{3N}$ telles que la fonction $q(t)$, définie par la dynamique

$$\frac{d\,q(t)}{d\,t} = F(q(t)), \quad q(0) = q^0, \tag{1.32}$$

converge vers q^i quand $t \to \infty$. Les équations (1.32) correspondent à la dynamique de plus grande pente le long de la surface de potentiel (cf. la figure 1.6).

Le comportement en temps long des moyennes temporelles dépend beaucoup du nombre de bassins d'énergie potentielle explorés, et surtout du temps passé dans chaque bassin par rapport au temps nécessaire pour changer de bassin. L'étude d'une particule dans un double puits de potentiel est à ce titre très instructive (cf. [P1]).

On considère une particule en deux dimensions, soumise au potentiel

$$V(q_x, q_y) = (q_x^2 - 1)^2 + (q_y + q_x^2 - 1)^2,$$

représenté sur la figure 1.7. Ce potentiel possède trois points critiques, deux minima globaux en $(\pm 1, 0)$ (pour lesquels $V(\pm 1, 0) = 0$), et un point selle en $(0, 1)$, pour lequel $V(0, 1) = 1$. L'énergie potentielle comporte donc deux bassins.

[13] Un système Hamiltonien presque intégrable est, de façon un peu imprécise, un système obtenu en perturbant un système Hamiltonien intégrable. La trajectoire reste alors au voisinage de tores

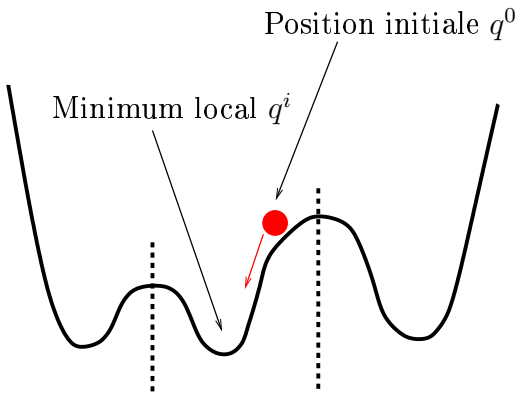

FIG. 1.6 – Exemple de potentiel avec trois minima locaux (coupe 1D). Les minima locaux sont séparés par des barrières de potentiel. Les équations (1.32) correspondent à une dynamique de plus grande pente, la boule représente une particule obéissant à cette dynamique. Les pointillés séparent les différents bassins.

Le Hamiltonien du système est

$$H = \frac{p_x^2}{2} + \frac{p_y^2}{2} + V(q_x, q_y).$$

Trois régimes peuvent être identifiés :

1. le cas d'une énergie strictement inférieure à l'énergie du point selle ($H_0 < 1$) ;
2. le cas d'une énergie bien plus grande que celle du point selle ($H_0 \geq 5$ dans le cas présent) ;
3. le cas d'une énergie légèrement supérieure à celle du point selle.

On travaille à énergie constante (dans l'ensemble thermodynamique NVE). Dans le premier cas ($H_0 < 1$), la particule ne peut pas franchir la barrière de potentiel et n'explore qu'un seul bassin. On observe que les moyennes temporelles (1.30) convergent au rythme $1/T^k$. Dans le second cas ($H_0 \gg 1$), la convergence est du même type et s'explique en considérant que la particule a suffisamment d'énergie pour franchir la barrière sans vraiment la "sentir".

Lorsque l'énergie de la particule est légèrement supérieure à la barrière (nous avons fait des simulations avec $H_0 = 1.25$), alors on constate que les moyennes temporelles convergent très lentement, au rythme $1/\sqrt{T}$. L'observation de la trajectoire montre que la particule reste très longtemps dans un bassin, puis, en un temps très court, va dans l'autre bassin, dans lequel elle réside encore très longtemps, ... La trajectoire est donc la succession de périodes de résidence dans chaque bassin, et le temps de résidence est très supérieur au temps de transit. Il est possible sur ce modèle simple de construire une chaîne de Markov à deux états, qui sont les deux bassins, et de paramétrer cette chaîne (temps de résidence dans chaque état, matrice de transfert entre états) à partir des résultats obtenus par la simulation de

invariants pendant un temps exponentiellement long en la pertubation.

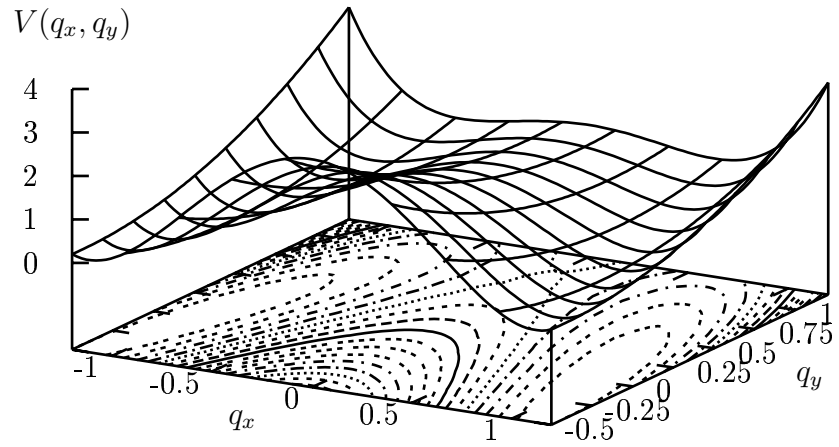

FIG. 1.7 – Double puits de potentiel en 2D. On a tracé dans le plan (q_x, q_y) les isovaleurs de l'énergie potentielle.

dynamique moléculaire. Une moyenne temporelle calculée à partir d'une trajectoire de cette chaîne de Markov converge au rythme universel $1/\sqrt{T}$. L'accord entre ces deux rythmes (celui obtenu par dynamique moléculaire et celui de la chaîne de Markov) montre que le modèle probabiliste donne une bonne compréhension de ce qui se passe (sur cet exemple du double puits, cf. aussi [113]).

Ces conclusions se généralisent à des exemples physiques plus complexes. Nous avons ainsi simulé des alcanes (cf. la figure 1.5 ; les atomes d'hydrogène sont agglomérés aux atomes de carbone, qui seuls sont simulés), pour lesquelles l'énergie potentielle s'écrit

$$V(q) = \sum_{i=1}^{N-1} V_2(q_i, q_{i+1}) + \sum_{i=1}^{N-2} V_3(q_i, q_{i+1}, q_{i+2}) + \sum_{i=1}^{N-3} V_4(q_i, q_{i+1}, q_{i+2}, q_{i+3}).$$

L'énergie potentielle fait donc apparaître trois types de terme, qui correspondent, de la raideur la plus importante à la raideur la plus faible, à

- des interactions à deux corps entre atomes de carbone voisins : ces interactions sont quadratiques, très raides, et imposent à la distance atomique entre deux atomes consécutifs de rester proche d'une constante ;
- des interactions à trois corps, qui dépendent de l'angle entre deux liaisons atomiques successives (cf. la figure 1.2), et qui sont quadratiques en cet angle.
- des interactions à quatre corps (qui dépendent des angles dihédraux, cf. la figure 1.2), qui font intervenir un potentiel présentant plusieurs bassins (cf. la

figure 1.8). Dans les unités avec lesquelles on travaille, les barrières pour la fonction V_4 sont de l'ordre de 10^{-4}.

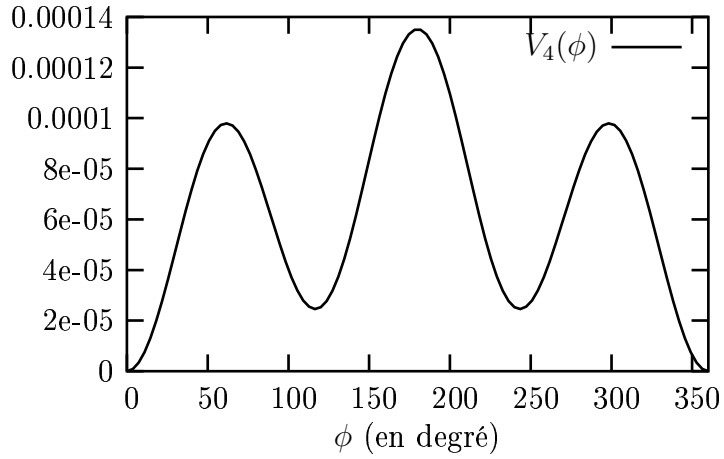

FIG. 1.8 – Potentiel de torsion V_4 en fonction de l'angle dihédral ϕ.

Si l'énergie est assez faible, alors le système ne peut franchir aucune barrière, et les moyennes temporelles (1.22) convergent au rythme $1/T$ (qu'on peut accélérer en $1/T^k$ avec la formule (1.30)). Au dessus d'un certain niveau d'énergie, les moyennes temporelles ne convergent qu'au rythme $1/\sqrt{T}$. On constate que la situation ne change pas même si on travaille avec une énergie de plus en plus grande. Autrement dit, il ne semble pas possible, sur cet exemple, d'identifier un régime haute énergie pour lequel on retrouve une convergence en $1/T$ (à la différence du cas du double puits en 2D). La manière dont se répartit l'énergie totale du système dans les différents degrés de liberté (lents, intermédiaires, rapides, qui correspondent respectivement aux interactions à 4 corps, 3 corps, 2 corps) fournit une explication. On note

$$V_{4,\mathrm{tot}}(q) = \sum_{i=1}^{N-3} V_4(q_i, q_{i+1}, q_{i+2}, q_{i+3})$$

le terme d'énergie potentielle qui correspond aux degrés de liberté les plus lents, et pour lequel plusieurs bassins sont présents. On constate que, au cours de la trajectoire, l'énergie potentielle $V_{4,\mathrm{tot}}(q(t))$ oscille entre deux valeurs, $V_{4,\mathrm{tot}}^{\min}(H_0)$ et $V_{4,\mathrm{tot}}^{\max}(H_0)$, qui dépendent de l'énergie totale H_0 du système (ce sont des fonctions croissantes de H_0).

Ces bornes $V_{4,\mathrm{tot}}^{\min}(H_0)$ et $V_{4,\mathrm{tot}}^{\max}(H_0)$ ont une limite finie quand H_0 tend vers l'infini, et elle est atteinte pour des énergies qui correspondent à une température de 10000 K. Ainsi, pour les températures[14] $T = 10000$ K et $T = 67000$ K, on trouve que

[14] On ne prétend pas ici que le modèle physique rend bien compte de l'expérience à de si hautes températures. L'idée est simplement de comprendre le comportement du système pour un modèle donné.

l'énergie moyenne par angle dihédral, soit $\frac{1}{N}V_4(q(t))$, est (respectivement) minorée et majorée par

$$\frac{1}{N}\,V_{4,\mathrm{tot}}^{\min} = 5.10^{-5}, \quad \frac{1}{N}\,V_{4,\mathrm{tot}}^{\max} = 6,25.10^{-5}.$$

Ces valeurs sont inférieures à la barrière de potentiel, qui est de 10^{-4} pour un angle dihédral. Donc il semble que l'énergie dans les degrés de liberté "angles dihédraux" soit trop faible, par rapport aux barrières à franchir, pour permettre un échantillonnage rapide de l'espace des phases. Bien sûr, l'énergie $V_{4,\mathrm{tot}}$ n'est pas répartie de manière équitable entre les différents angles dihédraux (la répartition change au cours du temps), ce qui permet à chaque angle dihédral, séparément, de franchir les barrières de potentiel. Cependant, les mouvements collectifs (plusieurs angles changent de puits en même temps) ne sont pas tous permis (le nombre d'angles qui peuvent changer de puits en même temps est limité).

En résumé, on peut dresser le tableau 1.1, qui donne le rythme de convergence des moyennes temporelles en fonction du régime énergétique dans lequel on travaille.

Régime	Convergence de $\frac{1}{N_m}\sum_{n=0}^{N_m-1} A(q_n,p_n)$	Convergence de $\sum_{n=0}^{N_m-1} w_n^{k,N_m} A(q_n,p_n)$
$E < V_b$	$1/T$	$1/T^k$
$E \gg V_b$	$1/T$	$1/T^k$
$E > V_b$	$1/\sqrt{T}$	$1/\sqrt{T}$

TAB. 1.1 – Rythme de convergence des moyennes temporelles en fonction du régime énergétique. L'énergie E est l'énergie du système qui se trouve dans les degrés de liberté pour lesquels l'énergie potentielle présente des barrières, et V_b est la hauteur des barrières correspondantes.

Pour certains systèmes (comme le cas des alcanes discutés ci-dessus), le régime $E \gg V_b$ n'existe pas. Par ailleurs, l'expérience numérique semble montrer que le système n'est dans le régime $E < V_b$ que pour des températures très basses. Autrement dit, il semble que, en dehors de conditions physiques particulières, **le régime générique** est le troisième, pour lequel **la convergence des moyennes temporelles n'a lieu qu'au rythme** $1/\sqrt{T}$.

1.1.5 Extension à d'autres ensembles thermodynamiques : le cas de l'ensemble NVT

Jusqu'à présent, seul le cas de l'ensemble thermodynamique NVE a été considéré, car c'est le cas le plus simple. On s'intéresse ici au problème de calcul de moyennes

thermodynamiques dans l'ensemble NVT. On rappelle qu'une telle moyenne, notée $\langle A \rangle_{NVT}$, est définie par (1.5), où l'espace des phases est $\Omega = \Omega_{NVT}$ défini par (1.8) et où la mesure est la mesure de Boltzmann $d\mu_{NVT}$ définie par (1.9).

L'approche dynamique moléculaire consiste à trouver une dynamique sur les variables (q, p), qu'on écrit ici sous la forme

$$\frac{d\,(q(t), p(t))}{d\,t} = \Gamma(q(t), p(t)), \tag{1.33}$$

telle que la seule mesure invariante de cette dynamique soit la mesure $d\mu_{NVT}$, ce qui laisse ensuite espérer un théorème ergodique, du type

$$\langle A \rangle_{NVT} = \lim_{T \to \infty} \frac{1}{T} \int_0^T A(q(t), p(t))\, dt,$$

où $(q(t), p(t))$ est une solution de (1.33).

En pratique, ce programme est loin d'être rempli. De nombreuses dynamiques ont été proposées, et il est facile de vérifier pour chacune d'elle que la mesure de Boltzmann est bien une mesure invariante. Par contre, pour certaines de ces dynamiques, la question de savoir si la mesure de Boltzmann est la seule mesure invariante est une question ouverte, et l'expérience numérique laisse entrevoir de sérieux problèmes.

Deux approches sont utilisées : des approches déterministes, fondées sur des systèmes étendus, et des approches probabilistes.

1.1.5.1 L'approche "systèmes étendus"

Cette approche s'appuie sur la signification physique de ce qu'est un système à température constante : c'est un système qui échange de l'énergie avec un thermostat extérieur, de façon à garder sa température constante. L'idée de cette approche consiste donc à rajouter aux variables physiques (q, p) des variables qui décrivent l'évolution de ce thermostat, et à *postuler* une forme d'interaction entre les variables physiques et les variables du thermostat. Nous insistons sur le fait qu'il ne faut pas chercher de signification physique précise aux équations dynamiques ainsi obtenues : la seule propriété qu'on demande est que la trajectoire solution de ces équations parcourt l'espace des phases avec la mesure souhaitée, soit la mesure de Boltzmann. Donnons un exemple très simple d'une telle dynamique dans un système étendu. On considère une particule unidimensionnelle, le thermostat est représenté par deux variables scalaires ξ et p_ξ, et le système dynamique s'écrit :

$$\frac{d\,q}{d\,t} = \frac{p}{m}, \tag{1.34}$$

$$\frac{d\,p}{d\,t} = F(q) - \frac{p_\xi}{Q}\,p, \tag{1.35}$$

$$\frac{d\,p_\xi}{d\,t} = \frac{p^2}{m} - k_B T, \tag{1.36}$$

$$\frac{d\,\xi}{d\,t} = \frac{p_\xi}{Q}. \tag{1.37}$$

Dans ces équations, T est la température à laquelle on souhaite simuler le système (c'est une constante), et Q est un paramètre, qui joue le rôle de la "masse" du thermostat. Les équations précédentes sont connues sous le nom d'équations de Nosé-Hoover [62,69]. Il est possible de les généraliser et d'utiliser plus de thermostats (on parle alors des chaînes de Nosé-Hoover [67]), ce qui revient à coupler le premier thermostat avec un second, ... Ni les équations de Nosé-Hoover ni les chaînes de Nosé-Hoover ne sont des équations Hamiltoniennes[15], et il existe des cas bien répertoriés [57, 67] dans la littérature pour lesquels on constate numériquement que la solution de (1.34-1.37) n'explore pas l'espace des phases suivant la mesure de Boltzmann (dans le cas $F = 0$, dans le cas d'un potentiel harmonique, ...). Cette approche est néanmoins utilisée, les cas problématiques étant bien connus, car le sentiment général est que ces problèmes sont liés à un trop faible nombre de particules, ou à un trop faible nombre de thermostats, et que augmenter l'un ou l'autre résout le problème. A notre connaissance, il n'y a aucune preuve que ce soit effectivement le cas, ni aucune preuve qui contredise ce sentiment.

Les équations (1.34-1.37), comme beaucoup de leurs généralisations, conservent une mesure. Il est naturel de chercher à construire des algorithmes d'intégration qui conservent eux aussi exactement cette mesure. De plus, on sait que la préservation de propriétés géométriques par l'algorithme lui confère en général de bonnes propriétés (par exemple, l'erreur numérique en fonction du temps total de simulation croît moins vite). Ainsi, le flot d'un système Hamiltonien est symplectique (ce qui implique que la mesure de Lebesgue $dq\,dp$ est préservée), et nous avons vu que les algorithmes symplectiques sont particulièrement intéressants (excellente conservation de l'énergie sur des temps longs, ...). Pour des EDO (non nécessairement Hamiltoniennes) du type

$$\frac{d\,x(t)}{d\,t} = \Gamma(x) \tag{1.38}$$

avec div $\Gamma = 0$, la mesure de Lebesgue $dq\,dp$ est conservée par le flot, et des expériences numériques montrent que des algorithmes qui conservent exactement cette mesure ont des propriétés meilleures que les autres [52] (par exemple, dans certaines applications, l'erreur entre la trajectoire exacte et la trajectoire numérique croît comme $\Delta t^{r_0}\ T$, où Δt est le pas de temps, r_0 l'ordre du schéma et T la longueur de la trajectoire, alors qu'elle croît comme $\Delta t^{r_0}\ T^2$ pour des algorithmes qui ne conservent pas la mesure de Lebesgue). Il est donc très souvent observé, et dans certains cas prouvé, que les algorithmes qui préservent certaines propriétés géométriques du système ont des comportements quantitatifs intéressants.

J'ai abordé avec Régis Monneau cette question de construire des algorithmes qui conservent une mesure en travaillant sur une généralisation récente des équations de Nosé-Hoover, les équations GGMT [65]. Ce système d'EDO conserve une me-

[15] A ce sujet, nous mentionnons l'existence de la méthode Nosé-Poincaré [61], qui est elle-aussi fondée sur la notion de système étendu, et qui est Hamiltonienne. A notre connaissance, c'est la seule qui soit dans ce cas.

sure notée $d\mu = h(x)\,dx$, où x représente l'ensemble des variables, soit les variables physiques (q,p) complétées par des variables non physiques décrivant le thermostat. Après intégration sur les variables non physiques, la mesure $h(x)\,dx$ donne la mesure de Boltzmann :

$$\forall f \in \mathcal{C}^0, \quad \int f(q,p)h(x)\,dx = \int f(q,p)\exp\left(-\frac{H(q,p)}{k_BT}\right)\,dq\,dp.$$

Nous avons montré que les algorithmes proposés dans la littérature ne conservent pas exactement $d\mu$ et nous avons proposé une méthode pour construire un schéma qui conserve la mesure [P3]. L'idée est de passer, par un changement de variables $x \mapsto y(x)$, du système (1.38) à un système d'EDO de la forme $\dfrac{d\,y(t)}{d\,t} = \gamma(y)$, avec $\partial_{y_i}\gamma_i = 0$ pour tout i et telle que la mesure $d\mu = h(x)\,dx$ s'écrive $d\mu = dy$. Sur un tel système, on sait facilement construire un algorithme qui préserve la mesure dy, il suffit ensuite de revenir dans les variables initiales. La construction d'un changement de variables explicite répondant aux critères ci-dessus n'est pas toujours possible. Pour le cas particulier qui nous intéresse, à savoir les équations GGMT, nous avons exhibé un tel changement de variables, ce qui nous a permis de construire un nouvel algorithme.

1.1.5.2 Les approches probabilistes

Nous avons souligné ci-dessus les difficultés liées à l'approche "système étendu", citons maintenant d'autres approches, fondées sur des méthodes probabilistes. Il est possible de travailler avec l'équation de Langevin, bien connue dans d'autres domaines, et dont la justification mathématique est plus claire.

Une autre méthode est celle dite "Hybrid Monte Carlo" (cf. [49,115]), qui consiste à générer une suite de positions $(q_n)_{n\in\mathbb{N}}$ qui échantillonnent l'espace $\mathbb{R}^{3N}$ suivant la densité $\exp\left(-V(q)/(k_BT)\right)$. On se donne un temps τ. Etant donné $q_n \in \mathbb{R}^{3N}$, la position q_{n+1} est calculée ainsi :

- On tire au hasard une impulsion p_n suivant la loi $e^{-\beta p^2/2m}$, où $\beta = 1/(k_BT)$;
- On calcule la solution de (1.4) (trajectoire à énergie constante) sur l'intervalle de temps $[0,\tau]$ en partant de la condition initiale (q_n,p_n), on obtient $(\tilde{q}_{n+1},\tilde{p}_{n+1})$.
- L'algorithme Metropolis est utilisé pour accepter ou rejeter $\tilde{q}_{n+1}$. On note $H_n = H(q_n,p_n)$ et $\tilde{H}_{n+1} = H(\tilde{q}_{n+1},\tilde{p}_{n+1})$. Si $\tilde{H}_{n+1} \leq H_n$, on accepte la nouvelle position : $q_{n+1} = \tilde{q}_{n+1}$. Sinon, on accepte la nouvelle position avec la probabilité $e^{-\beta(\tilde{H}_{n+1}-H_n)}$. Si la nouvelle position n'est pas acceptée, on pose $q_{n+1} = q_n$.

Cette méthode utilise donc à la fois des tirages aléatoires et le calcul de trajectoires à énergie constante.

1.1.6 Au delà de la dynamique moléculaire

Comme cela a été précisé dans la section 1.1.4, à partir du moment où le système explore plusieurs bassins d'énergie potentielle, et qu'il reste longtemps dans chaque bassin (par rapport au temps mis pour aller d'un bassin à un autre), les moyennes temporelles convergent au rythme $1/\sqrt{T}$, ce qui pose des problèmes pour la mise en œuvre numérique. Par ailleurs, il n'est pas certain que, sur le temps fini de simulation, tous les puits soient explorés. Par conséquent, la moyenne temporelle converge lentement, et de plus, à un instant T, sa valeur peut être assez éloignée de sa limite en temps infini, si certains puits n'ont pas encore été visités. Enfin, certains phénomènes intéressants pour les sciences des matériaux (la diffusion d'impuretés dans le réseau cristallin) ou en biologie (le repliement d'une protéine) ont des temps caractéristiques de l'ordre de la microseconde (voire plus), alors que le pas de temps utilisé en dynamique moléculaire est de l'ordre de la femtoseconde (il peut être plus ou moins grand suivant les différentes techniques utilisées pour traiter les échelles les plus rapides, cf. la section 1.1.2.3, mais la différence d'ordre de grandeur entre le pas de temps et la longueur souhaitée de la trajectoire demeure). Plusieurs approches, dépassant le simple cadre de la dynamique moléculaire, ont été proposées pour résoudre ce problème (on pourra trouver dans [27] une revue sur ce sujet, avec une bibliographie très complète).

L'idée essentielle est que la dynamique du système est composée de deux phases distinctes, la résidence dans un puits d'énergie potentielle d'une part et le transit d'un puits à l'autre d'autre part. Si on connaît

1. la localisation des puits de potentiel,
2. les chemins qui mènent le système d'un puits à un autre,
3. et les probabilités de passage, au bout d'un temps donné, d'un puits à un autre,

on peut alors représenter le système avec une précision acceptable par une chaîne de Markov dont les états sont justement les états métastables, correspondant aux puits d'énergie potentielle. Les méthodes que nous citons maintenant ont pour but de calculer les informations mentionnées ci-dessus, et la difficulté vient de la grande dimension ($3N$) dans laquelle on travaille.

Pour trouver où sont les puits d'énergie potentielle, l'idée est d'aider le système à franchir les barrières. Supposons que la barrière à franchir soit de hauteur V_b : alors, suivant la loi d'Arrhenius, la fréquence de sortie du puits est proportionnelle à

$$\exp\left(-\frac{V_b}{k_B T}\right), \tag{1.39}$$

où T est la température à laquelle on travaille. Plusieurs méthodes ont été proposées. La première, connue sous le nom de "Hyperdynamics method" [118] (cf. aussi [114] pour une application) consiste à modifier l'énergie potentielle du système, de façon à relever les fonds des puits (cf. la figure 1.9), sans modifier le potentiel au voisinage

de ses points-selle (par conséquent, si deux bassins A et C sont accessibles à partir du bassin B dans lequel le système réside initialement, le rapport des probabilités de transition $B \to A$ et $B \to C$ est conservé). Les barrières sont donc moins hautes, et plus facilement franchies (au bout d'un temps plus court).

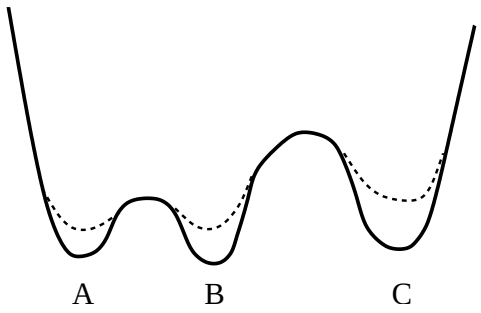

FIG. 1.9 – Méthode "Hyperdynamics" : les fonds de puits de l'énergie potentielle sont relevés (traits en pointillé), afin d'abaisser des barrières de potentiel.

La seconde méthode, connue sous la nom de "Temperature Accelerated Dynamics" [112, 117], consiste à travailler à une température plus élevée (ce qui encore une fois va accélérer le franchissement des barrières), et à extrapoler les résultats obtenus. Une autre méthode consiste à simuler plusieurs systèmes en parallèle [119].

Lorsque les différents bassins d'énergie potentielle sont connus, il faut ensuite déterminer le chemin de transition, ce qui va permettre de calculer la hauteur du col (i.e. la barrière) d'énergie potentielle à franchir (par définition, un col est un point q tel que les valeurs propres de la matrice hessienne du potentiel en ce point sont toutes positives, sauf une, qui est strictement négative). Cette hauteur de col intervient en effet pour le calcul de la probabilité de transition (cf. (1.39)). Là encore, plusieurs méthodes ont été proposées (la "String method" [109], la "Nudged elastic band method" [110]), qui reposent essentiellement sur la même idée : on se donne un chemin d'un puits A à un puits B, et on fait évoluer ce chemin suivant les lignes de plus grande pente de la surface de potentiel, de façon à ce qu'il passe par le col.

Toutes les méthodes précédentes procèdent puits de potentiel par puits de potentiel (pour chaque puits, détermination des puits voisins et des chemins de transition, ...). Une approche plus globale (et analysée sur le plan numérique) a été proposée dans [107, 108, 115, 116].

Toutes les méthodes que nous avons évoquées sont très récentes, et elles sont actuellement en train d'être testées par la communauté physicienne et chimiste sur des cas de plus en plus complexes.

1.2 Méthodes multi-échelles pour la simulation des matériaux

Nous abordons dans cette section un sujet tout à fait différent, celui des méthodes "multi-échelles" pour la simulation des matériaux, qui ont fait leur apparition depuis une dizaine d'années dans plusieurs domaines reliés au calcul scientifique. Néanmoins, le mot "multi-échelle" recouvre des réalités très variées, à la fois sur le plan des échelles en jeu (qui peuvent être d'espace et / ou de temps), des motivations pour de telles approches, des modèles considérés et des techniques mathématiques et numériques utilisées. Plusieurs de ces différentes méthodes sont discutées dans [133].

De façon générale, l'apparition de ces méthodes est motivée par la volonté (et la nécessité) de mieux comprendre l'impact, à l'échelle macroscopique, de phénomènes dont la description relève d'une échelle microscopique. Par ailleurs, le développement de la puissance de calcul des ordinateurs permet aujourd'hui de mettre en œuvre de telles méthodes, qui conduisent pour la plupart à des calculs lourds.

Citons ici un seul exemple de problème multi-échelle, celui de la dégradation des propriétés mécaniques des aciers de cuve des centrales nucléaires, qui a déjà été évoqué. Le phénomène à l'origine de cette dégradation est l'irradiation du métal et la désorganisation locale du réseau atomique : l'impact d'un neutron modifie le réseau sur une zone dont le diamètre est de l'ordre de 20.10^{-9} m, et cette modification prend un temps de l'ordre de 10^{-12} s. Les échelles auxquelles on observe cette dégradation sont des échelles macroscopiques (en espace, le mètre, et en temps, l'année). La compréhension, à partir de considérations microscopiques, d'un tel phénomène macroscopique est un problème excessivement difficile, et sur le plan pratique très important, car c'est un des phénomènes qui contrôlent la durée de vie des centrales.

Au cours de cette thèse, nous nous sommes intéressés à deux problématiques bien précises, qui relèvent d'approches multi-échelles *en espace*. La première concerne le couplage de modèles atomistiques avec des modèles de continuum (cf. la section 1.2.1) : l'échelle la plus fine est ici l'échelle atomistique, de l'ordre de 10^{-10} m, et il s'agit de coupler deux modèles physiques différents. L'autre thème étudié concerne l'homogénéisation numérique de matériaux polycristallins (cf. la section 1.2.2). L'échelle la plus fine est celle du grain (cf. ci-dessous), qui est de l'ordre de 10^{-6} m, et le même modèle est utilisé à toutes les échelles, celui de la mécanique du continuum.

1.2.1 Couplage de modèles atomistiques avec des modèles de continuum

L'approche la plus courante en mécanique des matériaux consiste à considérer que la matière est un continuum, dont l'état, localement, peut être décrit par plusieurs champs (déformations, contraintes, éventuellement déformations plastiques, endommagement, ...). Dans ce modèle, la nature atomistique de la matière est

ignorée.

Cette approche devient discutable lorsqu'on cherche à prendre en compte des phénomènes localisés dans le matériau, et dont les dimensions caractéristiques sont proches des dimensions atomiques.

La description de la propagation de fractures dans des matériaux cristallins [208] est un exemple de tel phénomène : loin de la fracture, le réseau atomique constituant le matériau est parfait. Au niveau de la pointe de la fracture, des liaisons atomiques se cassent, et sur les lèvres de la fracture, le réseau atomique se réorganise. La question qui se pose pour certains matériaux n'est pas d'éviter la formation de fractures (elles existent de toute façon), mais de savoir quelles sont les contraintes maximales que peut supporter le matériau au delà desquelles la fracture se propage (lorsque les contraintes sont inférieures à ce seuil, la fracture ne se propage pas et la situation est considérée comme acceptable).

La description des joints de grain dans les matériaux polycristallins nécessite aussi de prendre en compte des phénomènes très localisés. Un métal, par exemple, même s'il n'est pas fissuré, est rarement constitué par un réseau atomique parfait. Fréquemment, il est constitué d'un assemblage de grains, dans lesquels le réseau atomique est parfait. Le diamètre d'un grain est de l'ordre du micromètre, et l'orientation du réseau atomique change d'un grain à l'autre (cf. la figure 1.10). Si on considère un matériau constitué d'un petit nombre de grains, on peut alors isoler certaines zones (les interfaces entre les grains, aussi appelées joints de grain) où le réseau fait apparaître des singularités, alors que dans le reste du matériau, le réseau est régulier. Le problème consiste à comprendre comment le réseau atomique se réorganise au niveau de ces interfaces, ce qui en détermine la fragilité.

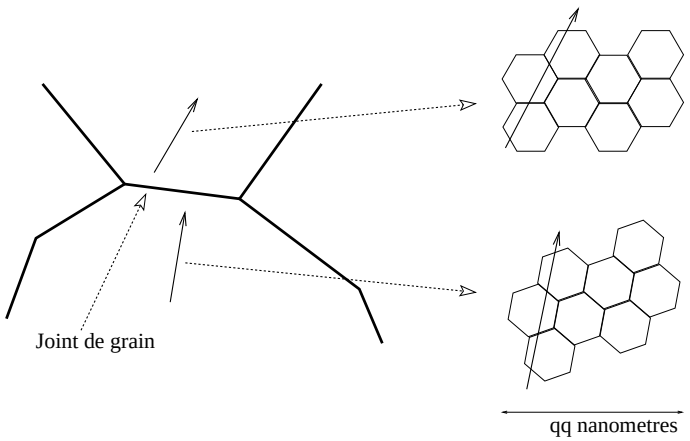

FIG. 1.10 – Un exemple de matériau polycristallin : la matière est formée de grains, dont le diamètre est de l'ordre du micromètre. L'interface entre deux grains est ici représentée. La flèche sur le dessin de gauche symbolise l'orientation du réseau atomique, qui est schématiquement reproduit à droite, dans deux orientations différentes (on a choisi une maille atomique hexagonale). A l'interface entre deux grains, le réseau se réorganise.

Supposons que l'on souhaite étudier des matériaux dans lesquels des singularités comme celles décrites ci-dessus apparaissent. Leur description précise nécessite l'utilisation d'un modèle à une échelle atomistique. D'autre part, pour éviter des effets de bord, il est nécessaire de considérer des matériaux suffisamment grands. Dans le cas de la fracture, par exemple, des ondes élastiques sont émises lorsque la pointe de fracture avance (c'est un des modes de relaxation d'énergie). Ces ondes ont une influence sur le champ de contrainte dans le matériau, qui en retour gouverne l'évolution de la fracture. Une description précise de la fracture doit donc prendre en compte la propagation de ces ondes, dont le calcul ne doit pas être altéré par la présence de bords.

Pour étudier les problèmes mentionnés ci-dessus, il faut donc à la fois simuler un grand domaine et décrire précisément des phénomènes très localisés. Plusieurs approches sont alors envisageables.

La première approche consiste à décrire la singularité avec un modèle phénoménologique à l'échelle du continuum. Dans le cas de la fracture, on modélise cette dernière comme une ligne de discontinuité du champ de déplacement [209], et on se donne aussi un modèle pour décrire l'avancée de la fracture : à partir de quelle contrainte, de quel gradient de déformation celle-ci se propage-t-elle ? dans quelle direction, et éventuellement à quelle vitesse ? Différents modèles sont discutés dans [207]. Cette approche est économe en temps calcul, puisque seule l'échelle du continuum est traitée. Elle repose néanmoins sur une loi phénoménologique, dont les paramètres doivent être calibrés sur l'expérience. L'utilisation de matériaux de plus en plus divers, et dans des conditions mécaniques de fonctionnement de plus en plus variées, rend cette étape de calibration difficile et coûteuse. De plus, comme pour toute loi calibrée sur l'expérience, il faut rester prudent lorsqu'on travaille en dehors du régime dans lequel elle a été ajustée.

Une deuxième approche consiste à tirer parti du fait que les phénomènes qui ne peuvent être décrits qu'à une échelle atomistique sont très localisés dans le matériau. Le domaine où il n'y a pas de singularité est grand par rapport à l'échelle atomistique, et l'approche classique de la mécanique du continuum est donc valable dans ce domaine. L'idée est alors de décrire la singularité par un modèle fin approprié, tandis que le reste du matériau est décrit avec un modèle de mécanique du continuum. La difficulté dans ce cas-là consiste d'une part à travailler avec deux modèles *consistants* (nous revenons sur ce point ci-dessous), et d'autre part à trouver un *critère* gouvernant le choix des zones : où utilise-t-on chacun des modèles, sachant que la localisation de la singularité est une inconnue du problème ? quelle est la condition d'interface ?

Nous nous intéressons dans cette section à la seconde approche, qui consiste à coupler deux modèles. Les deux échelles que nous considérons ici sont celle du continuum (dont les concepts sont supposés être connus du lecteur ; on pourra consulter [125, 126, 130, 137] pour l'exposé des modèles, et [124, 148] pour des méthodes numériques adaptées) et celle de l'atome. De nombreuses échelles intermédiaires

peuvent être identifiées, nous nous intéressons ici à des problèmes pour lesquelles elles n'interviennent pas. Nous supposons de plus que les phénomènes localisés peuvent être bien décrits par un modèle atomistique classique tel que celui exposé dans la section 1.1, dans laquelle la dynamique moléculaire a été présentée : dans ce modèle, la matière est décrite comme un ensemble d'atomes, considérés comme des particules classiques ponctuelles qui interagissent via des potentiels d'interaction.

Pour coupler ces deux modèles, plusieurs méthodes ont été récemment proposées. L'idée de base est de faire une décomposition de domaine, et d'employer sur chacun des domaines un modèle physique différent. Puisque l'on souhaite considérer des phénomènes très localisés, la taille totale des matériaux qui sont simulés est bien sûr très éloignée des échelles macroscopiques : les dimensions caractéristiques sont de l'ordre de 100 nm pour des simulations 2D (ce qui représente un système de 10^5 à 10^6 atomes), et 30 nm pour les simulations 3D (soit un système de 10^6 à 10^7 atomes).

Une notion très importante, lorsqu'on souhaite coupler deux modèles, est de s'assurer de leur *consistance*. Dans le cas présent, cela signifie que si la déformation du réseau atomique est donnée et régulière (et que donc le modèle de mécanique du continuum est valable), l'énergie du système calculée par le modèle atomistique doit être la même que celle calculée par le modèle de mécanique (à une erreur contrôlée par la maille atomique près). De même, si la déformation n'est plus donnée, mais qu'on cherche à la calculer, par exemple en considérant un problème d'équilibre, alors les équations des deux modèles doivent conduire à la même solution.

Dans la suite de cette section, nous décrivons plus précisément les problèmes liés au couplage d'un modèle atomistique avec un modèle de continuum, puis nous présentons plusieurs méthodes proposées dans la littérature pour réaliser ce couplage. Commençons par le cas le plus simple.

1.2.1.1 Le cas statique à température nulle

On considère un matériau décrit à l'échelle atomique, et on note ε le paramètre de maille atomique. On suppose que les atomes n'interagissent que par un potentiel de paire, qu'on note V_ε (V_ε ne dépend que de la distance interatomique, pour des raisons d'invariance galiléenne). Pour simplifier, on suppose que, dans sa configuration initiale, le réseau atomique est le réseau $\varepsilon\mathbb{Z}^3 \cap \Omega$, où Ω est le domaine (à l'échelle macroscopique) occupé par le matériau.

On suppose qu'on se donne une déformation du matériau à l'échelle atomique, c'est-à-dire qu'on se donne une fonction $u : \Omega \to \mathbb{R}^3$ telle que les positions des atomes dans la configuration déformée soient $(u(i\varepsilon))_i$, avec $i\varepsilon \in \varepsilon\mathbb{Z}^3 \cap \Omega$ (cf. la figure 1.11).

L'énergie par atome d'un tel réseau déformé est donnée par

$$E_\mu(u) = \frac{1}{2N} \sum_{i \in \mathbb{Z}^3 \cap \Omega/\varepsilon} \sum_{j \neq i} V_\varepsilon \left(|u(j\varepsilon) - u(i\varepsilon)| \right), \tag{1.40}$$

où $N = \mathrm{Card}\left(\mathbb{Z}^3 \cap \Omega/\varepsilon\right)$ est le nombre d'atomes, et où l'indice μ symbolise le caractère microscopique (ici, atomistique) de l'énergie. Le potentiel V_ε dépend de ε : en

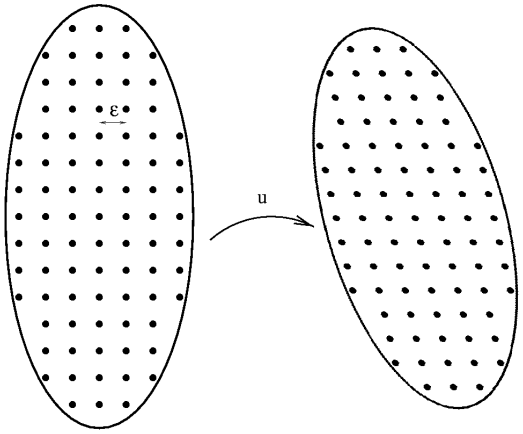

FIG. 1.11 – Configuration de référence et configuration déformée par u.

effet, puisque ε est la distance interatomique à l'équilibre, la fonction $r \in \mathbb{R}^+ \mapsto V_\varepsilon(r)$ est minimale en ε. On fait en général l'hypothèse que

$$V_\varepsilon(r) = V_0\left(\frac{r}{\varepsilon}\right),$$

où V_0 ne dépend pas de ε et atteint son minimum en 1. Si la déformation u est assez régulière, alors, quand $\varepsilon \to 0$, l'énergie atomistique (1.40) converge [142] vers

$$E_M(u) = \frac{1}{2|\Omega|} \int_\Omega W\left(\nabla u(x)\right)\, dx, \tag{1.41}$$

avec $W(M) = \sum_{k\in\mathbb{Z}^3\setminus\{0\}} V_0(|M\cdot k|)$. On retrouve donc une expression de l'énergie bien connue en mécanique du continuum, mais on dispose en plus d'un fondement atomistique pour la densité d'énergie élastique W.

Revenons maintenant au problème de la description d'une singularité. Une façon de comprendre le lien entre (1.40) et (1.41) consiste à partir de l'échelle du continuum, dans laquelle l'énergie est définie par (1.41). L'analyse à la limite précédente permet d'obtenir l'expression de W en fonction d'objets de nature atomistique (le potentiel d'interaction V_0), et de plus montre que définir l'énergie par (1.41) n'est valable que si la déformation considérée est régulière. Sinon, le passage de (1.40) à (1.41) n'est pas valable, et la seule expression dont on dispose pour calculer l'énergie est (1.40).

Etant donnée une déformation u qui n'est pas régulière dans l'ensemble du domaine Ω, une idée est alors de partitionner le domaine,

$$\Omega = \Omega_M(u) \cup \Omega_\mu(u),$$

où $\Omega_M(u)$ représente les zones de régularité de u et $\Omega_\mu(u)$ les zones de singularité[16]

[16]Les mots "zone de régularité" et "zone de singularité" sont volontairement laissés flous à ce stade, nous les préciserons plus loin.

de u. A partir de cette partition, on définit l'énergie par

$$\begin{aligned} E_c(u) &= \frac{1}{2|\Omega|}\int_{\Omega_M(u)} W\left(\nabla u(x)\right) dx \\ &+ \frac{1}{2N}\sum_{i\varepsilon\in\varepsilon\mathbb{Z}^3\cap\Omega_\mu(u)}\sum_{j\neq i} V_0\left(\frac{|u(j\varepsilon)-u(i\varepsilon)|}{\varepsilon}\right). \end{aligned} \tag{1.42}$$

Passer de (1.40) à (1.42) consiste donc à passer à la limite $\varepsilon \to 0$ uniquement là où cela est licite, c'est-à-dire, par définition, dans le domaine $\Omega_M(u)$. L'expression (1.42) est donc une approximation valable de (1.40), indépendamment de la régularité de u.

En pratique, la déformation u est l'inconnue du problème, et une possibilité est de la définir par un problème variationnel, c'est-à-dire comme le minimiseur d'une certaine énergie (mentionnons qu'une autre approche possible consiste à chercher des états critiques de l'énergie [189], plutôt qu'un minimiseur global). Nous considérons ici que la déformation de référence est la solution[17] u_μ du problème atomistique

$$\inf\left\{E_\mu(u);\, u \in X_\mu\right\}, \tag{1.43}$$

où l'espace X_μ englobe les contraintes que l'on impose à u (conditions aux limites, ...), et qui ne sont pas au centre du débat.

Puisqu'on dispose d'une approximation de l'énergie (l'expression (1.42)), il est naturel de chercher à approcher u_μ par la solution du problème variationnel

$$\inf\left\{E_c(u);\, u \in X_c\right\}, \tag{1.44}$$

où l'espace X_c tient compte des contraintes que l'on impose à u. Cependant, cette approche conduit à de grandes difficultés. Le problème (1.44) est très complexe à résoudre, car les domaines Ω_M et Ω_μ dépendent de u. A ce stade, nous n'avons pas encore précisément défini de $\Omega_M(u)$. Une définition naïve, pour $u \in \left(H^1(\Omega)\right)^3$, est

$$\Omega_M(u) = \left\{x \in \Omega;\ \|\nabla u(x)\|_\infty \leq C\right\},$$

où C est un certain seuil. Donc la fonction u est considérée comme régulière si son gradient est assez petit. Définie ainsi, la fonction $u \mapsto \Omega_M(u)$ est très irrégulière. Considérons en effet la suite de fonctions

$$\forall x \in \Omega, \quad u_n(x) = \left(C + \frac{1}{n}\right) x.$$

Cette suite converge fortement dans $\left(H^1(\Omega)\right)^3$ vers $u(x) = Cx$, et on a $\Omega_M(u_n) = \emptyset$ tandis que $\Omega_M(u) = \Omega$. Donc, même si u_n converge vers u de la manière la plus forte possible, il est possible que les ensembles $\Omega_M(u_n)$ et $\Omega_M(u)$ demeurent très

[17] Les problèmes d'existence et d'unicité sont abordés plus loin.

différents. Cette forte irrégularité de $u \mapsto \Omega_M(u)$ laisse penser que la résolution d'un problème tel que (1.44) va soulever de grandes difficultés théoriques et numériques.

Pour simplifier (1.44), une idée consiste à supprimer la dépendance de Ω_M avec u, i.e. à fixer *a priori* la partition $\Omega = \Omega_M \cup \Omega_\mu$: on définit alors l'énergie par

$$\begin{aligned} E_c(u, \Omega_M) &= \frac{1}{2|\Omega|} \int_{\Omega_M} W\left(\nabla u(x)\right) dx \\ &+ \frac{1}{2N} \sum_{i\varepsilon \in \varepsilon\mathbb{Z}^3 \cap \Omega_\mu} \sum_{j \neq i} V_0\left(\frac{|u^j - u^i|}{\varepsilon}\right), \end{aligned} \tag{1.45}$$

et cette fois-ci Ω_M est un paramètre, u est un champ dans le domaine Ω_M et un ensemble de variables discrètes (les vecteurs $u^i \in \mathbb{R}^3$, pour i tels que $i\varepsilon \in \varepsilon\mathbb{Z}^3 \cap \Omega_\mu$) dans le domaine Ω_μ.

Remarque 1.2.1 *Soulignons le fait que les énergies (1.40), (1.41), (1.42) et (1.45) sont toutes consistantes les unes avec les autres, en vertu de l'analyse limite exposée ci-dessus.*

Le problème variationnel associé à l'énergie (1.45) est

$$\inf \left\{ E_c(u, \Omega_M);\, u \in X_c(\Omega_M) \right\}, \tag{1.46}$$

et l'espace $X_c(\Omega_M)$ est défini de telle façon à ce que $u \in X_c(\Omega_M)$ soit régulier dans Ω_M. Le problème (1.46) est plus simple que le problème (1.44) (car la partition est fixée), néanmoins il fait intervenir en paramètre le domaine Ω_M qu'il faut donc spécifier. Idéalement, il s'agit de $\Omega_M(u_\mu)$ (car on espère que la solution de (1.46) approche la solution de référence u_μ), mais la solution du problème atomistique ne peut pas être calculée !

Nous présentons maintenant une méthode, la QuasiContinuum Method (QCM), proposée dans les années 1990 par une équipe de chercheurs américains (cf. [163, 165, 166] pour les premiers articles, et [159, 160, 162, 164, 167] pour des applications), et qui s'appuie sur la définition (1.45) de l'énergie et sur le problème variationnel (1.46) pour calculer une approximation de u_μ. La méthode repose sur l'algorithme itératif suivant, pour déterminer à la fois un domaine Ω_M et une fonction u : on se donne un domaine Ω_M, puis

1. à Ω_M fixé, on résout le problème (1.46) ;
2. les zones Ω_M et Ω_μ sont remises à jour en fonction des zones de régularité de la solution trouvée, suivant un certain critère, et on revient à l'étape 1.

Le premier critère proposé pour la remise à jour de la partition est basé sur le gradient de déformation, plus précisément sa partie symétrique, soit

$$\varepsilon(u) = \frac{1}{2}(\nabla u + \nabla u^T). \tag{1.47}$$

La déformation u est considérée comme irrégulière lorsque $\varepsilon(u)$ dépasse un certain seuil, qui est un paramètre de la méthode que l'utilisateur doit choisir. Précisons que seule une version discrétisée (par éléments finis) de la méthode a été proposée, ce qui élimine les questions d'existence et de régularité de $\varepsilon(u)$. Dans une version plus récente de la QCM, un autre critère a été proposé, basé sur les variations de $\varepsilon(u)$ d'un élément fini à un autre : si cette variation dépasse un certain seuil, alors u est considéré comme irrégulier sur les deux éléments finis, qui sont donc transférés du domaine Ω_M vers le domaine Ω_μ.

Un exemple typique d'utilisation de la méthode QCM est représenté sur la figure 1.12 : il s'agit du problème de la nanoindentation, qui consiste à enfoncer dans un matériau un obstacle[18] beaucoup plus dur (l'indenteur, dont le diamètre est de 25 nm), et à étudier la courbe donnant la force exercée sur l'indenteur en fonction de son enfoncement dans le matériau. Les aspects dynamiques ne sont pas pris en compte, la simulation est faite dans le régime quasistatique. La force maximale est atteinte lorsqu'un défaut (plus précisément, une dislocation) apparaît dans le réseau atomique, initialement parfait. On s'attend à ce que des singularités apparaissent sous l'indenteur : au départ, on se donne donc une zone Ω_μ telle que représentée sur la figure 1.12, et on la fait évoluer en fonction des critères décrits ci-dessus. La simulation comporte donc deux boucles : la boucle extérieure incrémente l'enfoncement de l'indenteur, et la boucle intérieure calcule la solution du problème variationnel avec un indenteur à une position fixe, en raffinant / déraffinant le maillage.

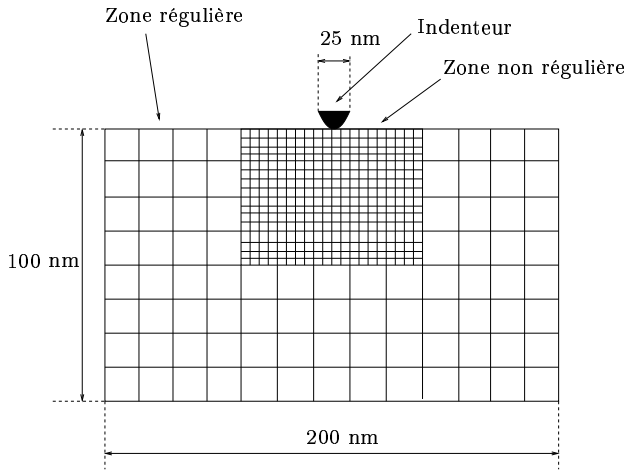

FIG. 1.12 – Un exemple d'application 2D de la méthode QCM : le problème de la nanoindentation. Dans la zone non régulière, le maillage correspond au réseau atomique. Dans la zone régulière, il s'agit d'un maillage élément fini traditionnel (ici, en quadrangle).

J'ai travaillé sur ce sujet avec Xavier Blanc et Claude Le Bris. Notre travail a

[18]En pratique, on ne traite pas un tel problème de frontière libre : on préfère représenter l'indenteur par un fort potentiel répulsif dont la zone d'application est très localisée, ce qui revient exactement à faire une méthode de pénalisation.

consisté à écrire les problèmes continus (c'est-à-dire non discrétisés en espace) (1.42)-(1.44) et (1.45)-(1.46), à essayer de comprendre le fondement d'une telle approche et à en faire l'analyse numérique. En particulier, quel critère doit gouverner la remise à jour de la partition ? Les approches couplées (1.42)-(1.44) et (1.45)-(1.46) sont-elles consistantes[19] avec l'approche atomistique (1.40)-(1.43), au sens où elles fournissent une approximation convergente de la solution atomistique ?

Pour ce faire, nous nous sommes placés dans un cadre très simplifié, en travaillant en une dimension, sur des matériaux décrits par un potentiel de paire V_0 à plus proches voisins, et soumis à des forces de volume f. Dans ce cas, l'énergie atomistique (1.40) s'écrit

$$E_\mu(u) = \frac{1}{N}\sum_{i=0}^{N-1} V_0\left(\frac{u^{i+1}-u^i}{\varepsilon}\right) - \frac{1}{N}\sum_{i=0}^{N} u^i\, f(i\,\varepsilon),$$

où u est l'ensemble de variables discrètes $(u^i)_{i=0}^N$ et où le nombre d'atomes N et le paramètre de maille atomique ε sont liés par $N\varepsilon = L$, où $L = |\Omega|$ est la longueur du matériau. L'énergie du modèle couplé (1.45) s'écrit

$$\begin{aligned} E_c(u,\Omega_M) &= \frac{1}{N}\sum_{i\in\mathbb{Z}\cap\Omega_\mu/\varepsilon} V_0\left(\frac{u^{i+1}-u^i}{\varepsilon}\right) - u^i f(i\varepsilon) \\ &+ \frac{1}{|\Omega|}\int_{\Omega_M} V_0(u'(x)) - u(x)f(x)\,dx. \end{aligned} \tag{1.48}$$

Les résultats obtenus dépendent de la convexité du potentiel de paire V_0 et sont les suivants [P4].

Si V_0 est strictement convexe, alors le problème atomistique (1.43) et le problème couplé (1.46), pour tout Ω_M fixé, sont bien posés et admettent un unique minimiseur. Il reste alors à fixer Ω_M, qui idéalement est $\Omega_M(u_\mu)$. A cause de la stricte convexité de V_0, les équations d'Euler-Lagrange des problèmes variationnels (1.43) et (1.46) sont des équations elliptiques, et par conséquent, on peut montrer que la zone d'irrégularité de u_μ est exactement la zone d'irrégularité des forces de volume f. Autrement dit, $\Omega_M(u_\mu)$ est déterminé par la simple connaissance de f. Il est alors possible de donner une définition précise de Ω_M en fonction de f, et d'estimer ensuite l'erreur entre la solution du problème couplé (1.46) et la solution du problème atomistique (1.43). Cette erreur converge vers 0 quand le paramètre de maille atomique ε tend vers 0.

Dans le cas non convexe, nous avons étudié un exemple particulier, celui du potentiel de Lennard-Jones (cf. la figure 1.13)

$$V_0(r) = V_{\mathrm{LJ}}(r) = \frac{1}{r^{12}} - \frac{2}{r^6},$$

qui possède les propriétés génériques d'un potentiel d'interaction :

[19] Nous avons déjà précisé ci-dessus que les énergies étaient consistantes, nous nous intéressons maintenant à la consistance des problèmes variationnels.

- à cause de la répulsion électronique, deux atomes ne peuvent être infiniment proches ; la propriété $V_{\mathrm{LJ}}(0) = +\infty$ rend compte de cette répulsion.
- lorsque deux atomes sont très éloignés l'un de l'autre, la force qu'ils exercent l'un sur l'autre est nulle ; le comportement de $V_{\mathrm{LJ}}(r)$ quand r tend vers $+\infty$ rend compte de ceci.
- il existe une distance d'équilibre pour la distance entre deux atomes. Le fait que $V'_{\mathrm{LJ}}(1) = 0$ et que $V''_{\mathrm{LJ}}(1) > 0$ rend compte de cette distance d'équilibre (l'unité de longueur a été choisie pour que la distance d'équilibre soit 1).

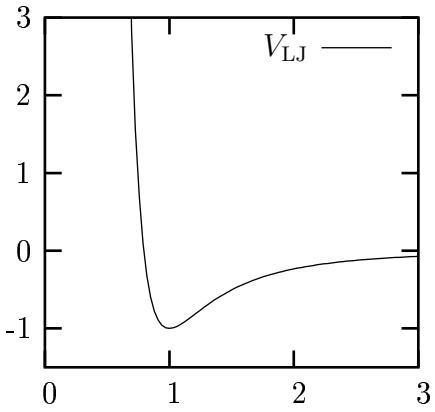

FIG. 1.13 – Potentiel de Lennard-Jones.

Lorsque le matériau est soumis à des conditions aux limites telles qu'il est en extension, on peut montrer que la solution du problème atomistique (1.43) fait apparaître une fracture (cf. la figure 1.14) : il existe un couple $(i, i+1)$ d'atomes tels que la distance $u_\mu^{i+1} - u_\mu^i$ entre ces deux atomes consécutifs est macroscopique (du même ordre de grandeur que la distance $u_\mu^N - u_\mu^0$ entre le premier et le dernier atome), alors que la distance entre tous les autres couples d'atomes consécutifs est proportionnelle à ε (le paramètre de maille atomique).

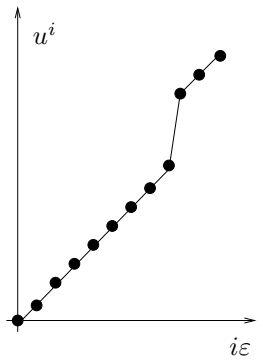

FIG. 1.14 – La solution du problème atomistique fait apparaître une fracture lorsque le matériau est mis en extension.

Le modèle macroscopique reproduit ce comportement, puisque, en extension, il

fait lui aussi apparaître une fracture, c'est-à-dire, à cette échelle, une discontinuité du déplacement.

L'utilisation d'une méthode multi-échelle doit permettre de calculer le réseau atomique au voisinage de la fracture par le modèle atomistique. Or, ceci est impossible si l'énergie du modèle couplé est définie par (1.48). En effet, supposons qu'il n'y a pas de force de volume, et considérons une déformation u faisant apparaître une fracture, qui est placée successivement dans la zone Ω_M puis dans la zone Ω_μ. Les énergies correspondantes sont notées $E_c(F \in \Omega_M)$ et $E_c(F \in \Omega_\mu)$, et on constate que

$$E_c(F \in \Omega_\mu) > E_c(F \in \Omega_M). \tag{1.49}$$

Donc le coût énergétique d'une fracture est plus grand lorsque celle-ci se trouve dans la zone Ω_μ. Lorsqu'on considére le problème variationnel (1.46), la fracture se place de telle façon à ce que son coût énergétique soit le plus faible : elle se place donc dans Ω_M, ce qui ne convient pas.

Nous proposons dans [P4] une façon de remédier à ce problème. Essentiellement, il s'agit de définir autrement l'énergie dans le modèle couplé, pour inverser l'inégalité (1.49). Au lieu de travailler avec (1.48), on travaille avec l'énergie

$$\begin{aligned} E_{\rm mod}(u,\Omega_M) &= \frac{1}{N}\sum_{i\in\mathbb{Z}\cap\Omega_\mu/\varepsilon} V_{\rm LJ}\left(\frac{u^{i+1}-u^i}{\varepsilon}\right) - u^i f(i\varepsilon) \\ &+ \frac{1}{|\Omega|}\int_{\Omega_M} V_{\rm LJ}^\varepsilon(u'(x)) - u(x)f(x)\,dx, \end{aligned} \tag{1.50}$$

avec

$$V_{\rm LJ}^\varepsilon(r) = V_{\rm LJ}(r) + \varepsilon^p\ \tau(r-r_0),$$

où p vérifie $0<p<1$ (on peut par exemple choisir $p=1/2$), τ est une régularisation de la fonction $t\in\mathbb{R}\mapsto t_+=\max(0,t)$, et où r_0 est un réel arbitraire choisi dans $]1,r_c[$ (par construction, le potentiel de Lennard-Jones $V_{\rm LJ}(r)$ est minimal en 1, convexe pour $r\le r_c=(13/7)^{1/6}$ et concave ensuite). La même comparaison des énergies des configurations fracturées donne cette fois-ci

$$E_{\rm mod}(F\in\Omega_\mu) < E_{\rm mod}(F\in\Omega_M). \tag{1.51}$$

L'énergie couplée (1.50) vérifiant cette inégalité, il est ensuite possible de construire une méthode multi-échelle. Le problème, comme dans le cas convexe, est de définir Ω_M sans passer par le calcul de u_μ. On montre en fait qu'on peut approximer $\Omega_M(u_\mu)$ par $\Omega_M(u_M)$, où u_M est la solution du problème variationnel à l'échelle macroscopique, qu'il est possible de résoudre. Une définition précise est proposée dans [P4].

1.2.1.2 Une méthode à température non nulle, dans un cadre dynamique

Comme souligné ci-dessus, la QCM est fondée sur une minimisation d'énergie, ce qui n'est pas compatible avec une température non nulle. Nous avons en effet

expliqué dans la section 1.1.1.2 (cf. la remarque 1.1.1) que travailler à la température $T > 0$, c'est considérer que le système peut être dans n'importe quel état énergétique (et pas seulement celui de plus basse énergie), et que la probabilité d'être dans un état (q, p) d'énergie $H(q, p)$ est donnée (à une constante de normalisation près) par la mesure de probabilité

$$d\mu_{NVT} = \exp\left(-\beta H(q,p)\right)\, dq\, dp,$$

avec

$$\beta = \frac{1}{k_B T},$$

où k_B est la constante de Boltzmann. La méthode QCM, fondée sur une minimisation d'énergie, est donc par construction une méthode pour traiter des problèmes statiques, à température nulle.

Nous détaillons maintenant une méthode [182] qui permet de travailler à température non nulle dans un cadre dynamique. Cette méthode est fondée sur l'idée de "coarse graining", i.e. de suppression de degrés de liberté.

Considérons un système unidimensionel de N atomes, de positions q_i, d'impulsions p_i et de masse m (pour simplifier, tous les atomes ont la même masse). On suppose que l'interaction se fait via un potentiel de paire V_ε qui ne fait intervenir que les plus proches voisins. L'énergie totale du système est

$$H(q,p) = \sum_{i=0}^{N} \frac{p_i^2}{2m} + \sum_{i=0}^{N-1} V_\varepsilon(q_{i+1} - q_i). \tag{1.52}$$

On travaille maintenant à température constante T, et on introduit la fonction de partition du système

$$Z = \int_{\Omega_{NVT}} d\mu_{NVT} = \int_{V_{occ}\times\mathbb{R}^{3N}} e^{-\beta H(q,p)}\, dq\, dp, \tag{1.53}$$

qui est une fonction de la température du système (V_{occ} est le volume occupé par le système). L'expression (1.52) permet de séparer l'intégrale en p de celle en q. On pose

$$Z_p = \int_{\mathbb{R}^{3N}} \exp\left(-\beta \sum_{i=0}^{N} \frac{p_i^2}{2m}\right)\, dp, \tag{1.54}$$

$$Z_q = \int_{V_{occ}} \exp\left(-\beta V(q)\right)\, dq, \tag{1.55}$$

avec $V(q) = \sum_{i=0}^{N-1} V_\varepsilon(q_{i+1} - q_i)$. Donc $Z = Z_p Z_q$ et Z_p peut être calculée analytiquement. Par contre, Z_q n'est pas facile à calculer, à cause du trop grand nombre de variables qui interviennent.

La fonction de partition Z permet le calcul de toutes les grandeurs thermodynamiques du système. Ainsi, en séparant les contributions de l'énergie cinétique et de l'énergie potentielle, l'énergie libre F s'écrit

$$F = -k_B T \ln Z = -k_B T \ln Z_p + F_q, \tag{1.56}$$

avec

$$F_q = -k_B T \ln Z_q, \tag{1.57}$$

tandis que l'entropie, qui est une mesure du désordre du système, est définie par

$$S = -\frac{\partial F}{\partial T}. \tag{1.58}$$

L'idée de la méthode que nous présentons ici est de supprimer des degrés de liberté, tout en conservant les propriétés thermodynamiques du système, ce qui est le cas dès que la fonction de partition est conservée. Comme l'expression de Z_p est connue, il faut donc s'attacher à conserver Z_q.

A titre d'exemple, nous expliquons ci-dessous comment la dépendance en q_{j+1} peut être supprimée. On considère donc le système composé des $N-1$ atomes $\{1, \ldots, j, j+2, \ldots, N\}$, et on définit $V_{(1)}(q_j, q_{j+2}, T)$ et $F_{(1)}(T, j+1)$ par

$$e^{-\beta\left(V_{(1)}(q_j, q_{j+2}, T) + F_{(1)}(T, j+1)\right)} = \int e^{-\beta(V_\varepsilon(q_j, q_{j+1}) + V_\varepsilon(q_{j+1}, q_{j+2}))}\, dq_{j+1}.$$

La dépendance de l'intégrale avec q_j et q_{j+2} est regroupée dans $V_{(1)}$, et le reste est dans $F_{(1)}(T, j+1)$, qui est l'énergie libre associée à la présence de l'atome $j+1$. Les fonctions $V_{(1)}$ et $F_{(1)}$ ne sont définies qu'à une constante près, mais ce n'est pas un problème car seules comptent leurs dérivées (une énergie potentielle est toujours définie à une constante près, la grandeur physique importante est son gradient, c'est-à-dire les forces ; pour ce qui est de l'énergie libre F, elle n'intervient dans le calcul des grandeurs thermodynamiques que par l'intermédiaire de ses dérivées, cf. par exemple (1.58)). Dans le nouveau système, le potentiel entre l'atome j et l'atome $j+2$ est $V_{(1)}$, il dépend *a priori* de la température. La quantité $F_{(1)}$ est la contribution à l'énergie libre de l'atome $j+1$.

L'originalité de cette approche est la prise en compte du terme $F_{(1)}$. La présence de l'atome $j+1$ apporte de l'entropie au système (cf. la relation (1.58)). Lorsqu'on cherche à construire un système ne contenant pas cet atome, il faut garder en mémoire cette contribution, ce qui est le rôle de $F_{(1)}$. Si on l'oublie, alors au fur et à mesure qu'on supprime des degrés de liberté, on retire au système de l'entropie, ce qui donne un résultat final faux.

Pour le système composé de $N-1$ atomes, l'énergie libre est la somme de 3 termes :

- un terme provenant de l'énergie cinétique, qui peut être calculé analytiquement ;

– le terme F_q provenant de l'énergie potentielle. Celle-ci s'écrit

$$V(q) = \sum_{i=0}^{j-1} V_\varepsilon(q_{i+1} - q_i) + V_{(1)}(q_j, q_{j+2}, T) + \sum_{i=j+2}^{N-1} V_\varepsilon(q_{i+1} - q_i), \tag{1.59}$$

et la contribution correspondante à la fonction de partition est donnée par (1.57), où Z_q est donné par (1.55) et V est donné par (1.59).

– le terme $F_{(1)}$.

Nous avons décrit dans le détail comment supprimer un degré de liberté. Il est tout à fait possible de supprimer un degré de liberté sur deux, en suivant la même méthode. Il est aussi possible de supprimer des degrés de liberté de façon non homogène dans le matériau. C'est intéressant lorsqu'on cherche à simuler un système pour lequel les champs sont singuliers dans certaines zones (on va alors conserver tous les degrés de liberté pour rendre compte au mieux de cette singularité), et réguliers dans d'autres zones (peu de degrés de liberté suffisent alors).

1.2.1.3 Couplage fondé sur la dynamique Hamiltonienne

Nous décrivons maintenant une autre méthode, initialement proposée par F.F. Abraham *et al* (les articles fondateurs sont [168, 169, 171], l'utilisation de cette méthode sur des problèmes concrets est décrite dans [174–181]). Cette méthode couple elle aussi un modèle de continuum avec un modèle atomistique. La démarche est toutefois différente, elle s'appuie sur la similarité entre les équations de la dynamique moléculaire et celles issues de la mécanique des milieux continus, après discrétisation en espace.

Cette méthode n'a pas été approfondie dans ce travail de thèse, aussi nous la décrivons telle que présentée par ses auteurs, i.e. dans une version déjà discrétisée en espace.

En dynamique moléculaire, l'évolution du système est gouvernée par son Hamiltonien, qui s'écrit, en supposant encore une fois que l'énergie potentielle ne fait intervenir que des interactions de paire,

$$H_\mu(q,p) = \frac{1}{2} \sum_{i=0}^{N-1} \sum_{j \neq i} V_\varepsilon(q_j - q_i) + \sum_{i=0}^{N} \frac{p_i^2}{2m}. \tag{1.60}$$

En mécanique des milieux continus, notant $u(x,t)$ le champ de déplacement et $\varepsilon(u)$ le gradient de déformation (donné par (1.47)), le Hamiltonien du système s'écrit, dans le cas de l'élasticité linéaire,

$$H_M(u,v) = \frac{1}{2} \int \left(\varepsilon(u) : \Lambda : \varepsilon(u) + \rho\, v^2 \right)\, dx, \tag{1.61}$$

où Λ est le tenseur d'élasticité (tenseur constant d'ordre 4), ρ la masse volumique et v la vitesse de déplacement. Après discrétisation en espace de H_M, on obtient

$$H_M^h(u_h, p_h) = \frac{1}{2}\, u_h^T(t) \cdot K \cdot u_h(t) + \frac{1}{2}\, p_h^T(t) \cdot M^{-1} \cdot p_h(t), \tag{1.62}$$

où $u_h(t)$ est le vecteur décrivant le champ $u(x,t)$ après discrétisation spaciale, K est la matrice de rigidité, et M est la matrice de masse. On fait l'approximation de la condensation de masse, si bien que la matrice M est diagonale. Le vecteur $p_h(t)$ est obtenu après discrétisation spatiale de l'impulsion $\rho\, v(x,t)$.

Les deux modèles (atomistique et continuum) ont donc la même structure, car ils sont tous les deux fondés sur une dynamique Hamiltonienne, à partir respectivement de (1.60) et de (1.62). Nous revenons ci-dessous sur l'hypothèse de l'élasticité linéaire pour le modèle de continuum et sur la consistance des deux Hamiltoniens.

Le couplage des deux modèles est fait via un unique Hamiltonien. On partitionne le système en deux zones, une zone Ω_μ dans laquelle on va utiliser la dynamique moléculaire et une zone Ω_M dans laquelle on va utiliser la mécanique des milieux continus : $\Omega = \Omega_M \cup \Omega_\mu$. Au niveau de l'interface, on définit une zone dite "Hand-Shake" (Ω_{HS}), là où on va coupler les deux modèles. Cette zone est à cheval entre les zones Ω_μ et Ω_M (cf. la figure 1.15).

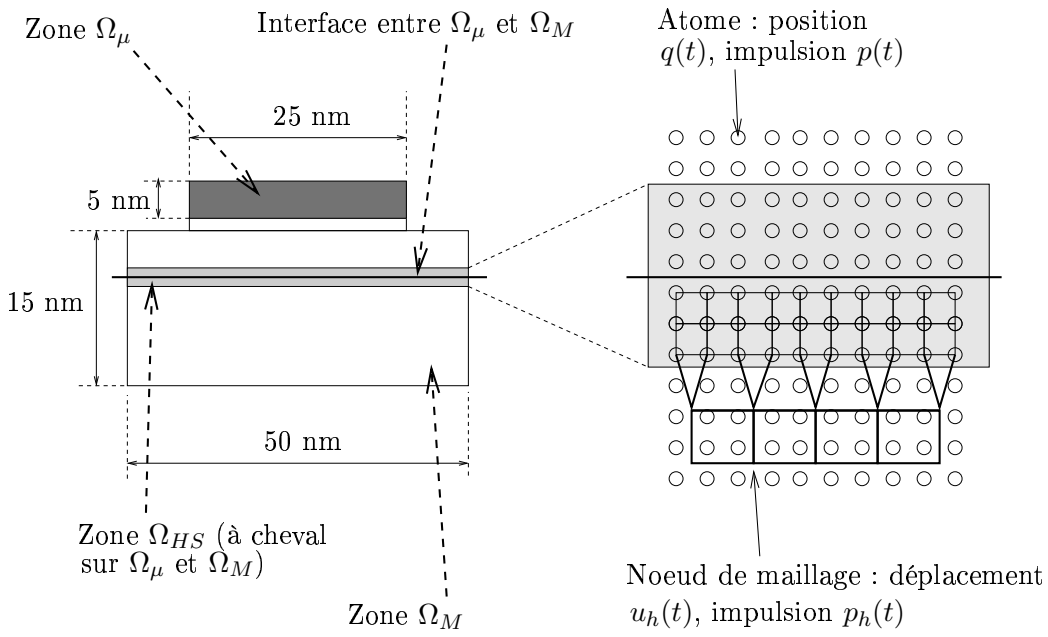

FIG. 1.15 – Exemple de décomposition $\Omega = \Omega_\mu \cup \Omega_M$, et schéma d'une zone intermédiaire Ω_{HS}. Le cas considéré est celui de la simulation 3D d'un film de Si_3N_4 (de dimension 25 nm $\times$ 5 nm, et représenté en grisé sur le dessin de gauche) sur un cristal de silicium Si (c'est le support situé sous le film). Le paramètre de maille atomique de Si est plus petit que celui de Si_3N_4 de 1.25 % : à l'interface entre les deux réseaux, on s'attend à de fortes distorsions, c'est pourquoi cette interface est immergée dans la zone Ω_μ. La figure de droite détaille les degrés de liberté.

Comme pour la méthode QCM, le domaine est partitionné. Cependant, la méthode décrite ici n'inclut pas de critère d'adaptation, et la partition n'évolue donc pas au cours de la simulation. Le problèmes traités ne sont cependant pas les mêmes :

ici, on étudie des cas pour lesquels on a une bonne idée *a priori* de la localisation des singularités. Les auteurs vérifient que la solution trouvée (pour une partition donnée) ne dépend pas de cette partition, en faisant un nouveau calcul avec une partition différente.

Comment choisir les zones Ω_M, Ω_μ et Ω_{HS} ? Dans la méthode telle que proposée par ses auteurs, ceci est fait de façon empirique. Cependant, tous les choix ne sont pas possibles. Nous avons en effet fait l'hypothèse de l'élasticité linéaire en écrivant le Hamiltonien (1.61) du modèle de continuum. La méthode n'a donc un sens que si la déformation dans la zone Ω_M est effectivement proche de l'identité. Nous avons aussi souligné ci-dessus l'importance de la consistance des modèles considérés. Cette consistance est assurée par le mode de calcul du tenseur d'élasticité Λ : à partir du potentiel V du modèle atomistique, il est possible de calculer les constantes élastiques du matériau (cf. la section 1.1.1.2), ce qui permet d'obtenir une valeur pour le tenseur Λ.

Dans la zone Ω_M, le champ u est discrétisé par éléments finis. Le maillage est tel que, dans la zone $\Omega_M \cap \Omega_{HS}$ (au voisinage de la zone Ω_μ), les nœuds du maillage et les atomes sous-jacents correspondent. Au fur et à mesure qu'on s'éloigne de la zone Ω_{HS}, la taille des éléments finis grandit, jusqu'à atteindre en pratique 4 à 8 unités atomiques.

Les degrés de liberté du problème sont les suivants (cf. la figure 1.15, à droite) :

- dans la zone Ω_μ, et en dehors de la partie Ω_{HS}, on utilise la dynamique moléculaire ; les degrés de liberté sont les positions $q_i(t)$ et les impulsions $p_i(t)$ des atomes présents dans cette zone.
- dans la zone Ω_M, et en dehors de la partie Ω_{HS}, on utilise la mécanique des milieux continus ; les degrés de liberté sont les déplacements $u_i^h(t)$ et les impulsions $p_i^h(t)$ aux nœuds du maillage.
- dans la zone Ω_{HS}, nœuds de maillage et atomes correspondent, on note $\chi_i(t)$ le déplacement (qui correspond à la différence entre position de l'atome et position initiale), et $w_i(t)$ l'impulsion. On va utiliser un modèle hybride.

On définit le Hamiltonien pour le système total par

$$\begin{aligned} H(q,p,u_h,p_h,\chi,w) &= \frac{1}{2}\sum_{i,j\in\Omega_\mu\setminus\Omega_{HS}} V_\varepsilon(q_j-q_i) + \sum_{i\in\Omega_\mu\setminus\Omega_{HS}} \frac{p_i^2(t)}{2m} \\ &+ \frac{1}{2}\,u_h^T(t)\cdot K\cdot u_h(t) + \frac{1}{2}\,p_h^T(t)\cdot M^{-1}\cdot p_h(t) \\ &+ V^{HS}(\chi(t)) + \sum_{i\in\Omega_{HS}} \frac{w_i^2(t)}{2m}, \end{aligned}$$

où l'expression de $V^{HS}(\chi(t))$ est donnée ci-dessous. Pour les degrés de liberté qui ne sont pas dans la zone Ω_{HS}, on reprend donc la même énergie que dans les deux modèles précédents. On explique maintenant la construction du terme hybride. A cause

de l'hypothèse de condensation de masse, et comme, dans la zone Ω_{HS}, le maillage est raffiné à l'échelle atomique, l'expression de l'énergie cinétique est naturelle (les modèles atomistique et de mécanique du continuum donnent la même expression).

L'expression de l'énergie potentielle n'est pas la même avec les deux modèles, et V^{HS} est définie comme une moyenne entre les deux expressions. On a noté par $\chi_i(t)$ le déplacement de l'atome i, donc sa position dans la configuration déformée est $R_i + \chi_i(t)$, où R_i est la position initiale de l'atome i. Le potentiel V^{HS} est défini par

$$V^{HS}(\chi) = \frac{1}{2} \sum_{i \in \Omega_{HS}} \sum_{j \neq i} \theta_{ij}\ \chi_i K_{ij} \chi_j \ + \ (1 - \theta_{ij})\ V_\varepsilon\big(\chi_i + R_i - \chi_j - R_j\big),$$

où K est la matrice de rigidité dans le modèle de mécanique du continuum tandis que V_ε est le potentiel de paire du modèle atomistique. La fonction θ_{ij} permet de donner un poids différent aux deux contributions, suivant que les deux degrés de liberté $\chi_i(t)$ et $\chi_j(t)$ sont plutôt dans la zone Ω_M ou plutôt dans la zone Ω_μ. Le choix précis de cette fonction est assez empirique.

Dans la méthode décrite ci-dessus, les descriptions atomistique et de continuum se recouvrent dans la zone Ω_{HS}. Dans [170], la méthode décrite ci-dessus est comparée à une méthode similaire, où cette fois-ci les deux descriptions ne se recouvrent pas.

1.2.2 Homogénéisation de matériaux polycristallins

La seconde approche multi-échelle abordée dans cette thèse, et que nous décrivons maintenant, concerne des échelles tout à fait différentes de celles qui sont en jeu dans l'approche précédente. On s'intéresse ici à des matériaux hétérogènes, décrits par la mécanique du continuum, et on cherche à en déterminer les propriétés effectives. La démarche qu'on présente ici est donc à rapprocher de la théorie de l'homogénéisation des EDP elliptiques.

Les matériaux hétérogènes sont depuis longtemps utilisés dans l'industrie. Ils permettent de faire des compromis entre différentes exigences : masse volumique (les matériaux légers sont plus intéressants), résistance mécanique, coût, ... Un exemple est celui des matériaux composites : un réseau de fibres, constituées d'un matériau résistant, est "plongé" dans un matériau moins résistant (la matrice), mais par exemple meilleur marché ou plus léger (cf. la figure 1.16). Comprendre comment les propriétés des matériaux constituants et la géométrie de l'assemblage influence les propriétés effectives du matériau ouvre la voix à la création de nouveaux matériaux possédant des propriétés mécaniques plus intéressantes.

Dans d'autres cas, l'hétérogénéité n'est pas un choix, mais une contrainte, venant soit du processus de fabrication, soit de l'utilisation du matériau. On peut penser aux mousses de caoutchouc utilisées dans l'industrie automobile (joints de portières, joints entre le moteur et le chassis, ...) : ces mousses comportent initialement de

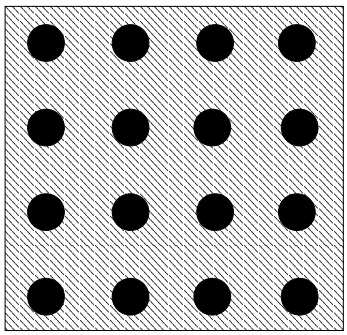

FIG. 1.16 – Schéma représentatif d'un matériau composite (coupe 2D).

petites cavités, et avec l'usure, ces cavités grossissent et parfois coalescent. Le calcul précis du comportement effectif d'un tel matériau est nécessaire pour mieux prévoir son temps de vie, par exemple.

La détermination des propriétés effectives de matériaux hétérogènes fait l'objet d'une importante littérature. Le cas des matériaux élastiques linéaires a été largement étudié. Dans le cas périodique, le problème modèle est l'étude du comportement quand ε tend vers 0 de la solution u^ε de l'équation

$$\begin{aligned} -\text{div}\ \left(a\left(x,\frac{x}{\varepsilon}\right)\nabla u_\varepsilon\right) &= f \text{ dans } \Omega, \\ u_\varepsilon &= 0 \text{ sur } \partial\Omega, \end{aligned} \tag{1.63}$$

où $a(x,y)$ est une matrice périodique en la variable y de période 1 et uniformément coercive. Ce problème a été traité sur le plan théorique, par plusieurs méthodes (cf. [140]). A la limite $\varepsilon \to 0$, on obtient un problème de même nature que (1.63) (une EDP elliptique), et des méthodes numériques standard peuvent être utilisées pour sa résolution. La solution de (1.63) fait apparaître des oscillations à l'échelle ε, et, dans certains cas, il est intéressant de calculer la solution u_ε à cette échelle (on ne souhaite pas passer à la limite $\varepsilon \to 0$ car on étudie un problème concret pour lequel ε a une valeur, certes petite, mais non nulle). Des méthodes numériques adaptées ont été développées pour répondre à cette question (cf. par exemple [147, 151, 153]).

Le cas de matériaux élastiques nonlinéaires et périodiques a été étudié dans [150]. Soit

$$\varepsilon(u) = \frac{1}{2}\left(\nabla u^T + \nabla u\right)$$

le tenseur des déformations (qu'on ne confondra pas avec la petite échelle ε destinée à tendre vers 0). Notant f les forces de volume auxquelles le matériau est soumis, $\Omega \subset \mathbb{R}^d$ le volume occupé par le matériau et $u : \Omega \to \mathbb{R}^d$ le déplacement, l'énergie s'écrit

$$I_\varepsilon(u) = \int_\Omega W\left(\frac{x}{\varepsilon}, \varepsilon(u(x))\right) - f(x)\,u(x)\,dx, \tag{1.64}$$

et on suppose que W est une fonction strictement convexe de la seconde variable et Y-périodique de la première. Sous des hypothèses de croissance sur W et de régularité sur f, le problème variationnel

$$\inf \left\{ I_\varepsilon(u);\, u \in \left(W_0^{1,p}(\Omega)\right)^d \right\} \tag{1.65}$$

est bien posé. Dans [150], on montre que le problème homogénéisé associé à (1.65) s'écrit

$$\inf \left\{ \overline{I}(u);\, u \in \left(W_0^{1,p}(\Omega)\right)^d \right\}, \tag{1.66}$$

où

$$\overline{I}(u) = \int_\Omega \overline{W}\left(\varepsilon(u(x))\right) - f(x)\, u(x)\, dx.$$

Dans cette expression, $\overline{W}$ est le potentiel effectif, défini pour toute matrice symétrique λ de taille 3×3 par

$$\overline{W}(\lambda) = \inf \left\{ \int_Y W\left(y, \lambda + \varepsilon(w(y))\right)\, dy;\; w \in \left(W_\#^{1,p}(Y)\right)^d \right\}, \tag{1.67}$$

où $W_\#^{1,p}(Y)$ est l'espace des fonctions appartenant à $W^{1,p}(Y)$ et qui sont Y-périodiques.

Le lien entre (1.65) et (1.66) est le suivant : comme W est strictement convexe de sa seconde variable, on peut montrer que $\overline{W}$ est aussi une fonction strictement convexe, si bien que le minimiseur $\overline{u}$ de (1.66) est unique. Alors le problème (1.65) converge vers le problème (1.66) au sens où, quand ε tend vers 0,

$$\begin{aligned} u_\varepsilon &\rightharpoonup \overline{u} \text{ faiblement dans } \left(W_0^{1,p}(\Omega)\right)^d, \\ I_\varepsilon(u_\varepsilon) &\to \overline{I}(\overline{u}). \end{aligned}$$

Résoudre directement un problème tel que (1.65) nécessite une discrétisation du domaine Ω à l'échelle ε, c'est donc un calcul très coûteux. L'analyse ci-dessus montre qu'une bonne approximation de u_ε est fournie par la résolution du problème (1.66), qui ne fait plus intervenir de petites échelles (on peut donc utiliser une discrétisation bien plus grossière).

Dans la littérature de mécanique, un grand nombre de bornes ont été proposées pour la fonction $\overline{W}(\lambda)$, c'est-à-dire de fonctions $\overline{W}_{inf}(\lambda)$ et $\overline{W}_{sup}(\lambda)$ telles que

$$\forall \lambda, \quad \overline{W}_{inf}(\lambda) \leq \overline{W}(\lambda) \leq \overline{W}_{sup}(\lambda).$$

On pourra par exemple consulter [192, 198–200]. Des approches multi-échelles plus complexes ont été développées dans [193, 195–197].

Ecrire l'énergie du matériau sous la forme (1.64) est équivalent à écrire que la loi constitutive du matériau, ou encore *loi de comportement*, qui relie le champ de contrainte $\sigma(x)$ au gradient de déformation $\varepsilon(x)$, s'écrit

$$\sigma(x) = \partial_2 W\left(\frac{x}{\varepsilon}, \varepsilon(x)\right), \tag{1.68}$$

où $\partial_2 W$ est la dérivée de W par rapport à sa deuxième variable. En notant $U(y,\sigma)$ la transformée de Legendre de $W(y,\lambda)$ par rapport à λ, la loi (1.68) se récrit

$$\varepsilon(x) = \partial_2 U\left(\frac{x}{\varepsilon}, \sigma(x)\right). \tag{1.69}$$

Cependant, pour beaucoup de matériaux, la loi constitutive ne s'écrit pas sous la forme (1.69). C'est le cas des matériaux élasto-viscoplastiques, dont la loi de comportement fait intervenir deux potentiels, un *potentiel viscoplastique* et un *potentiel élastique*. En relachant aussi l'hypothèse de périodicité, la loi de comportement de tels matériaux s'écrit

$$\dot{\varepsilon}(x,t) = \frac{\partial U_\varepsilon^{vp}}{\partial \sigma}(x, \sigma(x,t)) + \frac{\partial U_\varepsilon^{e}}{\partial \dot{\sigma}}(x, \dot{\sigma}(x,t)), \tag{1.70}$$

où $U_\varepsilon^{vp}(x,\sigma)$ est le potentiel viscoplastique et $U_\varepsilon^{e}(x,\dot{\sigma})$ est le potentiel élastique. La dépendance en x des potentiels rend ces matériaux hétérogènes, et on suppose que la distance caractéristique de variation de U_ε^{vp} et U_ε^{e} par rapport à la première variable est ε.

La principale différence entre (1.69) et (1.70) est que cette dernière est une équation dépendante du temps. Pour de tels matériaux, on ne dispose plus d'aucune borne sur le comportement effectif. Il n'est même pas évident *a priori* que la loi effective soit du même type que la loi (1.70).

Une partie du travail de cette thèse a consisté en l'étude numérique de tels matériaux [P5]. A l'échelle du micromètre, pour certains matériaux comme les alliages de zirconium, la matière est formée de grains. Chaque grain peut être supposé homogène et suivre une loi de comportement du type (1.70). D'un grain à un autre, les paramètres de la loi changent, si bien que globalement, un assemblage de grains (aussi appelé polycristal, car le réseau atomique dans un grain est supposé quasiment parfait, donc chaque grain est un monocristal) est un matériau hétérogène et non périodique (cf. la figure 1.17) qui suit une loi de comportement du type (1.70), où ε est la taille caractéristique des grains.

L'approche suivie consiste à travailler séparément sur chaque terme de la loi (1.70). Pour le potentiel viscoplastique comme pour le potentiel élastique, il est possible de définir un potentiel effectif, soit respectivement $U_{\text{eff}}^{vp}(\sigma)$ et $U_{\text{eff}}^{e}(\dot{\sigma})$, en utilisant (1.67). On *postule* alors que la loi effective s'écrit sous la forme

$$\dot{\varepsilon}(x,t) = \frac{\partial U_{\text{eff}}^{vp}}{\partial \sigma}(\sigma(x,t)) + \frac{\partial U_{\text{eff}}^{e}}{\partial \dot{\sigma}}(\dot{\sigma}(x,t)).$$

On compare ensuite le comportement d'un matériau décrit par une telle loi constitutive avec le comportement d'un matériau décrit la loi (1.70). Pour les cas considérés, on constate un bon accord.

1.3 Perspectives

L'exposé ci-dessus laisse beaucoup de questions non explorées, et pose aussi un certain nombre de questions nouvelles. Nous en citons quelques unes.

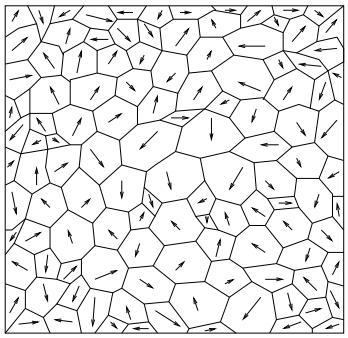

FIG. 1.17 – Structure d'un polycristal : la flèche symbolise l'orientation du réseau atomique dans chaque grain. La loi de comportement à l'intérieur d'un grain est homogène, et dépend de l'orientation du grain.

En dynamique moléculaire, nous avons étudié le cas le plus simple, celui du calcul de moyennes thermodynamiques dans l'ensemble NVE. Je souhaite poursuivre le travail en m'intéressant au calcul d'autres grandeurs (des coefficients d'auto-corrélation par exemple), dans d'autres ensembles thermodynamiques.

Le calcul de coefficients d'auto-corrélation (cf. la remarque 1.1.2) est un problème important en pratique. Il est donc intéressant d'étendre à ce type de calculs les schémas à convergence rapide développés jusqu'ici. La question est à la fois simple (du point de vue de l'implémentation) et difficile (du point de vue de l'analyse numérique). Concernant le second point, on ne sait pas par exemple démontrer que le coefficient de diffusion, qui est défini en dynamique moléculaire à partir des corrélations en vitesse, est un nombre mathématiquement bien défini. Il est néanmoins possible de traiter, à court terme, d'autres coefficients d'auto-corrélation, comme par exemple le calcul du spectre de phonons dans un solide [43], pour lesquels l'analyse semble plus simple.

A plus long terme, une autre piste concerne l'extension de notre travail au cas de l'ensemble thermodynamique NVT. Nous avons expliqué que l'approche "systèmes étendus" pose des problèmes (cf. la section 1.1.5.1), est-il possible de calculer plusieurs moyennes sur des trajectoires à énergie constante (ce qu'on sait bien faire) et de les combiner ensuite pour obtenir une moyenne dans l'ensemble NVT ? L'idée est donc de feuilleter l'espace des phases accessible dans l'ensemble NVT en fonction de l'énergie du système.

Sur le calcul proprement dit de la dynamique, beaucoup de méthodes ont été développées, soit pour réduire le coût calcul des forces, soit pour travailler avec un pas de temps plus grand. La méthode FMM, qui répond à la première question, est maintenant bien comprise sur le plan de l'analyse numérique, et elle est couramment utilisée en dynamique moléculaire. Nous avons cependant mentionné le fait que l'énergie potentielle du système telle qu'approximée par la méthode FMM est une fonction discontinue de la position des particules. Or, toute l'analyse numérique

pour le calcul des moyennes en temps long repose sur la théorie KAM, qui suppose une régularité très forte du Hamiltonien. Donc, lorsqu'on travaille avec le potentiel FMM, la théorie ne s'applique plus. Il est donc intéressant de chercher à comprendre comment les résultats obtenus dépendent de la régularité du potentiel. Des premières observations numériques semblent montrer que lorsque le potentiel est de classe $\mathcal{C}^1$, le système se comporte bien. Que se passe-t-il si le potentiel n'est que de classe $\mathcal{C}^0$? ou bien discontinu ? Ce sujet est important dans la pratique (la méthode FMM étant très utilisée), mais aussi sans doute très difficile sur le plan de l'analyse numérique. Dans un premier temps, nous envisageons surtout de faire des expérimentations numériques, afin d'essayer de comprendre ce qui se passe et quelles peuvent être de bonnes stratégies.

Enfin, comme expliqué dans la section 1.1.6, un des problèmes majeurs de la dynamique moléculaire est la différence des échelles de temps entre les temps sur lesquels il est possible de simuler un système et les temps caractéristiques de certains phénomènes qu'on souhaite étudier. Plusieurs méthodes pour résoudre ce problème ont été proposées. Je me suis intéressé à la méthode développée dans [107, 108, 115, 116] au cours du CEMRACS 2004[20], et aussi via l'encadrement d'un stagiaire de DEA, Mohamed El Makri[21].

Concernant les méthodes multi-échelles, au moins deux pistes de travail ont été dégagées. La première concerne le développement d'éléments finis adaptés à des densités d'énergie élastique qui présentent des singularités, et l'autre concerne l'extension de notre travail sur les méthodes couplant modèle atomistique et modèle de continuum à un cadre dynamique.

Supposons que la solution du problème atomistique soit régulière *à l'échelle atomistique* : alors l'expression intégrale de l'énergie (celle du modèle de mécanique) est une bonne approximation pour calculer l'énergie du modèle atomistique (qui est une somme discrète portant sur tous les atomes du système), et dans le cas où le potentiel est convexe, nous avons expliqué que la solution du problème macroscopique est une approximation convergente de la solution du problème atomistique. On suppose maintenant que le potentiel du modèle de mécanique du continuum est singulier en 0 (ce qui provient de l'impossibilité pour des atomes d'être au même point). Alors il est possible que la solution du problème de mécanique exhibe des irrégularités *à l'échelle macroscopique*, à cause de la singularité du potentiel. Par conséquent, une approche basée sur des éléments finis usuels peut conduire à une mauvaise approximation de la solution. Une collaboration avec Claude Le Bris et Yvon Maday est en cours, pour construire un élément fini adapté à cette situation.

L'étude du cas non convexe montre que l'écriture naturelle de l'énergie couplée

[20] CEMRACS est un acronyme pour Centre d'été Mathématique de Recherche Avancée en Calcul Scientifique. Le site web de l'édition 2004 de cette école d'été est `http ://smai.emath.fr/cemracs/cemracs04/index.php`.

[21] Elève du DEA *Equations Aux Dérivées Partielles et Applications*, Paris Dauphine, que j'ai encadré pendant son stage en collaboration avec Eric Cancès.

conduit à des problèmes, au moins dans le cas simplifié que nous avons considéré. Qu'en est-il en dimension supérieure ?

Le modèle atomistique avec lequel nous avons travaillé montre que le matériau casse dès qu'il est mis en extension. Ceci n'est pas très physique, on s'attend plutôt à ce que le matériau casse au delà d'une certaine extension. Ce comportement vient des propriétés génériques du potentiel de Lennard-Jones, et du fait que nous avons posé le problème comme un problème de minimisation. Comme expliqué ci-dessus, les propriétés génériques du potentiel semblent motivées par des arguments physiques. Par conséquent, c'est peut-être la notion de minimisation globable qu'il faut revoir.

Si on ne souhaite pas chercher un minimiseur global, une autre approche consiste à définir la solution comme un minimiseur local de l'énergie. Une telle approche a peut être plus de sens physique, mais soulève des difficultés sur le plan de l'analyse numérique : en général, il y a beaucoup de minimiseurs locaux ; faut-il en privilégier un par rapport aux autres ? et dans ce cas, sous quel critère ? Une autre idée est de considérer un problème dynamique, et d'essayer de coupler une dynamique Hamiltonienne décrivant un système discret avec l'équation de l'élastodynamique.

Chapitre 2

Schémas d'ordre élevé pour le calcul de moyennes en dynamique moléculaire

Ce chapitre reprend l'intégralité d'un article écrit en collaboration avec Eric Cancès, François Castella, Philippe Chartier, Erwan Faou, Claude Le Bris et Gabriel Turinici, et accepté dans *Journal of Chemical Physics* [P1].

La dynamique moléculaire repose sur l'équivalence entre moyennes dans l'espace des phases d'une part et moyennes temporelles le long d'une solution d'une équation différentielle ordinaire d'autre part, et permet d'approcher les premières en calculant les secondes. Nous nous intéressons à la vitesse de convergence des moyennes temporelles vers leur limite en temps infini et nous proposons des schémas numériques permettant d'accélérer cette convergence. Dans ce chapitre, ces schémas sont mis en œuvre sur plusieurs exemples de systèmes physiques, tandis que leur analyse numérique est présentée au chapitre 3.

High-order averaging schemes with error bounds for thermodynamical properties calculations by molecular dynamics simulations

Eric Cancès[a], François Castella[b,c], Philippe Chartier[c], Erwan Faou[c], Claude Le Bris[a], Frédéric Legoll[a,d] and Gabriel Turinici[a]

[a] *CERMICS, Ecole Nationale des Ponts et Chaussées, 6 et 8 avenue Blaise Pascal, 77455 Marne-la-Vallée Cedex 2*
and
MICMAC, INRIA Rocquencourt, Domaine de Voluceau, 78153 Le Chesnay Cedex
[b] *IRMAR, Université de Rennes 1, Campus de Beaulieu, 35042 Rennes Cedex*
[c] *IPSO, INRIA Rennes, Campus de Beaulieu, 35042 Rennes Cedex*
[d] *EDF R & D, Analyse et Modèles Numériques, 1, avenue du Général de Gaulle, 92140 Clamart*

{cances,lebris,legoll}@cermics.enpc.fr,
francois.castella@univ-rennes1.fr,
{chartier,efaou}@irisa.fr,
Gabriel.Turinici@inria.fr

We introduce high-order formulae for the computation of statistical averages based on the long-time simulation of molecular dynamics trajectories. In some cases, this allows us to significantly improve the convergence rate of time averages toward ensemble averages. We provide some numerical examples that show the efficiency of our scheme. When trajectories are approximated using symplectic integration schemes (such as velocity Verlet), we give some error bounds that allow to fix the parameters of the computation in order to reach a given desired accuracy in the most efficient manner.

2.1 Introduction

The properties of a given physical system at thermodynamical equilibrium (radial distribution functions, free energies, transport coefficients, ...) can be computed as averages of some observables over the phase space of a representative microscopic system [2, 7].

In most applications of interest, this microscopic system is composed of a high number of particles (typically more than 100,000 atoms for biological systems), so that the dimension of the phase space, which is 6 times the number of particles, is also very large. This makes the computation of averages a challenging issue.

Two families of methods are commonly used in order to sample the phase space according to the convenient probability density (which depends on the thermodynamic ensemble in which the calculation is performed : NVE, NVT, NPT, ...) : Monte Carlo methods [2,7] on the one hand, and Molecular Dynamics (MD) on the other hand. In the latter case, the time evolution of the microscopic system is simulated (possibly coupled with a bath) ; under the ergodicity assumption, the trajectory of the system samples the phase space and the time average of the observable over the trajectory converges toward the ensemble average when time goes to infinity [57].

When time averages are estimated by MD calculations, a numerical scheme is needed to compute the dynamics of the system. The numerical trajectory covers a finite but large range of time and in addition may deviate significantly from the exact trajectory for large times. It is however of common belief and indeed commonly observed that, in many situations, results obtained by time averages are satisfactory.

In this article, we introduce and analyze a new method for calculating time averages, that can be used either to speed up the convergence at a fixed given accuracy, or to achieve a better accuracy at a given computational time.

The motivation for the introduction of such a method is the observation, complemented by a rigorous numerical analysis on simple cases, that the error between the time average and the ensemble average often decreases as $1/T$ for long times when a standard discretization of $\frac{1}{T}\int_0^T A(t)\,dt$ is employed (here and below, $A(t)$ denotes the value of the observable under study at time t). On the contrary, if the time average is evaluated by the formula we introduce, namely $\int_0^T A(t)\,f_{k,T}(t)\,dt$ where $f_{k,T}(t)$ is some *filtering function* (to be made precise below, see formulae (2.10) and (2.11)), then the convergence rate can be in general improved to $1/T^k$, for arbitrary large k. This rate can again be rigorously established in simple cases, and the precise estimation of the error (as a function of the time integration step, the trajectory length, the order of the numerical scheme and other parameters of the method) can be used in order to efficiently calibrate all these parameters.

Our method is presented in Sec. 2.2. Under some strong hypotheses collected in Appendix 2.6 (in particular, we assume that the Hamiltonian system is integrable), and if the numerical scheme used to compute the trajectory is symplectic, we are able to rigorously establish the error estimate mentioned above. The rather intricate and mathematically demanding proof of this estimate can be read in a companion article [P2].

We next turn in Sec. 2.3 to applications to examples of physical interest. Test cases are performed in order to illustrate the method and show the sharpness of the error bounds obtained analytically. We provide in Secs. 2.3.1 and 2.3.2 numerical examples on integrable systems that show that our estimate is indeed sharp. Of course, very few of the systems used in physics or chemistry are integrable. Fortunately, even if we are not able to prove it rigorously, our estimate seems to also hold true in more realistic and complex situations, as it is shown in Sec. 2.3.3,

where convincing results on a system of particles interacting through a Lennard-Jones potential are provided, and in Sec. 2.3.4, where alkane chains are studied. In Sec. 2.3.5, we simulate a 2D particle in a double well potential, in the regimes of low or high energy. This toy model mimicks larger systems whose potential energy surface presents several basins corresponding to metastable states.

For all the examples that we study in Sec. 2.3, time averages computed by MD converge to ensemble averages at the rate $1/T$. In this case, we show that our method allows one to speed up this convergence rate to $1/T^k$. In Sec. 2.4, we continue the study of systems that present several metastable states, and we focus on the regime when transitions between these states are possible but they are rare events in comparison with the faster characteristic time scale of the system. This energy regime is outside the scope of our theoretical analysis, and we observe that the rate of convergence of time averages calculated by MD is slowed down to $1/\sqrt{T}$.

We wish to emphasize that we only deal here with the NVE thermodynamical ensemble. The equations of motion which correspond to this ensemble are the Newton equations (governing the "physical" dynamics of the system), which can be recast as a Hamiltonian dynamical system. Although we have no definite conclusions to date, similar results are likely to hold for other statistical ensembles, provided the underlying dynamical system is Hamiltonian; that is the case when Nosé-Poincaré thermostats [61] are used to perform computations in the NVT ensemble.

2.2 Main setting and result

Let us consider M particles in 3D. Each particle i is described by its position $\mathbf{q}_i \in \mathbb{R}^3$, its momentum $\mathbf{p}_i \in \mathbb{R}^3$ and its mass m_i. Let $H(\mathbf{q}, \mathbf{p})$ be the Hamiltonian of the system, defined on $\mathbb{R}^{3M} \times \mathbb{R}^{3M}$ by

$$H(\mathbf{q}, \mathbf{p}) = \sum_{i=1}^{M} \frac{\mathbf{p}_i^2}{2m_i} + V(\mathbf{q}_1, \ldots, \mathbf{q}_M),$$

where V is the interaction potential, and $\mathbf{q}$ and $\mathbf{p}$ are notations for $(\mathbf{q}_1, \ldots, \mathbf{q}_M)$ and for $(\mathbf{p}_1, \ldots, \mathbf{p}_M)$ respectively. The Hamiltonian dynamical system associated to H is given by

$$\left\{ \begin{array}{rcl} \dfrac{d\mathbf{q}}{dt} & = & \dfrac{\partial H}{\partial \mathbf{p}}(\mathbf{q}(t), \mathbf{p}(t)), \\ \dfrac{d\mathbf{p}}{dt} & = & -\dfrac{\partial H}{\partial \mathbf{q}}(\mathbf{q}(t), \mathbf{p}(t)). \end{array} \right. \tag{2.1}$$

We suppose that the dynamical system (2.1) is integrable (see Appendix 2.6) and that the numerical scheme we use to integrate it is symplectic (recall that, for instance, the velocity Verlet scheme is symplectic). Since the dynamical system is integrable, it has $3M$ invariant functions $I_j(\mathbf{q}, \mathbf{p})$, $j = 1, \ldots, 3M$, which means that $I_j(\mathbf{q}(t), \mathbf{p}(t))$, where $(\mathbf{q}(t), \mathbf{p}(t))$ is solution of (2.1), is constant with respect to time.

Of course, $H(\mathbf{q},\mathbf{p})$ is one of these invariants, since the energy is preserved by the dynamics. We denote by

$$S(\mathbf{q},\mathbf{p}) = \left\{(\mathbf{x},\mathbf{y}) \in \mathbb{R}^{3M} \times \mathbb{R}^{3M} \text{ s.t.} \forall j \in [1,3M], I_j(\mathbf{x},\mathbf{y}) = I_j(\mathbf{q},\mathbf{p})\right\} \tag{2.2}$$

the level set of the invariant functions $\{I_j\}_{1\leq j\leq 3M}$ containing the phase space point $(\mathbf{q},\mathbf{p})$.

Let $(\mathbf{q}_0,\mathbf{p}_0)$ be an initial condition, and let $A(\mathbf{q},\mathbf{p})$ be an observable whose average is well-defined in the NVE ensemble. As we suppose that the dynamical system has $3M$ invariants, the actual trajectory starting from the initial point $(\mathbf{q}_0,\mathbf{p}_0)$ remains for all times on the level set $S(\mathbf{q}_0,\mathbf{p}_0)$. Thus, the ensemble average of the observable A reads

$$\langle A\rangle = \frac{\displaystyle\int_{S(\mathbf{q}_0,\mathbf{p}_0)} A(\mathbf{q},\mathbf{p})\, d\mu(\mathbf{q},\mathbf{p})}{\displaystyle\int_{S(\mathbf{q}_0,\mathbf{p}_0)} d\mu(\mathbf{q},\mathbf{p})}, \tag{2.3}$$

where $d\mu(\mathbf{q},\mathbf{p})$ is the invariant measure [P2] on $S(\mathbf{q}_0,\mathbf{p}_0)$.

Under some hypotheses collected in Appendix 2.6, the so-called ergodic theorem [P2,3] holds :

$$\lim_{T\to\infty} \frac{1}{T}\int_0^T A(\mathbf{q}(t),\mathbf{p}(t))\, dt = \langle A\rangle, \tag{2.4}$$

where $(\mathbf{q}(t),\mathbf{p}(t))$ is the trajectory of the system (2.1) starting from the initial point $(\mathbf{q}_0,\mathbf{p}_0)$. As an approximation to (2.3), it is a standard approach to use the discrete sum

$$\langle A\rangle_{num}^{Rie}(\delta t, T) = \frac{1}{\mathcal{N}}\sum_{j=0}^{\mathcal{N}-1} A\left(\mathbf{q}_j,\mathbf{p}_j\right), \tag{2.5}$$

where $(\mathbf{q}_j,\mathbf{p}_j)_{j=0}^{\mathcal{N}-1}$ is the numerical trajectory given by an integration scheme with time step $\delta t = T/\mathcal{N}$ applied on Eqs. (2.1). However, as the convergence in (2.4) occurs [3] only at the rate $1/T$, we cannot expect a faster convergence rate of (2.5) toward $\langle A\rangle$.

In order to improve this asymptotic convergence rate, we suggest to replace the uniformly-weighted time average $\dfrac{1}{T}\displaystyle\int_0^T A(\mathbf{q}(t),\mathbf{p}(t))\, dt$ by

$$\int_0^T A(\mathbf{q}(t),\mathbf{p}(t))\, f_T(t)\, dt, \tag{2.6}$$

where f_T acts as a filtering function of the signal $t \mapsto A(\mathbf{q}(t),\mathbf{p}(t))$. We present two different types of filtering functions, that both improve the convergence rate. The first one allows for a better understanding of our method, whereas the second one (see formulae (2.10) and (2.11)) is easier to implement. Denoting by

$B_1(t_1) = \dfrac{2}{T}\displaystyle\int_{t_1}^{t_1+T/2} A(\mathbf{q}(t),\mathbf{p}(t))\,dt$ the time average in the interval $[t_1, t_1+T/2]$ (with $t_1 \in [0, T/2]$), one can consider the average

$$\begin{aligned} \langle A\rangle^{(2*)}(T) &= \frac{2}{T}\int_0^{T/2} B_1(t_1)dt_1 \\ &= \left(\frac{2}{T}\right)^2 \int_{t_1=0}^{T/2}\int_{t_2=0}^{T/2} A(\mathbf{q}(t_1+t_2),\mathbf{p}(t_1+t_2))\,dt_1\,dt_2. \end{aligned} \tag{2.7}$$

Let us define the function $\alpha_{[0,\mathcal{T}]}$ by

$$\alpha_{[0,\mathcal{T}]} \ : \ t \in \mathbb{R} \ \mapsto \ 1/\mathcal{T} \ \text{ if } t \in [0,\mathcal{T}], \ \ 0 \text{ otherwise.}$$

Then $\langle A\rangle^{(2*)}(T)$ can be written in the form (2.6) with $f_T \equiv f^*_{2,T}$, where $f^*_{2,T}$ is the convolution of $\alpha_{[0,T/2]}$ with itself. So $\langle A\rangle^{(2*)}(T)$ can be understood as a filtered average and also as a mean value, over initial conditions chosen along a MD trajectory (here, the trajectory $(\mathbf{q}(t_1),\mathbf{p}(t_1))$, $t_1 \in [0,T/2]$), of standard MD averages (see the right hand side of (2.7)). More generally, denoting by $f^*_{k,T}$ the k-th convolution of the function $\alpha_{[0,T/k]}$ with itself, one can define the time average by $\langle A\rangle^{(k*)}(T) = \displaystyle\int_0^T A(\mathbf{q}(t),\mathbf{p}(t))\, f^*_{k,T}(t)\,dt$. We can prove [P2], under the assumptions collected in Appendix 2.6, that

$$\langle A\rangle^{(k*)}(T) = \langle A\rangle + O\left(\frac{1}{T^k}\right). \tag{2.8}$$

Thus, for any **arbitrary** positive integer k, it is possible to design an averaging scheme which converges to the ensemble average at the rate $1/T^k$.

The discrete version of $\langle A\rangle^{(k*)}(T)$ is

$$\langle A\rangle^{(k*)}_{num}(\delta t, T) = \frac{1}{N^k}\sum_{j=0}^{k(N-1)} C(k,N-1,j) A\left(\mathbf{q}_j,\mathbf{p}_j\right), \tag{2.9}$$

where $\mathcal{N} = kN$ is the number of time steps that have been computed, $(\mathbf{q}_j,\mathbf{p}_j)_{j=0}^{kN}$ is the numerical trajectory given by an integration scheme with time step $\delta t = T/(kN)$ and where $C(k,N-1,j)$ is the number of k-tuples of non-negative integers $j_1,\ldots,j_k$ whose sum is fixed to be j and which all belong to $[0,N-1]$. These coefficients can be computed by recursion, in such a way that their computation is almost for free in comparison with the computation of the trajectory $(\mathbf{q}_j,\mathbf{p}_j)_{j=0}^{kN}$.

However, formula (2.9) is not incremental, due to the presence of the coefficients $C(k,N-1,j)$ in the summation. In other words, the average after $k(N+1)$ steps are not simply deduced from the knowledge of that after kN steps. This difficulty leads us to use other filtering functions and to define the time average by

$$\langle A\rangle^{(k)}(T) = \int_0^T A(\mathbf{q}(t),\mathbf{p}(t))\, f_k\left(\frac{t}{T}\right) dt\,, \tag{2.10}$$

with

$$f_k(t) = \alpha_k t^{k-1}(1-t)^{k-1}, \tag{2.11}$$

where α_k is a normalizing constant ($k \geq 1$). Again, $\langle A \rangle^{(k)}(T)$ satisfies an estimate similar to (2.8). The discrete version of $\langle A \rangle^{(k)}(T)$ is

$$\langle A \rangle^{(k)}_{num}(\delta t, T) = \frac{\displaystyle\sum_{j=0}^{\mathcal{N}-1} A\left(\mathbf{q}_j, \mathbf{p}_j\right) f_k\left(\frac{j}{\mathcal{N}}\right)}{\displaystyle\sum_{j=0}^{\mathcal{N}-1} f_k\left(\frac{j}{\mathcal{N}}\right)}, \tag{2.12}$$

where $T = \mathcal{N}\delta t$. Decomposing f_k as a sum of monomial functions, one can see that the averages after $\mathcal{N}+1$ steps can be computed from the averages after $\mathcal{N}$ steps and the observable values at time step $\mathcal{N}+1$. So, just like the standard formula (2.5), formula (2.12) allows one to compute time averages "on the fly". In the test cases below, we will make use of (2.12) with several values of k to compute an approximation of the ensemble average.

Formulae (2.9) and (2.12) are generalizations of (2.5) since

$$\langle A \rangle^{(k*=1)}_{num}(\delta t, T) = \langle A \rangle^{(k=1)}_{num}(\delta t, T) = \langle A \rangle^{Rie}_{num}(\delta t, T).$$

To the best of our knowledge, formulae (2.9) or (2.12) do not seem to be used in Molecular Dynamics simulations, or at least are not reported on in the literature.

In addition, if a symplectic scheme of order r_0 is used to compute the trajectory ($r_0 = 2$ for the velocity Verlet scheme), we can prove [P2] the following error estimate :

$$\left|\langle A \rangle^{(k)}_{num}(\delta t, T) - \langle A \rangle\right| \leq C\left(\frac{1}{T^k} + (\delta t)^{r_0}\right) \tag{2.13}$$

for some constant C which does not depend on T or δt (a similar estimate holds for $\langle A \rangle^{(k*)}_{num}(\delta t, T)$).

We emphasize the fact that the polynomial decay of the error with respect to the total time of simulation T is made possible by the fact that the values $A\left(\mathbf{q}_j, \mathbf{p}_j\right)$ are correlated in time, in the sense that they are obtained by the simulation of one single trajectory. When one uses a Monte Carlo method, the values $A\left(\mathbf{q}_j, \mathbf{p}_j\right)$ are not correlated any longer, and the error between the computed empirical average and the ensemble average $\langle A \rangle$ only decays at the universal rate $1/\sqrt{T}$.

The proof of (2.13), together with details on the filter function used in formula (2.12), may be read in a companion article [P2].

2.3 Numerical examples

We now turn to numerical examples showing that :
- estimate (2.13) is sharp for integrable Hamiltonian systems satisfying (at least some of) the assumptions under which the estimate has been rigorously proved.
- in some particular conditions, estimate (2.13) seems to also hold for some non-integrable systems of practical interest, although our proof does not carry through such non-integrable systems.

We first present results in the case of a collection of independent harmonic oscillators (see Sec. 2.3.1), then on the Kepler problem (see Sec. 2.3.2). Both systems are integrable, and fall within the hypotheses of our theoretical result. In Sec. 2.3.3, we study a system of particles subjected to the Lennard-Jones potential, and in Sec. 2.3.4, we study alkane chains described by the "united atom" (UA) model [47]. Finally, in Sec. 2.3.5, we study a particle submitted to a double well potential in 2D, in the case when the particle energy is lower than or much larger than the saddle point energy (the study of the other cases is postponed to Sec. 2.4). In all these cases, time averages computed by (2.5) converge to ensemble averages at rate $1/T$, and we show the efficiency of the method (2.12) that we propose.

With a view to showing that our approach is insensitive to the numerical integration scheme, provided that it is symplectic, we have used, depending on the case at hand, three different algorithms [9], of different orders, to integrate the dynamical system (2.1) :
- the standard velocity Verlet algorithm, which is, as recalled above, symplectic and of order 2; we denote by $\phi_{\delta t}^{VV}$ a step of the algorithm of time step δt : $(\mathbf{q}_{n+1}, \mathbf{p}_{n+1}) = \phi_{\delta t}^{VV}(\mathbf{q}_n, \mathbf{p}_n)$;
- a composition of the velocity Verlet algorithm with itself so that the order of the algorithm increases to 4; namely,

$$(\mathbf{q}_{n+1}, \mathbf{p}_{n+1}) = \phi_{\delta t}^{(4)}(\mathbf{q}_n, \mathbf{p}_n) = \phi_{b_1\delta t}^{VV} \circ \phi_{b_0\delta t}^{VV} \circ \phi_{b_1\delta t}^{VV}(\mathbf{q}_n, \mathbf{p}_n)$$

  with $b_1 = 1/(2 - 2^{1/3})$ and $b_0 = 1 - 2b_1$.
- a composition of the velocity Verlet algorithm with itself so that the order of the algorithm increases to 6; namely,

$$\begin{aligned}(\mathbf{q}_{n+1}, \mathbf{p}_{n+1}) &= \phi_{\delta t}^{(6)}(\mathbf{q}_n, \mathbf{p}_n) \\ &= \phi_{c_3\delta t}^{VV} \circ \phi_{c_2\delta t}^{VV} \circ \phi_{c_1\delta t}^{VV} \circ \phi_{c_0\delta t}^{VV} \circ \phi_{c_1\delta t}^{VV} \circ \phi_{c_2\delta t}^{VV} \circ \phi_{c_3\delta t}^{VV}(\mathbf{q}_n, \mathbf{p}_n)\end{aligned}$$

  with $c_1 = -1.17767998417887$, $c_2 = 0.235573213359357$, $c_3 = 0.784513610477560$ and $c_0 = 1 - 2(c_1 + c_2 + c_3)$.

2.3.1 Collection of harmonic oscillators

In this section, we consider the dynamics of a system of five particles in 1D described by the Hamiltonian

$$H = \sum_{i=1}^{5} \frac{p_i^2}{2} + \omega_i^2 \frac{q_i^2}{2},$$

with the following frequencies :

$$\omega_1 = 1.0,\ \omega_2 = 8.01256,\ \omega_3 = 12.25245,\ \omega_4 = 17.234,\ \omega_5 = 20.98765. \qquad (2.14)$$

As an instance, we study the average of $A = p_4^2$ with the algorithm of order 6. The exact value $\langle A \rangle = 0.5$ of the ensemble average can be obtained by analytical calculations.

All but one hypothese needed to prove our estimates are satisfied : the integration algorithm is symplectic and the analyticity and integrability assumptions on the Hamiltonian function (see Appendix 2.6.1) are satisfied. The frequencies (2.14) do not satisfy the diophantine assumption (see Appendix 2.6.2) for they are rational numbers, but the level set (defined by (2.2)) of the invariant functions $I_j(\mathbf{q}, \mathbf{p}) = \frac{p_j^2}{2} + \omega_j^2 \frac{q_j^2}{2}$ is well-enough sampled, in comparison to the precision needed for the computation of $\langle A \rangle$.

We use formula (2.12) with different values of k ($k = 1$ to 4), for a given δt ($\delta t = 0.025$). The numerical error as a function of the length of the MD trajectory is displayed in Fig. 2.1. A least-square fit provides the following result : the error $\left| \langle A \rangle_{num}^{(k)}(\delta t, T) - \langle A \rangle \right|$ decreases with rate $1/T^{\tilde{k}}$, with

$$\tilde{k}(k=1) = 0.978, \quad \tilde{k}(k=2) = 1.91, \quad \tilde{k}(k=3) = 3.05, \quad \tilde{k}(k=4) = 3.87.$$

So, the time average defined by (2.12) converges to the ensemble average with a rate very close to $1/T^k$.

2.3.2 The Kepler problem

We next turn to the simulation of the Kepler problem. We consider the Hamiltonian

$$H = \frac{\mathbf{p}^2}{2} - \frac{1}{|\mathbf{q}|},$$

describing a 3D particle in an attractive Coulomb potential.

Let us for instance study the average of $A = |\mathbf{q}|$. As in the previous example, the ensemble average can be computed analytically :

$$\langle |\mathbf{q}| \rangle = \left(3 + 2\, E_0 L_0^2\right) / (4\, |E_0|),$$

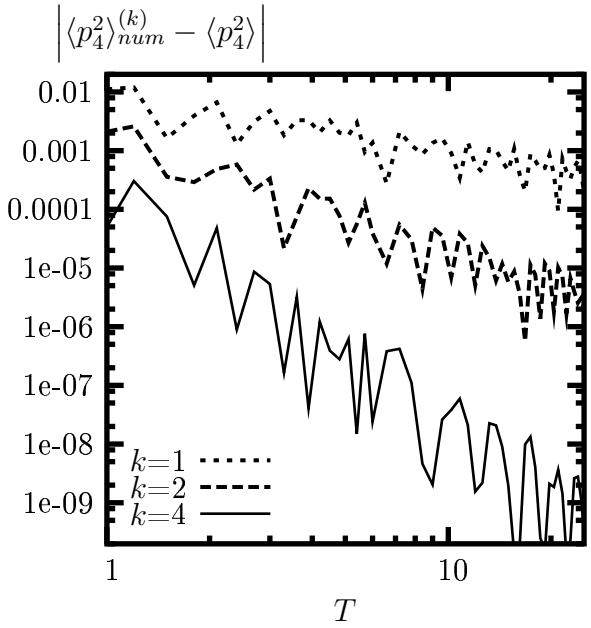

FIG. 2.1 – Convergence of $\langle p_4^2\rangle_{num}^{(k)}(\delta t, T)$ to $\langle p_4^2\rangle$ for different values of k, in the case of a collection of 5 harmonic oscillators. Initial conditions are $q_1 = 4$, $p_1 = -1.2$, $q_i = 0$ and $p_i = 1$ for $i = 2, \ldots, 5$.

where E_0 is the initial (negative) energy and L_0 is the norm of the initial kinetic momentum. The simulation time step is $\delta t = 0.01$, and we use the algorithm of order 6. As for the harmonic oscillators, the Hamiltonian function is analytic and integrable and the integration scheme is symplectic.

We compute the averages (2.12) with k ranging from 1 to 4. We see that, for $T \geq 40$ (approximately beyond 4 revolution periods), the larger k, the larger the convergence rate (see Fig. 2.2; the convergence rate is quantitatively studied below). Remark that, as k increases, the amplitude of the oscillations at the beginning of the simulation (which is not the focus here) increases (see Fig. 2.3). This fact seems to be generic, and not to be restricted to the Kepler problem.

Let us now check that the estimate (2.13) is sharp. For this purpose, we study the dependence of the error $e^{(k)}(\delta t, T)$ with respect to T and δt in order to check whether there exist constants $C_1(k)$ and $C_2(r_0)$ such that

$$e^{(k)}(\delta t, T) = \left|\langle A\rangle_{num}^{(k)}(\delta t, T) - \langle A\rangle\right| \approx \frac{C_1(k)}{T^k} + C_2(r_0)(\delta t)^{r_0}.$$

The initial conditions are such that $\langle A\rangle = 1.37$. We first determine $C_2(r_0)$. To do this, we choose several values of the time step δt in the interval [0.025 ;0.2], and for each of them, we compute a long-time trajectory. We estimate that the trajectory

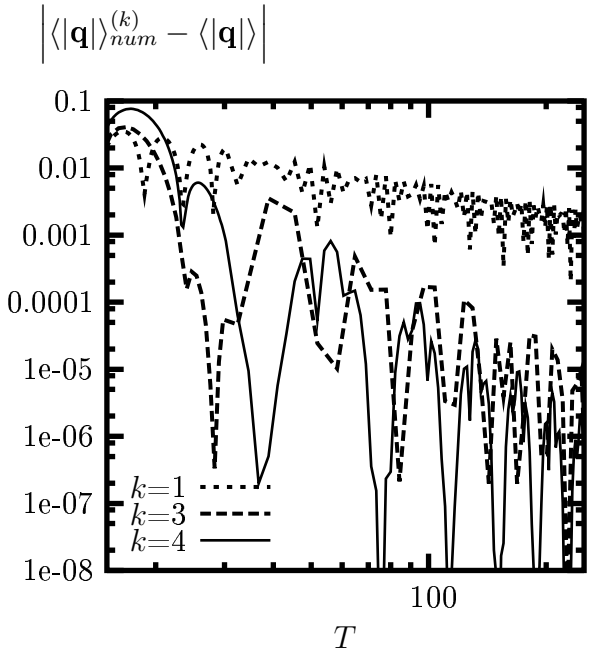

FIG. 2.2 – Convergence of $\langle|\mathbf{q}|\rangle_{num}^{(k)}(\delta t, T)$ to $\langle|\mathbf{q}|\rangle$ for different values of k, in the case of the Kepler problem. Initial conditions are $\mathbf{q} = (0.9, 0, 0)$ and $\mathbf{p} = (0, 1.1, 0.5)$.

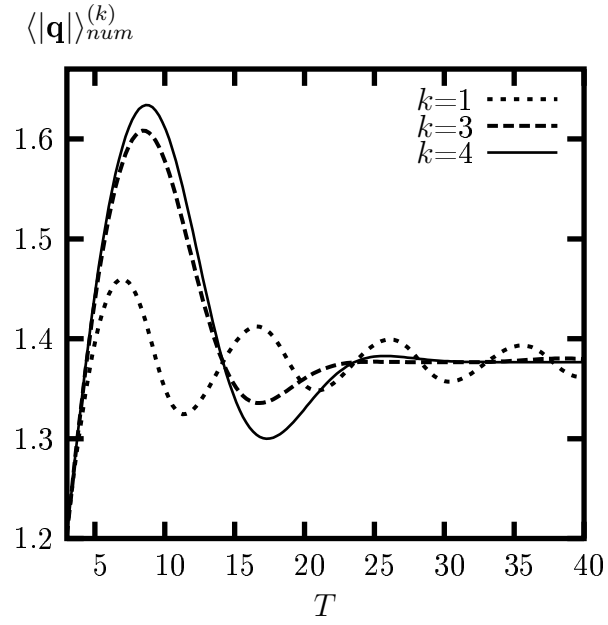

FIG. 2.3 – Oscillations of $\langle|\mathbf{q}|\rangle_{num}^{(k)}(\delta t, T)$ at the beginning of the simulation (Kepler problem), for different values of k. Initial conditions are $\mathbf{q} = (0.9, 0, 0)$ and $\mathbf{p} = (0, 1.1, 0.5)$.

is long enough when the oscillations of the function $T \mapsto \langle A \rangle_{num}^{(k)}(\delta t, T)$ are both smaller than 10^{-8} and more than 100 times as small as the difference between $\langle A \rangle$ and $\langle A \rangle_{num}^{(k)}(\delta t, T)$. So the leading term of the error is $C_2(r_0)(\delta t)^{r_0}$. We observe indeed that $\ln e^{(k)}(\delta t, T)$ is a linear function of $\ln \delta t$ of slope $-r_0$, and find values for $C_2(r_0)$ which **do not depend on** k and are given by

$$C_2(r_0 = 6) = 0.163, \quad C_2(r_0 = 4) = 1.09, \quad C_2(r_0 = 2) = 0.622.$$

In order to determine the values of $C_1(k)$, we work with very small values of δt (respectively $\delta t = 10^{-2}$, 5.10^{-3} and 5.10^{-4} when we use the algorithm of order 6, 4 and 2) so that the term $C_2(r_0)(\delta t)^{r_0}$ is lower than 10^{-7}. We also observe that the time average converges to the exact ensemble average up to an error of 10^{-7}. So, as long as the error is larger than 10^{-7}, its leading term is $C_1(k)/T^k$. A least-square fit of $\ln e^{(k)}(\delta t, T)$ as a function of $\ln T$ shows that the error decreases with rate $1/T^{\tilde{k}}$, with

$$\tilde{k}(k = 1) = 1.02, \quad \tilde{k}(k = 2) = 2.01, \quad \tilde{k}(k = 3) = 2.82, \quad \tilde{k}(k = 4) = 3.85.$$

Likewise, we observe that the rescaled error $T^k \left| \langle A \rangle_{num}^{(k)}(\delta t, T) - \langle A \rangle \right|$ is bounded independently of T, and we also notice that the bound is independent of r_0. We obtain

$$C_1(k = 1) \approx 0.6, \quad C_1(k = 2) \approx 14.4, \quad C_1(k = 3) \approx 212, \quad C_1(k = 4) \approx 8270.$$

Not surprisingly, the values of $C_1(k)$ quickly increase with k (this is in agreement with the fact that oscillations at the beginning of the simulation are larger for larger values of k).

These numerical experiments therefore confirm the sharpness of the estimate (2.13).

2.3.3 Particles in a truncated Lennard-Jones potential

We study here a system of $M = 288$ particles in 3D interacting through a truncated Lennard-Jones potential. The Hamiltonian of the system reads

$$H = \sum_{i=1}^{M} \frac{\mathbf{p}_i^2}{2} + \sum_{i=1}^{M} \sum_{j>i} V\left(|\mathbf{q}_j - \mathbf{q}_i|\right),$$

with $V_{LJ}(z) = \dfrac{4}{z^{12}} - \dfrac{4}{z^6}$ and $V(z) = V_{LJ}(z) - V_{LJ}(z_c) - V'_{LJ}(z_c)(z - z_c)$ if $z \leq z_c$, 0 otherwise. Thus, both the potential and the forces are continuous at the cutoff radius $z_c = 3.08$. We apply periodic boundary conditions (the simulation box is $[-3.2; 3.2]^3$) and use the minimum image convention [2]. The initial conditions correspond to a solid state and are such that the density of the system is 1.12 and the average kinetic

temperature is 0.884. Such an initial condition ensures that the system remains in the solid phase. The hypotheses of Appendix 2.6 are **not satisfied** (in particular, there are not enough invariant functions). The observable under examination is the pressure :

$$A(\mathbf{q},\mathbf{p}) = P = \frac{1}{3V}\sum_{i=1}^{M}\left(\mathbf{p}_i^2 + \sum_{j>i}\mathbf{q}_{ij}\cdot\mathbf{F}_{ij}\right),$$

where $\mathbf{q}_{ij} = \mathbf{q}_i - \mathbf{q}_j$ and $\mathbf{F}_{ij} = -\dfrac{\partial H}{\partial \mathbf{q}_{ij}}$. We work with the velocity Verlet algorithm, using a time step of $\delta t = 5.10^{-3}$.

The results are displayed in Figs. 2.4 and 2.5. We have checked that the error $\left|\langle A\rangle_{num}^{(k)}(\delta t, T) - \langle A\rangle\right|$ decreases with the rate $1/T^{\tilde{k}}$, with

$$\tilde{k}(k=1) = 0.996, \quad \tilde{k}(k=2) = 1.85, \quad \tilde{k}(k=3) = 2.67, \quad \tilde{k}(k=4) = 3.76.$$

So, as in the previous examples, the rate of convergence is improved from $1/T$ to almost $1/T^k$, even if the system at hand is not integrable. Indeed, even if the system remains in the solid phase, the potential cannot be approximated by a harmonic potential : along the trajectory of the system, we have observed that some eigenvalues of the Hessian of the potential are negative, and that some other ones vary by 20%.

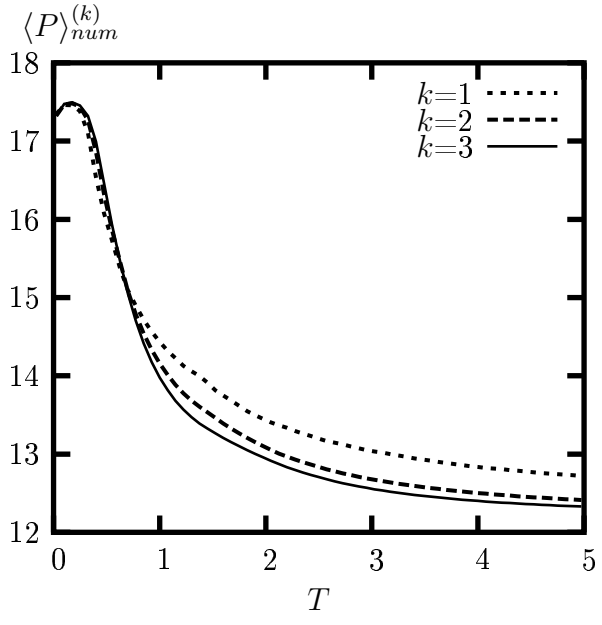

FIG. 2.4 – Average pressure $\langle P\rangle_{num}^{(k)}(\delta t, T)$ at the beginning of the simulation of the Lennard-Jones system, for different values of k.

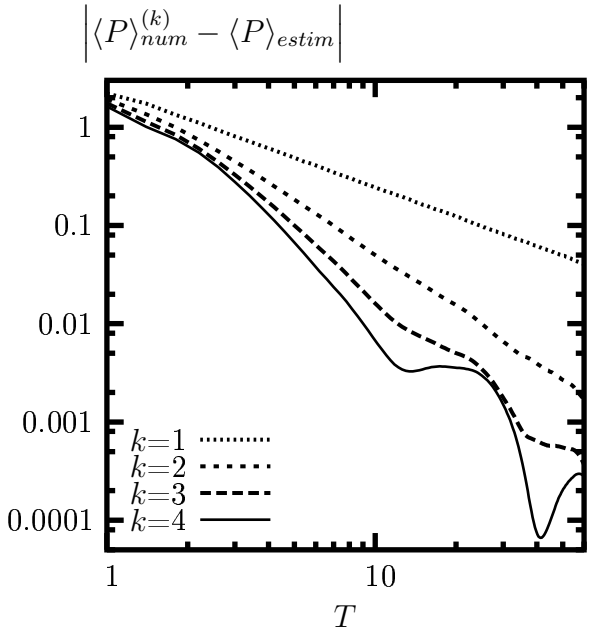

FIG. 2.5 – Convergence of the average pressure $\langle P\rangle_{num}^{(k)}(\delta t, T)$ to $\langle P\rangle_{estim}$ for different values of k, in the case of the Lennard-Jones system.

2.3.4 Alkane chains

We now study chains of particles interacting through the Ryckaert- Bellemans united atom (UA) model [47]. These chains represent alkane molecules. The observable under examination is the normalized end-to-end distance

$$A(\mathbf{q}) = \frac{|\mathbf{q}_1 - \mathbf{q}_M|^2}{M d_{C-C}^2},$$

where $d_{C-C} = 1.53$ Å is the constant bond length. We work with the velocity Verlet algorithm. As in the previous example, the hypotheses of Appendix 2.6 are not satisfied.

The first example is a chain of $M = 40$ particles. For an extremely low temperature (T= 2.13 K), we obtain the convergence of the time averages (see Fig. 2.6, obtained with a time step of $\delta t = 0.1$ fs) with the rate $1/T^{\tilde{k}}$, with the following values for $\tilde{k}$:

$$\tilde{k}(k=1) = 0.98, \quad \tilde{k}(k=2) = 1.70, \quad \tilde{k}(k=3) = 3.02, \quad \tilde{k}(k=4) = 4.21.$$

So we again observe convergence at a rate close to $1/T^k$. In Fig. 2.7, we display the results obtained with a chain of $M = 100$ particles in 3D : results on this larger system are similar to those for the 40 particles chain.

The case of higher temperatures will be studied in Sec. 2.4.

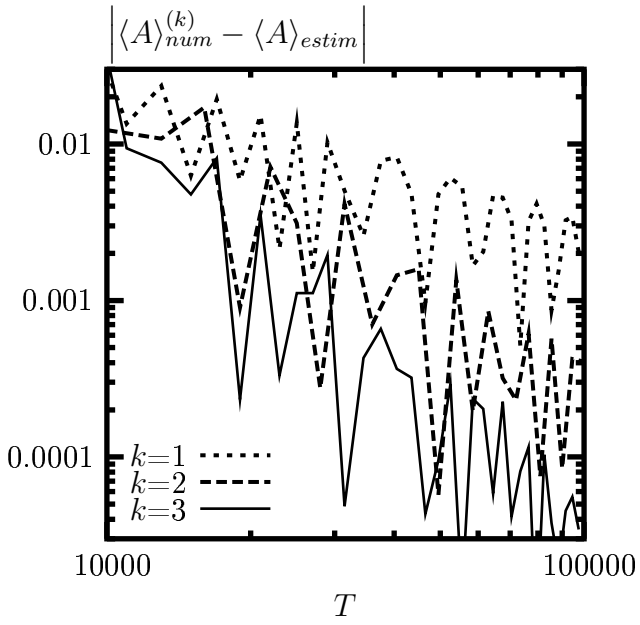

FIG. 2.6 – Convergence of the end-to-end distance of an alkane chain ($M = 40$ particles) for different values of k, in the low temperature case ($T = 2.13$ K).

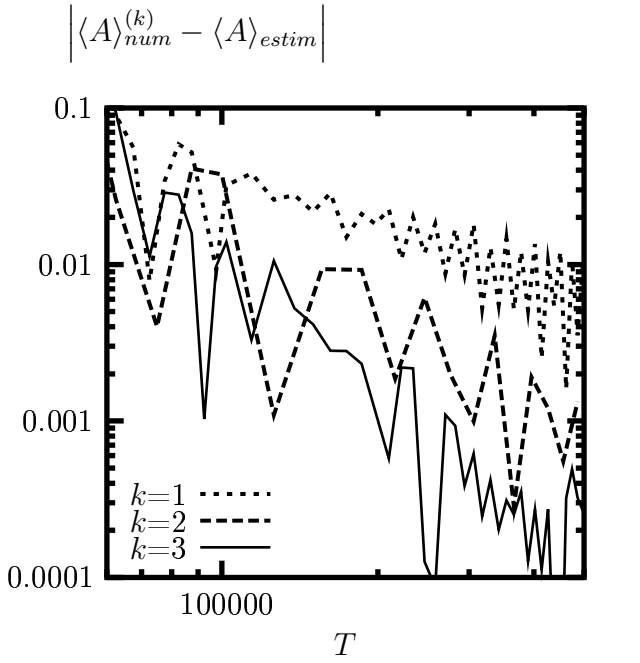

FIG. 2.7 – Convergence of the end-to-end distance of an alkane chain ($M = 100$ particles) for different values of k, in the low temperature case.

2.3.5 One particle in a double well potential

In this section, we deal with the Hamiltonian

$$H = \frac{p_x^2}{2} + \frac{p_y^2}{2} + V(q_x, q_y),$$

where

$$V(q_x, q_y) = (q_x^2 - 1)^2 + (q_y + q_x^2 - 1)^2,$$

describing a 2D particle in a double well potential. The potential V goes to $+\infty$ at infinity and has three critical points : two global minima located at $(\pm 1, 0)$, at which $V(\pm 1, 0) = 0$, and one saddle point located at $(0, 1)$, at which $V(0, 1) = 1$.

The behaviour of time averages strongly depends on the value of the particle energy E_0 with respect to the saddle point energy (here equal to 1). Three different regimes can be identified :

1. for E_0 smaller than 1, the trajectory only explores a single basin. Time averages converge to a limit and using the filtering function f_k and formula (2.12) improves the convergence rate from $1/T$ to $1/T^k$ (for an arbitrary k) as in the previous examples. These results hold even if the energy is large enough (for instance, $E_0 = 0.8$) so that, on the region sampled by the particle, the potential cannot be approximated by a harmonic potential ;
2. for E_0 much larger than 1 (say $E_0 \geq 5$), the particle is so energetic that it does not really "feel" the barrier and the convergence of the time averages is similar to the one observed in the first case ; once again, we have checked that we are far from the harmonic approximation ;
3. for E_0 a little larger than 1, the system really "feels" the presence of two distinct basins ; we postpone the study of this case to Sec. 2.4.

Let us end this section by a discussion on the ergodicity of the 2D double well system under study. It is easy to check that the isoenergetic surface $H(q_x, q_y, p_x, p_y) = E_0$ can be parametrized by three angles :

$$\begin{array}{rclcrcl} q_x & = & \pm\sqrt{1 + \sqrt{E_0}\cos(\theta)}, & & q_y & = & \sqrt{E_0}\ (\sin(\theta)\cos(\phi) - \cos(\theta)), \\ p_x & = & \sqrt{2E_0}\ \sin(\theta)\sin(\phi)\cos(\psi), & & p_y & = & \sqrt{2E_0}\ \sin(\theta)\sin(\phi)\sin(\psi), \end{array}$$

with $(\theta, \phi, \psi) \in [0, \theta_M] \times [0, \pi] \times [0, 2\pi]$, where $\theta_M = \pi$ if $E_0 \leq 1$ and $\theta_M = \arccos(-1/\sqrt{E_0})$ if $E_0 > 1$. It is therefore possible to compute numerically the NVE ensemble average of any (regular) observable with a high accuracy. It reads

$$\langle A \rangle(E_0) = \int_{[0,\theta_M]\times[0,\pi]\times[0,2\pi]} A\,(\theta, \phi, \psi)\ w(\theta, \phi)\ d\theta\, d\phi\, d\psi, \tag{2.15}$$

where $w(\theta, \phi) = C\,\dfrac{\sin(\phi)\sin(\theta)^2}{\sqrt{1 + \sqrt{E_0}\cos(\theta)}}$ is the relevant invariant probability density of the dynamical system on (θ, ϕ, ψ) (C is a normalizing constant). In Table 2.1, we

compare this ensemble average with several time averages computed from trajectories with various initial conditions that all correspond to the energy $E_0 = 0.5$. In general, these time averages are different from the ensemble average. This difference can be explained either by the fact that the isoenergetic surface has invariant subsets (this may be due to the existence of a second invariant that we did not manage to identify), or by the fact that the system is ergodic but on very long time scales, much longer than the simulation times we can afford.

A	$\langle A \rangle (E_0 = 0.5)$	$\langle A \rangle_{num}^{(k=5)}(\mathbf{q}_1, \mathbf{p}_1)$	$\langle A \rangle_{num}^{(k=5)}(\mathbf{q}_2, \mathbf{p}_2)$	$\langle A \rangle_{num}^{(k=5)}(\mathbf{q}_3, \mathbf{p}_3)$
q_x	-0.94459	-0.93118	-0.95585	-0.92386
q_x^2	0.92843	0.90077	0.95170	0.88565
q_x^4	0.98964	0.92855	1.04074	0.89565
q_y	0.071562	0.099221	0.048292	0.114346
q_y^2	0.25517	0.14298	0.34725	0.085192
p_x^2	0.24482	0.33949	0.16794	0.38717
p_y^2	0.24482	0.14615	0.32574	0.095278
$V(q_x, q_y)$	0.25517	0.25717	0.25315	0.25878

TAB. 2.1 – Computed values for the ensemble average and the time average of different observables A (double well potential). Ensemble averages are computed from (2.15). For the time averages, computed from (2.12), initial conditions are $(\mathbf{q}_1, \mathbf{p}_1) = (-1, 0.5; 0.5, -0.5)$, $(\mathbf{q}_2, \mathbf{p}_2) = (-0.9, 0.772\ldots; 0, -0.5)$ and $(\mathbf{q}_3, \mathbf{p}_3) = (-1.05, 0.549; 0.3606\ldots, 0)$, corresponding to the energy 0.5, and the time step is $\delta t = 0.005$. We checked that a simulation of length $T = 5000$ was long enough and that time average values do not depend on k.

From a practical point of view, one way to solve this issue is to compute the mean value of the time averages over several different initial conditions. First, as in a Monte Carlo method, we choose many initial conditions $(\mathbf{q}_i, \mathbf{p}_i)_i$ on the surface of constant energy E_0, according to the probability density $w(\theta, \phi)$ (to do so, one can use the Acceptance-Rejection method [14]). For all these initial conditions, we compute a MD trajectory and a time average, and then the mean value $\mu(E_0)$ and the standard deviation $\sigma_{MC+MD}(E_0)$ of these data. Results are reported in Fig. 2.8 : this method yields a good approximation of the ensemble average.

Thus, to compute an ensemble average, we can suggest to choose several different initial conditions on the constant energy surface, and to use each of them to compute a trajectory and a time average. A good approximation of the ensemble average is the mean value of these time averages.

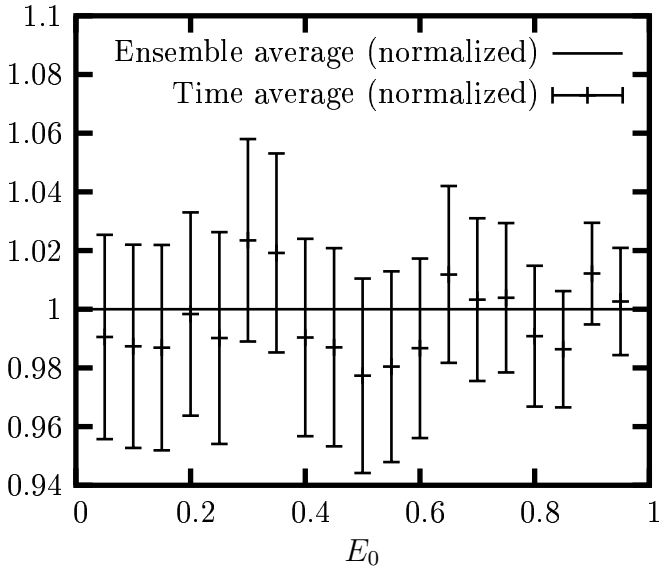

FIG. 2.8 – Comparison between the ensemble average of $A = q_y^2$, computed from (2.15), and the expected time average (error bars correspond to a 95% confidence interval; 1000 initial conditions have been used), on the double well potential, for different energy levels E_0. All quantities have been normalized by the ensemble average $\langle q_y^2 \rangle(E_0)$.

2.4 The case of systems with multiple metastable states

Alkane chains and the double well example are systems presenting several metastable states. In Sec. 2.3, we have studied them in the regimes of either low or high energy (see Secs. 2.3.4 and 2.3.5). In such regimes, either a unique well is explored (because the system does not have enough energy to visit other wells), or the system is so energetic that it does not really "feel" the barriers. We now focus on the case of an intermediate energy : transitions between metastable states are possible but they are rare events in comparison with the faster characteristic time scale of the system. In this regime, we show that time averages converge to ensemble averages at the slower rate $1/\sqrt{T}$.

2.4.1 One particle in a double well potential

In this section, we simulate the same system as in Sec. 2.3.5, but we now choose to work with an energy $E_0 = 1.25$ that is slightly larger than the saddle point energy (which is equal to 1). In this case, one does not observe any convergence of the time averages, even for simulation times 10,000 times as large as those which lead to convergence in the cases of low or high energy (see Fig. 2.9).

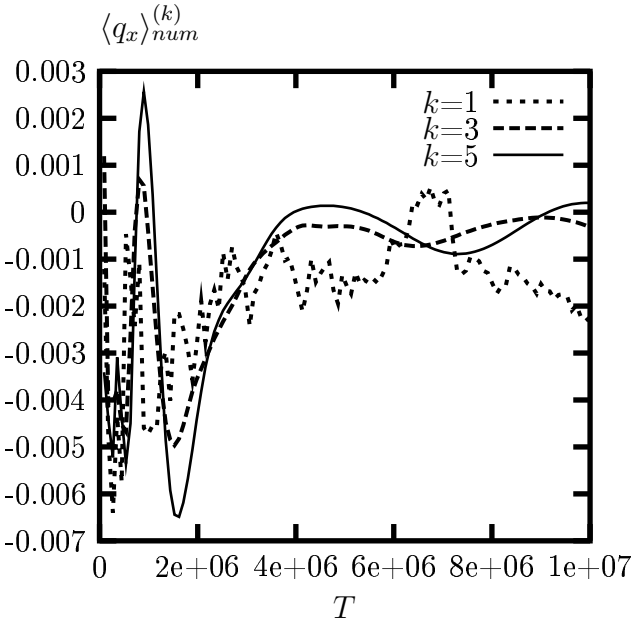

FIG. 2.9 – Evolution of $\langle q_x \rangle_{num}^{(k)}(\delta t, T)$ as a function of T (double well potential, energy of 1.25 close to the barrier energy), for a time step $\delta t = 0.05$ and 2.10^8 time steps.

Actually, as shown on Fig. 2.10, the particle spends "a long time" in one basin, then quickly undergoes a transition into the other basin, in which it spends another "long time", and so on. In this simple example, the basin locations are known and the mean exit time can be easily estimated from MD simulations. These data can be used to parametrize a two-state Markov chain model [116] mimicking the transition between the two basins. A comparison between the convergence rate of time averages (2.12) computed by MD on the one hand and by the so-obtained Markov chain model on the other hand, for the observable q_x (the average of which is zero due to the symmetry of the potential), is reported in Fig. 2.11. The good quantitative agreement between these two convergence rates confirms that the bottleneck in the computation of averages by brute force MD simulations is the presence of several well-separated basins ; in this case, the convergence rate is no longer of $1/T$ or better, but falls down to $1/\sqrt{T}$, at least in the range of accessible simulation times to date.

2.4.2 Alkane chains

In this section, we simulate the same chains of particles as in Sec. 2.3.4 (with $M = 40$ particles), but we now work at higher temperature.

We first simulate the chain at an energy such that the temperature is $T = 135$ K, and study the end-to-end distance. Although the simulation time is 400 times longer that in the case studied in Sec. 2.3.4, time averages have not yet reached their

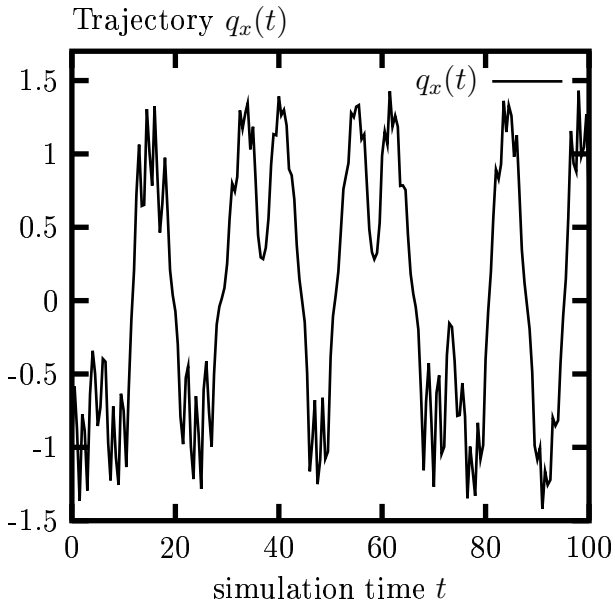

FIG. 2.10 – Evolution of $q_x(t)$ as a function of time t (double well potential, energy of 1.25 close to the barrier energy, time step $\delta t = 0.01$).

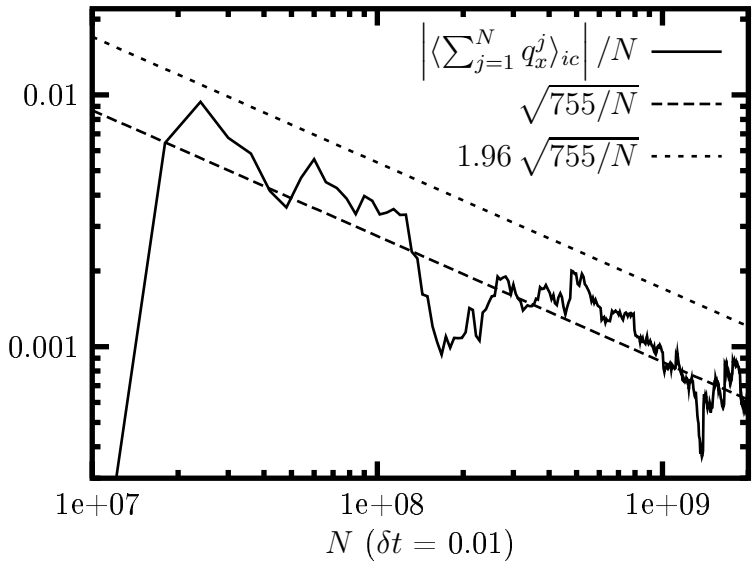

FIG. 2.11 – Estimation of the convergence of $\frac{1}{N}\sum_{j=1}^{N} q_x^j$ to 0 with respect to N with a Markov chain model (double well potential, energy of 1.25 close to the barrier energy, $\delta t = 0.01$; $\langle\cdot\rangle_{ic}$ denotes average over 16 initial conditions).

infinite time limit (see Fig. 2.12). We study in Fig. 2.13 their convergence rate : it is very close to $1/\sqrt{T}$.

The reason for this slow convergence is the same as in the double well potential case : the system has enough energy to explore many wells, but it spends a long time in each well (in comparison with the time needed to go from one well to another one). Thus, once again, time average convergence rate is slowed down to $1/\sqrt{T}$.

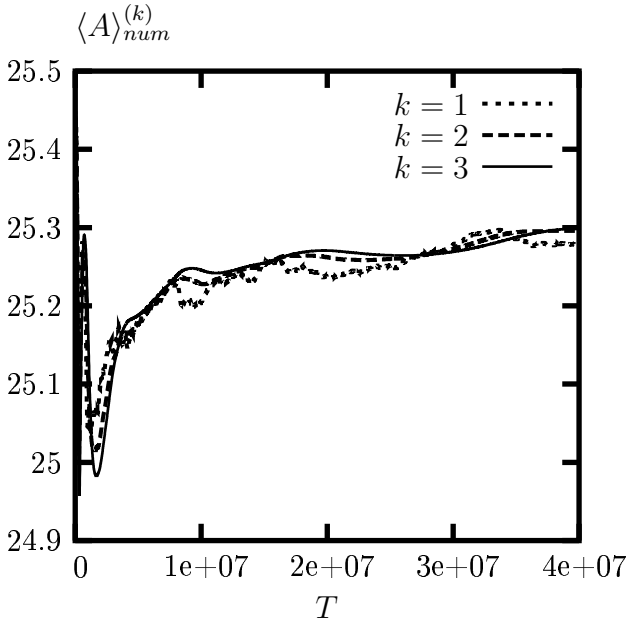

FIG. 2.12 – Evolution of the end-to-end distance time average as a function of T (alkane chain with $M = 40$ particles, temperature of $T = 135$ K).

Let us now simulate the chain at a temperature somewhat higher ($T = 4.16$ K) than that in the case studied in Sec. 2.3.4 (where $T = 2.13$ K), but still very low. The situation is exactly the same as that in the case $T = 135$ K : neither time averages computed by (2.5) nor by (2.12) converge (see Fig. 2.14).

2.5 Conclusions

In order to compute ensemble averages of observables, one can use Molecular Dynamics and compute a time average of these observables over one (or several) trajectory of a dynamical system. When the ergodic assumption is satisfied, this time average converges, as the trajectory length goes to infinity, towards the ensemble average. We have presented a numerical method which allows, when this convergence occurs at the rate $1/T$, to speed up the convergence up to $1/T^k$ (k arbitrary) so that a trajectory of reduced length (and thus needing less computational effort) is sufficient

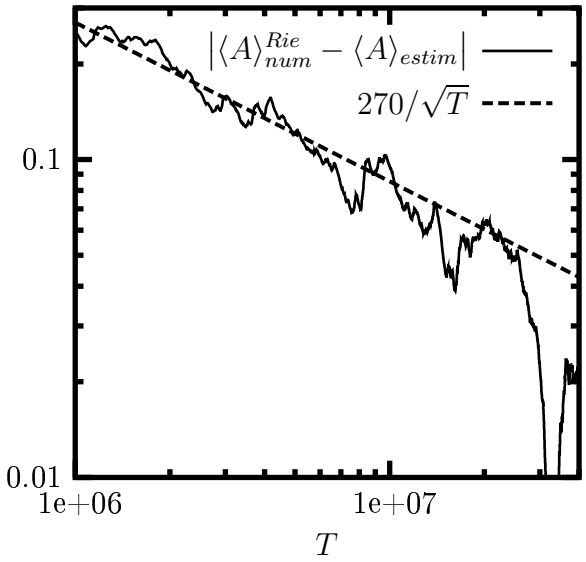

FIG. 2.13 – Estimation of the convergence rate of the averaged end-to-end distance (alkane chain with $M = 40$ particles, temperature of $T = 135$ K).

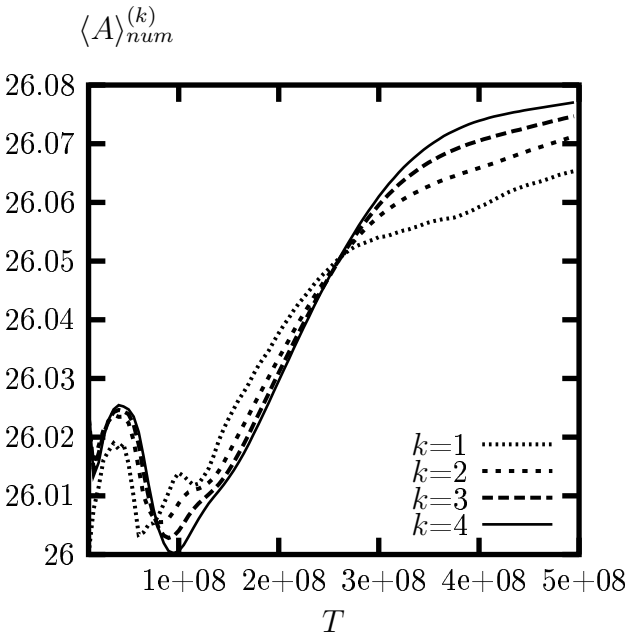

FIG. 2.14 – Evolution of the end-to-end distance time average as a function of T (alkane chain with $M = 40$ particles, temperature of $T = 4.16$ K).

to achieve a given precision. Together with this numerical method, we have presented some error bounds that may be used to optimize the simulation parameters. We have demonstrated by test cases the sharpness of these bounds. However, a limitation of our method must be pointed out.

A typical problem in Molecular Dynamics is the exploration of the whole phase space when the potential energy of the system has several basins corresponding to distinct metastable states, and when the energy is such that transitions between these states are possible but are rare events. In this case, the convergence rate of time averages is very slow (it is close to $1/\sqrt{T}$). The method we have presented in this paper does not solve this issue (it does not improve the phase space exploration, but the computation of the time average). Other methods have then to be used, for instance those suggested in Refs. [109, 111, 116, 117, 120], and maybe combined with our method.

Acknowledgments

We would like to thank Gilles Zérah for his very helpful suggestions and Christophe Chipot for stimulating discussions. We also gratefully acknowledge the financial support of INRIA through the contract grant "Action de Recherche Coopérative PRESTISSIMO".

2.6 Appendix

In this appendix, we detail the hypotheses needed to prove estimates (2.4) and (2.8). As we have mentioned it in Sec. 2.2, we consider a system of M particles in 3D, described by the Hamiltonian $H(\mathbf{q}, \mathbf{p})$. Let $(\mathbf{q}_0, \mathbf{p}_0)$ be the initial condition.

2.6.1 Assumptions on the Hamiltonian function H :

We suppose that the Hamiltonian H is an analytic function, and that it is *completely integrable*, that is to say that (see Ref. [3], pp 214 and 272) :

- there exist $3M$ invariant functions $I_j(\mathbf{q}, \mathbf{p})$, $j = 1, \ldots, 3M$, of the dynamical system (2.1) (the definition of an invariant function is recalled in Sec. 2.2) ; we denote by $S(\mathbf{q}, \mathbf{p})$ the level set of the invariant functions $\{I_j\}_{1 \leq j \leq 3M}$ (see (2.2)) ;
- these invariants I_j are in involution, i.e. satisfy the condition

$$\nabla_{\mathbf{q}} I_{j_1}(\mathbf{q}, \mathbf{p}) \cdot \nabla_{\mathbf{p}} I_{j_2}(\mathbf{q}, \mathbf{p}) = \nabla_{\mathbf{p}} I_{j_1}(\mathbf{q}, \mathbf{p}) \cdot \nabla_{\mathbf{q}} I_{j_2}(\mathbf{q}, \mathbf{p})$$

  for all $(\mathbf{q}, \mathbf{p}) \in \mathbb{R}^{3M} \times \mathbb{R}^{3M}$ and all j_1 and j_2, where $\nabla_{\mathbf{q}}$ is the gradient with respect to the position variables and $\nabla_{\mathbf{p}}$ is the gradient with respect to the momentum variables ;
- there exists a neighbourhood of the initial condition $(\mathbf{q}_0, \mathbf{p}_0)$ such that, for all $(\mathbf{q}, \mathbf{p})$ in this neighbourhood, (i) the level set $S(\mathbf{q}, \mathbf{p})$ is compact and connec-

ted [17], and (ii) the gradients ∇I_j of the invariant functions are linearly independent.

2.6.2 Diophantine assumption :

Let $\mathbb{T}$ be the torus $\mathbb{R}/2\pi\mathbb{Z}$. Under the above hypotheses, it is possible [3,9] to find a bounded open set $B \subset \mathbb{R}^{3M}$ containing $(I_1(\mathbf{q}_0, \mathbf{p}_0), \ldots, I_{3M}(\mathbf{q}_0, \mathbf{p}_0))$, a bounded open set $B' \subset \mathbb{R}^{3M} \times \mathbb{R}^{3M}$ containing the initial condition $(\mathbf{q}_0, \mathbf{p}_0)$ and a local change of variables ψ,

$$\psi : (\mathbf{a}, \theta) \in B \times \mathbb{T}^{3M} \quad \mapsto \quad (\mathbf{q}, \mathbf{p}) \in B' \subset \mathbb{R}^{3M} \times \mathbb{R}^{3M},$$

such that the new variables $(\mathbf{a}(t), \theta(t)) = \psi^{-1}(\mathbf{q}(t), \mathbf{p}(t))$ obey the simple dynamics

$$\frac{d\mathbf{a}}{dt} = 0, \qquad \frac{d\theta}{dt} = \omega(\mathbf{a}(t)), \tag{2.16}$$

supplied with initial condition $(\mathbf{a}(0), \theta(0)) = (\mathbf{a}_0, \theta_0) = \psi^{-1}(\mathbf{q}_0, \mathbf{p}_0)$. In (2.16), ω is a function from $\mathbb{R}^{3M}$ to $\mathbb{R}^{3M}$. The trajectory in the new variables is $\mathbf{a}(t) = \mathbf{a}_0$, $\theta(t) = \theta_0 + \omega(\mathbf{a}_0)t$.

We then make the following additional assumption (the so-called diophantine assumption) on the vector $\omega(\mathbf{a}_0)$: there exist a constant $C_0 > 0$ and an exponent $\gamma_0 > 0$ such that

$$\forall \alpha \in \mathbb{Z}^{3M} \setminus \{0\}, \quad |\alpha \cdot \omega(\mathbf{a}_0)| \ \geq \ \frac{C_0}{|\alpha|^{\gamma_0}}, \tag{2.17}$$

where we note $|\alpha| = \alpha_1 + \ldots + \alpha_{3M}$ and $\alpha \cdot \omega(\mathbf{a}_0) = \alpha_1\omega_1(\mathbf{a}_0) + \ldots + \alpha_{3M}\omega_{3M}(\mathbf{a}_0)$. This assumption means that the quantity $\alpha \cdot \omega(\mathbf{a}_0)$ can get close to zero only if $|\alpha|$ goes to infinity.

Under the assumption that the Hamiltonian function H is analytic and integrable, and the diophantine assumption, one can prove that time averages converge to ensemble averages at the rate $1/T$, and that this rate can be improved up to $1/T^k$ by using filtering functions. In addition, if the numerical scheme used to integrate the equations of motion is symplectic, and if a non-resonance condition is added to the diophantine assumption (2.17), then one can prove estimate (2.13) (see Sec. 2.2 and Refs. [P2, 3]).

Chapitre 3

Calcul de moyennes en temps long pour des systèmes dynamiques Hamiltoniens intégrables

Ce chapitre reprend l'intégralité d'un article écrit en collaboration avec Eric Cancès, François Castella, Philippe Chartier, Erwan Faou, Claude Le Bris et Gabriel Turinici, et soumis à *Numerische Mathematik* [P2].

Nous nous intéressons ici à l'analyse numérique de schémas pour le calcul de moyennes temporelles en temps long sur des trajectoires de systèmes dynamiques Hamiltoniens intégrables. L'application de ces schémas au domaine de la dynamique moléculaire a été présentée au chapitre 2. Les schémas étudiés procèdent d'une double discrétisation, celle de la moyenne temporelle et celle de la solution du système dynamique. Nous étudions l'effet des deux discrétisations. Le schéma couramment utilisé en dynamique moléculaire pour calculer une moyenne temporelle est analysé. De plus, nous proposons et analysons de nouveaux schémas, qui permettent d'accélérer la vitesse de convergence de la moyenne temporelle vers sa limite en temps infini.

Long-time averaging for integrable Hamiltonian dynamics

Eric Cancès[a], François Castella[b,c], Philippe Chartier[c], Erwan Faou[c], Claude Le Bris[a], Frédéric Legoll[a,d] and Gabriel Turinici[a]

[a] *CERMICS, Ecole Nationale des Ponts et Chaussées, 6 et 8 avenue Blaise Pascal, 77455 Marne-la-Vallée Cedex 2*
and
MICMAC, INRIA Rocquencourt, Domaine de Voluceau, 78153 Le Chesnay Cedex
[b] *IRMAR, Université de Rennes 1, Campus de Beaulieu, 35042 Rennes Cedex*
[c] *IPSO, INRIA Rennes, Campus de Beaulieu, 35042 Rennes Cedex*
[d] *EDF R & D, Analyse et Modèles Numériques, 1, avenue du Général de Gaulle, 92140 Clamart*

{cances,lebris,legoll}@cermics.enpc.fr,
francois.castella@univ-rennes1.fr,
{chartier,efaou}@irisa.fr,
Gabriel.Turinici@inria.fr

Given a Hamiltonian dynamical system, we address the question of computing the space-average of an observable through the limit of its time-average. For a completely integrable system, it is known that ergodicity can be characterized by a diophantine condition on its frequencies and that the two averages then coincide. In this paper, we show that we can improve the rate of convergence upon using a filter function in the time-averages. We then show that this convergence persists when a symplectic numerical scheme is applied to the system, up to the order of the integrator.

3.1 Introduction

Consider a Hamiltonian dynamical equation in $\mathbb{R}^d \times \mathbb{R}^d$

$$\left\{ \begin{array}{rcl} \dot{p}(t) & = & -\nabla_q H(p(t),q(t)), \quad p(0) = p_0, \\ \dot{q}(t) & = & \nabla_p H(p(t),q(t)), \quad q(0) = q_0. \end{array} \right. \tag{3.1}$$

Let $M(p_0,q_0)$ be the manifold $\{(p,q) \in \mathbb{R}^{2d} \,|\, H(p,q) = H(p_0,q_0)\}$. The solution of (3.1) is a dynamical system on $M(p_0,q_0)$ with the invariant measure

$$d\rho(p,q) = \frac{d\sigma(p,q)}{\|\nabla H(p,q)\|_2},$$

where $d\sigma(p,q)$ is the measure induced on $M(p_0,q_0)$ by the Euclidean metric of $\mathbb{R}^{2d}$ (see for instance [6]), and $\|\cdot\|_2$ the Euclidean norm in $\mathbb{R}^{2d}$.

It is a common problem to estimate the *space* average of an observable[1] A over the manifold $M(p_0,q_0)$

$$\frac{\displaystyle\int_{M(p_0,q_0)} A(p,q)d\rho(q,p)}{\displaystyle\int_{M(p_0,q_0)} d\rho(q,p)}, \tag{3.2}$$

through the limit of the *time* average

$$\lim_{T\to\infty}\frac{1}{T}\int_0^T A(p(t),q(t))dt, \tag{3.3}$$

where $(p(t),q(t))$ is the solution of (3.1). Our wish is here to give a sound ground to (and in some cases improve [P1]) the numerical simulations of (3.3) commonly used in the field of molecular dynamics.

The conditions under which the two quantities (3.2) and (3.3) coincide cannot be stated *in general*. In contrast, a dynamical system is known to have an ergodic behavior in two clearly identified -and somewhat opposite- situations :

- in the case of a differential equation with an hyperbolic structure, giving rise to mixing, the convergence of (3.3) toward (3.2) for T going to infinity is insured at a typical rate of $1/\sqrt{T}$. It is the belief of the authors that not much can be gained in this situation due to the presence of chaos ;
- in the case of an *integrable* system, a well-known result of Bohl, Sierpinski and Weyl (see [3] p. 287 and references therein) states that, under a *non-resonant* condition on the frequency vector associated with the initial condition, the space average of a continuous function on the manifold

$$\begin{aligned} S(p_0,q_0) \;=\; & \{(p,q)\in\mathbb{R}^d\times\mathbb{R}^d\ ; \\ & I_1(p,q)=I_1(p_0,q_0),\ldots,I_d(p,q)=I_d(p_0,q_0)\}, \end{aligned} \tag{3.4}$$

  where $I_1,\ldots,I_d$ are the d invariants of the problem (3.1), coincide with the long-time average of this function. Moreover, if the frequencies satisfy a *diophantine* condition, the convergence is of order T^{-1}. Being more analytically tractable, this case allows for the design of more elaborated averaging methods than the straightforward numerical simulation of (3.3).

[1] Properties of a physical system at thermodynamical equilibrium such as *radial distributions, free energies, transport coefficients* can be computed as averages of some observables over the phase space of a representative microscopic system. In most applications of interest, this microscopic system is composed of a high number of particles, making the computation of averages a challenging issue.

In realistic situations, Hamiltonian systems belong neither to the first category, nor to the second one : they typically exhibit different behaviors for different energy levels. Nevertheless, the acceleration techniques presented in this paper are relevant to actual computations for the following two reasons :

- their efficiency shows off also in situations where integrability assumptions are not satisfied (see the companion paper [P1]).
- their induced computational overhead is only marginal and thus not penalizing when integrability assumptions are violated. Meanwhile, when the explored energy level is such that the system can be (locally) considered as integrable, a significant acceleration is observed.

Integrable systems under some diophantine condition will thus constitute a natural framework for this work. Besides, all the results presented here could be extended with only minor modifications to the case of near-integrable systems.

In the following, we consider a completely integrable Hamiltonian system (3.1) in the sense of the Arnold-Liouville theorem [3, 9] : there exist d invariants $I_1 = H, I_2, \ldots, I_d$ in involution (i.e. their Poisson Bracket $\{I_i, I_j\} = 0$) such that their gradient are everywhere independent, and the trajectories of the system remain bounded. Under these conditions, there exist action-angles variables (a, θ) in a neighborhood U of $S(p_0, q_0)$. We have $(p, q) = \psi(a, \theta)$, where ψ is a symplectic transformation

$$\psi : D \times \mathbb{T}^d \ni (a, \theta) \mapsto (p, q) \in U,$$

with $\mathbb{T}^d = (\mathbb{R}/2\pi\mathbb{Z})^d$ the standard d-dimensional flat torus, and D a neighborhood in $\mathbb{R}^d$ of the point a_0 such that $(a_0, \theta_0) = \psi^{-1}(p_0, q_0)$. By definition of action-angle variables, the Hamiltonian $H(p, q)$ of (3.1) is written $H(p, q) = K(a)$ in the coordinates (a, θ), and thus the dynamics reads

$$\left\{ \begin{array}{rcl} \dot{a}(t) & = & 0, \\ \dot{\theta}(t) & = & \omega(a(t)), \end{array} \right. \tag{3.5}$$

where $\omega = \partial K/\partial a$ is the frequency vector associated with the problem. The solution of this system for initial data (a_0, θ_0) is simply written $a(t) = a_0$ and $\theta(t) = \omega(a_0)t + \theta_0$.

For fixed $(a_0, \theta_0) = \psi(p_0, q_0)$, the image of $S(p_0, q_0)$ under ψ^{-1} is the torus $\{a_0\} \times \mathbb{T}^d$. On this torus, the measure $d\theta$ is invariant by the flow of (3.5). Considering the pull-back of this measure by the transformation ψ, we thus get a measure $d\mu(p, q)$ on $S(p_0, q_0)$ which is invariant by the flow of (3.1). For any function $A(p, q)$ defined on $S(p_0, q_0)$ we define the *space* average :

$$\langle A \rangle := \frac{\displaystyle\int_{S(p_0,q_0)} A(p,q) d\mu(p,q)}{\displaystyle\int_{S(p_0,q_0)} d\mu(p,q)} = \frac{1}{(2\pi)^d} \int_{\mathbb{T}^d} A \circ \psi(a_0, \theta) d\theta. \tag{3.6}$$

For a fixed time T, the *time* average is defined as

$$\langle A \rangle(T) := \frac{1}{T} \int_0^T A(p(t), q(t)) dt. \tag{3.7}$$

In a first step, we will investigate the extent to which the convergence of the time average (3.7) toward the space average (3.6) can be accelerated through the use of weighted integrals of the form

$$\langle A \rangle_\varphi(T) := \frac{\displaystyle\int_0^T \varphi\left(\frac{t}{T}\right) A(p(t), q(t)) dt}{\displaystyle\int_0^T \varphi\left(\frac{t}{T}\right) dt}, \tag{3.8}$$

where φ is a positive smooth function with compact support in $[0, 1]$ (later on, we will refer to φ as the *filter* function ; it is sometimes refereed as a *window* function in the context of signal processing [11]). In a second step, we will consider the time-discretization of (3.8), i.e. the discretization of both the integral through Riemann sums and the trajectory with symplectic integrators. In particular, we will derive estimates of the convergence with respect to T and the size h of the time-grid, which are in perfect agreement with the numerical experiments conducted in [P1].

3.2 The complete analysis of the d-dimensional harmonic oscillator

In this section, we illustrate the main ideas of the paper in the rather simple situation of the d-dimensional harmonic oscillator, where most of the analysis can be conducted in an explicit way. Hereafter, $H(p, q)$ is thus the Hamiltonian function from $\mathbb{R}^d \times \mathbb{R}^d$ to $\mathbb{R}$ defined as

$$H(p, q) = \frac{1}{2} \sum_{k=1}^{d} (\omega_k^2 q_k^2 + p_k^2), \tag{3.9}$$

and the corresponding dynamics is governed by the equations

$$\left\{ \begin{array}{rcl} \dot{p}_k &=& -\omega_k^2 q_k \\ \dot{q}_k &=& p_k \end{array} \right. , \ k = 1, \ldots, d.$$

The exact trajectory lies on the d-dimensional manifold $S(p_0, q_0)$ defined by (3.4) where the $I_k(p, q) = \frac{1}{2}(\omega_k^2 q_k^2 + p_k^2)$ are the conserved energies of the d oscillators. Hence, denoting $r_k^0 = \sqrt{2 I_k(p_0, q_0)}$, $k = 1, \ldots, d$ and $z = (\omega_1 q_1 + i p_1, \ldots, \omega_d q_d + i p_d)$ the aggregated vector of rescaled positions and momenta, the exact solution is of the form

$$z(t) = \left(r_1^0 e^{i(\omega_1 t + \phi_1)}, \ldots, r_d^0 e^{i(\omega_d t + \phi_d)} \right), \tag{3.10}$$

where $\phi = (\phi_1, \ldots, \phi_d)$ depends on the initial conditions (p_0, q_0). As a consequence, the space average (3.6) we wish to approximate may be written here as :

$$\langle A \rangle = \frac{1}{(2\pi)^d} \int_{\mathbb{T}^d} (A \circ \Delta)(r^0, \theta) d\theta,$$

where $\Delta(r^0, \theta) = \left(\frac{r_1^0}{\omega_1} \cos(\theta_1), r_1^0 \sin(\theta_1), \ldots, \frac{r_d^0}{\omega_d} \cos(\theta_d), r_d^0 \sin(\theta_d) \right)$. As for the time-average (3.7), it reads :

$$\langle A \rangle(T) = \frac{1}{T} \int_0^T (A \circ \Delta)(r^0, \omega t + \phi) dt.$$

In order to estimate the rate of convergence of (3.7) toward (3.6), we expand $A \circ \Delta$ in a Fourier series (the conditions under which this expansion is valid will be detailed in the following sections) :

$$(A \circ \Delta)(r^0, \theta) = \sum_{\alpha \in \mathbb{Z}^d} \widehat{A \circ \Delta}(r^0, \alpha) e^{i\alpha \cdot \theta},$$

where $\alpha \cdot \theta = \alpha_1 \theta_1 + \ldots + \alpha_d \theta_d$ and with :

$$\widehat{A \circ \Delta}(r^0, \alpha) = \frac{1}{(2\pi)^d} \int_{\mathbb{T}^d} (A \circ \Delta)(r^0, \theta) e^{-i\alpha \cdot \theta} d\theta.$$

In particular, $\widehat{A \circ \Delta}(r^0, 0) = \langle A \rangle$. Hence, we have :

$$|\langle A \rangle - \langle A \rangle(T)| \leq \frac{1}{T} \sum_{\alpha \in \mathbb{Z}^d, \alpha \neq 0} \frac{2 \left| \widehat{A \circ \Delta}(r^0, \alpha) \right|}{|\alpha \cdot \omega|}. \tag{3.11}$$

This infinite sum can then be bounded if we assume, on one hand, that the vector of frequencies $\omega = (\omega_1, \ldots, \omega_d)$ satisfies *Siegel's diophantine condition*

$$\exists \gamma, \nu > 0, \quad \forall \alpha \in \mathbb{Z}^d, \quad |\alpha \cdot \omega| > \gamma |\alpha|^{-\nu}, \tag{3.12}$$

and on the other hand, that the Fourier coefficients decay sufficiently rapidly. This relatively poor rate of convergence $(1/T)$ may now be considerably improved by considering *iterated* averages of the form :

$$\langle A \rangle_k(T) := \frac{1}{T^k} \int_0^T \ldots \int_0^T (A \circ \Delta)(r^0, (t_1 + \ldots + t_k)\, \omega + \phi) dt_1 \ldots dt_k. \tag{3.13}$$

Using Fourier expansions as in (3.11), we indeed obtain in a very similar way the following error estimate for (3.13) :

$$|\langle A \rangle - \langle A \rangle_k(T)| \leq \frac{1}{T^k} \sum_{\alpha \in \mathbb{Z}^d, \alpha \neq 0} \frac{2 \left| \widehat{A \circ \Delta}(r^0, \alpha) \right|}{|\alpha \cdot \omega|^k}, \tag{3.14}$$

and under slightly more stringent bounds on the $\left|\widehat{A\circ\Delta}(r^0,\alpha)\right|$, (3.14) leads to a rate of convergence of $1/T^k$. Inspired by these computations, and noticing that (3.13) is a special case of (3.8) (more precisely $\langle A\rangle_k(T/k) = \langle A\rangle_\varphi(T)$ with $\varphi \equiv \chi^{*k}_{[0,1/k]}$, the k^{th}-convolution of the characteristic function of $[0,1/k]$), we will consider in the sequel more general *filter* functions and demonstrate that the rate of convergence can be further improved.

Now, a natural question that arises is whether the techniques explained above are amenable to numerical computations, when both the trajectory $z(t)$ and the integrals (3.7) or (3.13) are approximated using numerical schemes. In the case of the harmonic oscillator, it turns out that the numerical trajectory $z^h(t_n)$ (i.e. the approximation at time $t_n = nh$ of $z(t_n)$), when the underlying scheme is a symplectic (or symmetric) Runge-Kutta method, may be interpreted as the exact solution of a harmonic oscillator with *modified* frequencies $\omega_k^h = \omega_k\Theta(h\omega_k)$. In particular, the numerical trajectory lies on the same manifold $S(q_0,p_0)$ as the exact one. For the velocity-Verlet scheme (and partitioned methods), the same interpretation is possible, though the numerical trajectory would lie on an invariant torus $\mathcal{O}(h^2)$-close to $S(q_0,p_0)$: this situation is more typical of what happens for general integrable Hamiltonian systems.

In our situation, we have :

$$z^h(t_n) = \left(r_1^0 e^{i(\omega_1\Theta(h\omega_1)t_n+\phi_1)}, \ldots, r_d^0 e^{i(\omega_d\Theta(h\omega_d)t_n+\phi_d)}\right),$$

where Θ is a smooth function defined by

$$\Theta(y) = \frac{1}{y}\arctan\left(\frac{R(iy)-R(-iy)}{i(R(iy)+R(-iy))}\right),$$

$R(z)$ being the *stability function* of the method (in fact, Θ is real analytic as soon as R has no pole on the imaginary axis and satisfies $\Theta(y) = 1+\mathcal{O}(y^r)$ where r denotes the order of convergence of the Runge-Kutta method). As a consequence, the Riemann sum associated with (3.13) (note that (3.15) with $k=1$ corresponds to (3.7)) reads, for $T = nh$, $n\in\mathbb{N}$,

$$\langle A\rangle_k^{\text{Rie}}(T) := \frac{1}{n^k}\sum_{j_1=0}^{n-1}\cdots\sum_{j_k=0}^{n-1}(A\circ\Delta)(r^0,(j_1+\ldots+j_k)h\,\omega\,\Theta(\omega h)+\phi), \qquad (3.15)$$

where $\omega\Theta(\omega h) = (\omega_1\Theta(\omega_1 h),\ldots,\omega_d\Theta(\omega_d h))$, so that using once again Fourier expansions, we get straightforwardly :

$$|\langle A\rangle - \langle A\rangle_k^{\text{Rie}}(T)| \le \frac{1}{n^k}\sum_{\alpha\in\mathbb{Z}^d,\,\alpha\neq 0}\left|\widehat{A\circ\Delta}(r^0,\alpha)\right|\left|\frac{e^{inh\alpha\cdot(\omega\Theta(\omega h))}-1}{e^{ih\alpha\cdot(\omega\Theta(\omega h))}-1}\right|^k. \qquad (3.16)$$

Bounding the above infinite sum now requires to bound the term $|e^{inx}-1|/|e^{ix}-1|$ for x of the form $x = h\alpha\cdot(\omega\Theta(\omega h))$. To this aim, we use the following two inequalities

$$\exists\, C_0, x_0 > 0,\ \forall\, n \in \mathbb{N},\ \forall\, |x| \le x_0,\ \left|\frac{e^{inx}-1}{e^{ix}-1}\right| \le C_0\frac{1}{|x|}, \tag{3.17}$$

$$\forall\, n \in \mathbb{N},\ \forall\, x \in \mathbb{R},\ \left|\frac{e^{inx}-1}{e^{ix}-1}\right| \le n, \tag{3.18}$$

according to whether $|x|$ is small (3.17) or not (3.18). The bound we are looking for is now based on the following lemma :

Lemma 3.2.1 *Assume that the vector of frequencies ω satisfies the diophantine condition* (3.12) *and the Runge-Kutta method is of order r. Then, there exist strictly positive constants c and h_0 such that*

$$\forall\, h \le h_0 \quad \forall\, \alpha \in \mathbb{Z}^d, \quad |\alpha\cdot(\omega\,\Theta(\omega h))| \le \frac{\gamma}{2}\,|\alpha|^{-\nu} \Longrightarrow |\alpha| \ge c\,h^{-\frac{r}{\nu+1}}.$$

Proof. Assume that there exists $\alpha \in \mathbb{Z}^d$ such that

$$|\alpha\cdot(\omega\,\Theta(\omega h))| \le \frac{\gamma}{2}\,|\alpha|^{-\nu}.$$

Then, from $\Theta(h\omega_k) = 1 + \mathcal{O}(|h\,\omega_k|^r)$, we obtain for h sufficiently small :

$$\begin{aligned}\frac{\gamma}{2}\,|\alpha|^{-\nu} &\ge |\omega\cdot\alpha| - C|\alpha|\,|h\,\omega|^r,\\ &\ge \gamma|\alpha|^{-\nu} - C|\alpha|\,|h\,\omega|^r,\end{aligned}$$

where C is the strictly positive constant contained in the term $\mathcal{O}$ (note that if $\Theta \equiv 1$, although the constant C is zero, there is no α violating condition (3.12) and the lemma remains valid). Hence,

$$|\alpha| \ge \left(\frac{\gamma}{2C|\omega|^r}\,h^{-r}\right)^{\frac{1}{\nu+1}}.$$

■

But for $|\alpha| \le ch^{-r/(\nu+1)}$ we have $|h\alpha\cdot\omega\Theta(\omega h)| \le \tilde{c}h^{1-r/(\nu+1)}$ for a constant $\tilde{c}$ independent of h. Hence if $\nu > r-1$, then for small enough h we have $|h\alpha\cdot\omega\Theta(\omega h)| \le x_0$ defined in (3.17). Now we can split the sum in (3.16) into

$$\sum_{1\le|\alpha|\le ch^{-\frac{r}{\nu+1}}} \left|\widehat{A\circ\Delta}(r^0,\alpha)\right| \frac{C_0^k}{n^k h^k |\alpha\cdot(\omega\,\Theta(\omega h))|^k} + \sum_{|\alpha|\ge ch^{-\frac{r}{\nu+1}}} \left|\widehat{A\circ\Delta}(r^0,\alpha)\right|. \tag{3.19}$$

Using Lemma 3.2.1 for the first term and assuming that the Fourier coefficients $\left|\widehat{A\circ\Delta}(r^0,\alpha)\right|$ decay exponentially with $|\alpha|$, an estimate of the form

$$|\langle A\rangle - \langle A\rangle_k^{\mathrm{Rie}}(T)| = \mathcal{O}\left(\frac{1}{T^k} + \exp\Big(-ch^{-s}\Big)\right),$$

with $s = r/(\nu+1)$. Whenever a symplectic partitioned method is used, the quadratic invariants I_k might be preserved only up to the order of the scheme, and an additional term h^r then comes into play which becomes dominant : for general Runge-Kutta methods, the best possible estimate is thus of the form

$$|\langle A\rangle - \langle A\rangle_k^{\mathrm{Rie}}(T)| = \mathcal{O}\left(\frac{1}{T^k} + h^r\right). \tag{3.20}$$

The term $1/T^k$ is the intrinsic error component of the *iterated*-average, whereas the term h^r reflects the use of a numerical scheme of order r. It is worth noticing that there is **no secular component** in h^r (neither in the bound e^{-c/h^s}) : symplectic schemes (partitioned or not) preserve quadratic invariants for *all times* (either exactly or up to the order of the method). Our aim in next sections is to prove that (3.20) remains true over *exponentially long times* for averages with general filter functions and for general integrable Hamiltonian systems with bounded trajectories.

3.3 Approximation of the average : The continuous case

The function φ considered in Formula (3.8) is somewhat arbitrary. The most commonly used function in practice is $\varphi \equiv 1$, which corresponds to the usual time-average as defined in (3.7), for which convergence when T tends to infinity is rather slow (with rate $1/T$). For the harmonic oscillator, we have seen that the use of iterated-averages (which can be seen as a special case of filtered-averages) allows for a significant acceleration of the convergence. Theorem 3.3.1 below shows that with increasingly smooth functions φ satisfying appropriate zero boundary conditions, it is possible to improve the rate of convergence to $1/T^k$ for any integer $k > 1$, not only for the harmonic oscillator, but for a general integrable Hamiltonian system. It is then natural to investigate what happens in the limit when k tends to infinity. To this aim, we shall consider, as an example of infinitely differentiable functions φ with compact support $[0,1]$ that satisfy $\varphi^{(k)}(0) = \varphi^{(k)}(1) = 0$ for any $k \in \mathbb{N}$, the function ξ defined below :

$$\begin{array}{rcl} \xi : [0,1] & \longrightarrow & [0,+\infty[\\ x & \longmapsto & \exp\left(-\dfrac{1}{x(1-x)}\right). \end{array} \tag{3.21}$$

In the sequel, we shall assume that the estimates

$$\|\xi^{(k)}\|_{\mathrm{L}^1} := \int_0^1 |\xi^{(k)}(x)|dx \leq \mu\beta^k k^{\delta k}, \tag{3.22}$$

$$\|\xi^{(k)}\|_{\mathrm{L}^\infty} := \sup_{x\in[0,1]} |\xi^{(k)}(x)| \leq \mu\beta^k k^{\delta k}, \tag{3.23}$$

hold for some strictly positive constants μ, β and δ. The existence of such constants will be shown in Appendix (Lemma 3.7.1).

Theorem 3.3.1 *Consider the completely integrable system* (3.1)*, and assume that the diophantine condition* (3.12) *is satisfied for* $\omega(a_0)$ *defined in* (3.5) *by the initial condition* (q_0, p_0)*, with* $(q_0, p_0) = \psi(a_0, \theta_0)$*. Consider a function* A *real analytic on* $\mathbb{R}^d \times \mathbb{R}^d$ *(the observable). Recall that to this function we associate the space-average* $\langle A \rangle$*, the time-average* $\langle A \rangle(T)$ *and the filtered time-average* $\langle A \rangle_\varphi(T)$ *respectively defined in* (3.6), (3.7) *and* (3.8)*, where* $\varphi \in C^0(0,1)$ *(the filter function) is assumed to be positive. Then we have the following convergence estimates :*

1. *There exists a constant* c *depending on* A, d, ν *and* γ *such that*

$$|\langle A \rangle(T) - \langle A \rangle| \leq \frac{c}{T}.$$

2. *Let* $k \geq 1$*. If* φ *is* $C^{k+1}(0,1)$ *with* $\varphi^{(j)}(0) = \varphi^{(j)}(1) = 0$ *for all* $j = 0, \ldots, k-1$*, then there exist positive constants* c_0 *and* R *depending on* A, φ, d, ν *and* γ*, such that (here* $\nu \in \mathbb{N}$*, though a similar formula holds for general* ν *using the* Γ *function)*

$$|\langle A \rangle(T) - \langle A \rangle| \leq \frac{c(k,\varphi)}{T^{k+1}},$$

where

$$c(k,\varphi) = c_0 R^{k+1} (\nu(k+1)+1)! \ \ \frac{|\varphi^{(k)}(0)| + |\varphi^{(k)}(1)| + \|\varphi^{(k+1)}\|_{\mathrm{L}^1}}{\|\varphi\|_{\mathrm{L}^1}}.$$

3. *If* ξ *defined in* (3.21) *is taken as the filter function, then there exist strictly positive constants* c_1, κ *and* ρ *depending on* A, d, ν *and* γ*, such that*

$$|\langle A \rangle_\xi(T) - \langle A \rangle| \leq c_1 e^{-\kappa T^{1/\rho}}.$$

Proof. Statement 1. is proved in [3]. It may also be obtained as a special case of 2. with $\varphi \equiv 1$. Now, if A is real analytic on $\mathbb{R}^d \times \mathbb{R}^d$, then so is $A \circ \psi$ on the d-dimensional torus $\mathbb{T}^d$ and we can expand it as a Fourier series

$$(A \circ \psi)(a_0, \alpha) \ = \ \sum_{\alpha \in \mathbb{Z}^d} \widehat{A \circ \psi}(a_0, \alpha) e^{i\alpha \cdot \theta},$$

with exponentially decaying coefficients :

$$\forall\, \alpha \in \mathbb{Z}^d, \ \left| \widehat{A \circ \psi}(a_0, \alpha) \right| \leq C e^{-\frac{|\alpha|}{C}},$$

where C is a strictly positive real constant. The integral over $\mathbb{T}^d$ of the first coefficient of the series ($\alpha = 0$) is straightforwardly identified as the space-average

$$\widehat{A \circ \psi}(a_0, 0) = \frac{1}{(2\pi)^d} \int_{\mathbb{T}^d} (A \circ \psi)(a_0, \theta) d\theta.$$

Writing $\int_0^T \varphi(\frac{t}{T})dt = T\|\varphi\|_{\mathrm{L}^1} := \chi^{-1}$, the error can be computed as follows :

$$\begin{aligned}\langle A\rangle_\varphi(T) - \langle A\rangle &= \chi \sum_{\alpha\in\mathbb{Z}^d,\ \alpha\neq 0} \widehat{A\circ\psi}(a_0,\alpha)\int_0^T \varphi\Big(\frac{t}{T}\Big)e^{i\alpha\cdot(\theta_0+t\omega(a_0))}dt \\ &= \chi \sum_{\alpha\in\mathbb{Z}^d,\ \alpha\neq 0} \widehat{A\circ\psi}(a_0,\alpha)e^{i(\alpha\cdot\theta_0)}\int_0^T \varphi\Big(\frac{t}{T}\Big)e^{it(\alpha\cdot\omega(a_0))}dt. \quad (3.24)\end{aligned}$$

Now, the integral in each term of the series can be integrated by parts

$$\begin{aligned}\int_0^T \varphi\Big(\frac{t}{T}\Big)e^{it(\alpha\cdot\omega(a_0))}dt &= \left[\frac{\varphi\big(\frac{t}{T}\big)e^{it(\alpha\cdot\omega(a_0))}}{i(\alpha\cdot\omega(a_0))}\right]_0^T \\ &\quad -\frac{1}{Ti(\alpha\cdot\omega(a_0))}\int_0^T \varphi'\Big(\frac{t}{T}\Big)e^{it(\alpha\cdot\omega(a_0))}dt.\end{aligned}$$

Integrating repeatedly by parts, this last term can be written as

$$\begin{aligned}&\frac{e^{iT(\alpha\cdot\omega(a_0))}\varphi(1)-\varphi(0)}{i(\alpha\cdot\omega(a_0))} - \frac{1}{Ti(\alpha\cdot\omega(a_0))}\int_0^T \varphi'\Big(\frac{t}{T}\Big)e^{it(\alpha\cdot\omega(a_0))}dt \\ &\qquad = \ldots = \frac{(-1)^k}{(Ti(\alpha\cdot\omega(a_0)))^k}\int_0^T \varphi^{(k)}\Big(\frac{t}{T}\Big)e^{it(\alpha\cdot\omega(a_0))}dt,\end{aligned}$$

and eventually,

$$\begin{aligned}\int_0^T \varphi\Big(\frac{t}{T}\Big)e^{it(\alpha\cdot\omega(a_0))}dt &= \frac{(-1)^k}{(Ti(\alpha\cdot\omega(a_0)))^{k+1}}T\left[\varphi^{(k)}\Big(\frac{t}{T}\Big)e^{it(\alpha\cdot\omega(a_0))}\right]_0^T \\ &\quad - \frac{(-1)^k}{(Ti(\alpha\cdot\omega(a_0)))^{k+1}}\int_0^T \varphi^{(k+1)}\Big(\frac{t}{T}\Big)e^{it(\alpha\cdot\omega(a_0))}dt.\end{aligned}$$

Inserting this expression in equation (3.24) and taking the moduli of both sides, we finally get the bound

$$\begin{aligned}|\langle A\rangle_\varphi(T) - \langle A\rangle| &\leq \frac{|\varphi^{(k)}(0)| + |\varphi^{(k)}(1)| + \|\varphi^{(k+1)}\|_{\mathrm{L}^1}}{T^{k+1}\|\varphi\|_{\mathrm{L}^1}} \\ &\quad \times \sum_{\alpha\in\mathbb{Z}^d,\ \alpha\neq 0} \frac{\Big|\widehat{A\circ\psi}(a_0,\alpha)\Big|}{|\alpha\cdot\omega(a_0)|^{k+1}}\end{aligned}$$

It remains to justify the convergence of the series considered above (and to bound its limit). This is a consequence of the diophantine condition $|\alpha\cdot\omega(a_0)| \leq \frac{\gamma}{|\alpha|^\nu}$, which

gives here

$$\sum_{\alpha\in\mathbb{Z}^d,\ \alpha\neq 0} \frac{\left|\widehat{A\circ\psi}(a_0,\alpha)\right|}{|\alpha\cdot\omega(a_0)|^{k+1}} \leq \sum_{\alpha\in\mathbb{Z}^d,\ \alpha\neq 0} Ce^{-\frac{|\alpha|}{C}}\left(\frac{|\alpha|}{\gamma^{1/\nu}}\right)^{\nu(k+1)},$$
$$\leq C\eta^{\nu(k+1)}\sum_{\alpha\in\mathbb{Z}^d,\ \alpha\neq 0} e^{-\frac{|\alpha|}{C}}\left(\frac{|\alpha|}{\eta\gamma^{1/\nu}}\right)^{\nu(k+1)}.$$

We now take $\eta = 2C/\gamma^{1/\nu}$ so that $1/(\gamma^{1/\nu}\eta) = 1/(2C)$ and we obtain :

$$\sum_{\alpha\in\mathbb{Z}^d,\ \alpha\neq 0} \frac{\left|\widehat{A\circ\psi}(a_0,\alpha)\right|}{|\alpha\cdot\omega(a_0))|^{k+1}} \leq C\eta^{\nu(k+1)}(\nu(k+1)+1)!\sum_{\alpha\in\mathbb{Z}^d} e^{-\frac{|\alpha|}{2C}},$$
$$\leq \frac{C(2C)^{\nu(k+1)}(8C)^d}{\gamma^{k+1}}(\nu(k+1)+1)!,$$

where we have used $x^n \leq e^x(n+1)!$. Statement 3. is a consequence of Statement 2. with a suitably chosen k : since $\xi^{(k)}(0) = \xi^{(k)}(1) = 0$ for any $k \in \mathbb{N}$, we have indeed that for all $k \geq 0$:

$$|\langle A\rangle_\xi(T) - \langle A\rangle| \leq c_1\left(\frac{r_1}{T}\right)^{k+1}(k+1)^{\delta(k+1)}(\nu(k+1)+1)!,$$

with $c_1 = c_0\mu$ and $r_1 = R\beta$, μ and β being the constants of (3.22). Now let $\tilde{\nu}$ be the nearest integer to ν toward infinity. This gives :

$$|\langle A\rangle_\xi(T) - \langle A\rangle| \leq c_1\left(\frac{r_1\tilde{\nu}^{\tilde{\nu}}}{T}\right)^{k+1}(k+1)^{(\delta+\tilde{\nu})(k+1)},$$
$$\leq c_1 e^{f(k+1)},$$

where $f(\ell) = \ell[\ln(r_1\tilde{\nu}^{\tilde{\nu}}/T) + (\delta+\tilde{\nu})\ln(\ell)]$. The minimum of f for positive ℓ is attained for $\ell = \frac{1}{e}\left(\frac{T}{r_1\tilde{\nu}^{\tilde{\nu}}}\right)^{1/(\tilde{\nu}+\delta)}$ and is

$$f_{min} = -\frac{(\delta+\tilde{\nu})}{e}\left(\frac{T}{r_1\tilde{\nu}^{\tilde{\nu}}}\right)^{\frac{1}{(\delta+\tilde{\nu})}}.$$

■

Remark 3.3.1 *In the proof of Theorem 3.3.1, one gets* $c_0 = C(8C)^d$, $R = (2C)^\nu/\gamma$, $c_1 = \mu c_0$, $\kappa = (\delta+\tilde{\nu})e^{-1}\tilde{\nu}^{-\frac{\tilde{\nu}}{\tilde{\nu}+\delta}}$ *and* $\rho = (\delta+\tilde{\nu})$, *where* $\tilde{\nu} = \nu+1$. *The values of these constants rely heavily on the sharpness of estimates (3.22) and it is likely that they might be improved. Nevertheless, the convergence behavior would be essentially the same for large dimensions : even if* ξ *were analytic, one would get* $\rho = 1+\tilde{\nu}$. *More noticeably, since almost all frequencies* $\omega(a_0)$ *satisfy the diophantine condition for some* γ *as soon as* $\nu > d-1$, *we may think of* $\tilde{\nu}$ *as being* d *and thus* δ *as being approximately* $1+d$. *The rate of convergence thus directly depends on the dimension*

of the phase-space.

3.4 Semi-discrete averages

We now wish to investigate whether the estimates of Theorem 3.3.1 persist when one replaces the integrals by Riemann sums. It turns out, quite remarkably, that its proof can be almost readily adapted.

Theorem 3.4.1 *Assume that the conditions of Theorem* 3.3.1 *are satisfied and let* $T = nh > 0$ *for a given integer* $n \geq 2$. *Let us further define the Riemann sums corresponding to the* continuous *time-average*

$$\langle A \rangle^{\mathrm{Rie}}(T) := \frac{1}{n} \sum_{j=0}^{n-1} A(p(jh), q(jh)),$$

and the filtered time-average

$$\langle A \rangle_{\varphi}^{\mathrm{Rie}}(T) := \frac{\sum_{j=0}^{n-1} \varphi(j/n) A(p(jh), q(jh))}{\sum_{j=0}^{n-1} \varphi(j/n)},$$

where $\varphi \in C^0(0,1)$ *is the filter function. Then we have the following convergence estimates :*

1. *There exist constants* c *and* c^* *depending on* A, d, ν, γ *and* $\omega = \omega(a_0)$ *such that*

$$\left| \langle A \rangle^{\mathrm{Rie}}(T) - \langle A \rangle \right| \leq \frac{c}{T} + c^* \exp\left(-\frac{1}{c^* h} \right).$$

2. *Let* $k \geq 1$. *If* φ *is* $C^{k+1}(0,1)$ *with* $\varphi^{(j)}(0) = \varphi^{(j)}(1) = 0$ *for all* $j = 0, \ldots, k-1$, *then there exist strictly positive constants* c^*, c_0 *and* R *depending on* A, φ, d, ν, γ *and* ω *such that*

$$\left| \langle A \rangle^{\mathrm{Rie}}(T) - \langle A \rangle \right| \leq \frac{c(k,\varphi)}{T^{k+1}} + c^* \exp\left(-\frac{1}{c^* h} \right),$$

where

$$c(k,\varphi) = c_0 R^{k+1} k^k (\nu(k+1)+1)! \times \frac{1}{\|\varphi\|_{\mathrm{L}^1}} \left(|\varphi^{(k)}(0)| + |\varphi^{(k)}(1)| + \|\varphi^{(k+1)}\|_{\mathrm{L}^\infty} \right).$$

3. *If* ξ *is taken as the filter function, then there exist strictly positive constants* c^*, c_1, κ *and* ρ *depending on* A, d, ν, γ *and* ω, *such that*

$$\left| \langle A \rangle_{\xi}^{\mathrm{Rie}}(T) - \langle A \rangle \right| \leq c_1 e^{-\kappa T^{1/\rho}} + c^* \exp\left(-\frac{1}{c^* h} \right).$$

Remark 3.4.1 *In the proof of Theorem 3.4.1, one gets* $\rho = (\delta + 1 + \tilde{\nu})$ *and* $\kappa = (\delta + 1 + \tilde{\nu})e^{-1}(\tilde{\nu})^{-\frac{\tilde{\nu}}{\tilde{\nu}+\delta+1}}$, *where* $\tilde{\nu} = \nu + 1$.

Proof. Statement 1. is a special case of Statement 2. with $\varphi \equiv 1$, so that we focus on the error estimate for the filtered average. Expanding $(A \circ \psi)$ in Fourier series as in Theorem 3.3.1 and denoting $S_n = \sum_{j=0}^{n-1}(1/n)\varphi(j/n)$, we have :

$$\langle A\rangle_{\varphi}^{\mathrm{Rie}}(T) - \langle A\rangle = \frac{1}{nS_n} \sum_{\alpha\,\in\,\mathbb{Z}^d,\ \alpha\neq 0} \widehat{A\circ\psi}(a_0,\alpha)e^{i(\alpha\cdot\theta_0)} \times \sum_{j=0}^{n-1} \varphi\Big(\frac{j}{n}\Big)e^{i\alpha\cdot jh\omega},$$

where $\omega = \omega(a_0)$. We use the following result, whose proof is given in Appendix :

Lemma 3.4.1 *For a given filter-function* φ *in* $C^{k+1}(0,1)$ *with* $\varphi^{(j)}(0) = \varphi^{(j)}(1) = 0$ *for all* $j = 0,\dots,k-1$, *and a given integer* $n \geq k+2$, *let* φ_j *be the real numbers defined by* $\varphi_j = \varphi(j/n)$ *for* $j = 0,\dots,n$. *If* $b \neq 1$ *is a complex number of modulus* 1, *then we have the estimate*

$$\left|\sum_{0\leq j\leq n-1} \varphi_j b^j\right| \leq \frac{2e^2k^k}{n^k|1-b|^{k+1}}\left(|\varphi^{(k)}(0)| + |\varphi^{(k)}(1)| + \|\varphi^{(k+1)}\|_{\mathrm{L}^\infty}\right).$$

Now, we can bound the previous sum by using the following splitting

$$|\langle A\rangle_{\varphi}^{\mathrm{Rie}}(T) - \langle A\rangle| \leq \frac{C(k,\varphi)}{n^{k+1}S_n} \sum_{\alpha\,\in\,\mathbb{Z}^d,\ 0<|\alpha|\leq(h|\omega|)^{-1}} \frac{\left|\widehat{A\circ\psi}(a_0,\alpha)\right|}{|1-e^{i\alpha\cdot h\omega}|^{k+1}} + \sum_{\alpha\,\in\,\mathbb{Z}^d,\ |\alpha|>(h|\omega|)^{-1}} \left|\widehat{A\circ\psi}(a_0,\alpha)\right|, \quad (3.25)$$

where we have denoted

$$C(k,\varphi) = 2e^2k^k\left(|\varphi^{(k)}(0)| + |\varphi^{(k)}(1)| + \|\varphi^{(k+1)}\|_{\mathrm{L}^\infty}\right).$$

Note that, since $0 < |\alpha| \leq (h|\omega|)^{-1}$ in the first term, we have $0 < |h\alpha\cdot\omega| \leq 1$, so that $b = e^{ih\alpha\cdot\omega} \neq 1$.

The second term in the right-hand side can be straightforwardly bounded by $c^*\exp(-\frac{1}{c^*h})$. Using (3.12), we have for all $|\alpha| \leq (h|\omega|)^{-1}$:

$$\frac{1}{|1-e^{i\alpha\cdot h\omega}|} \leq C_0\frac{1}{h|\alpha\cdot\omega|}.$$

The first term in the right-hand side can be estimated as

$$\frac{C(k,\varphi)C_0^{k+1}}{T^{k+1}S_n} \sum_{\alpha\,\in\,\mathbb{Z}^d,\ 0<|\alpha|\leq(h|\omega|)^{-1}} \frac{\left|\widehat{A\circ\psi}(a_0,\alpha)\right|}{|\alpha\cdot\omega|^{k+1}}$$

and we can conclude as in the proof of Theorem 3.3.1. ∎

3.5 Fully discrete averages

We consider the numerical trajectory (p_n, q_n) for $n \geq 0$ obtained by a symplectic r^{th}-order numerical scheme Φ_h from the initial point $(p_0, q_0) = \psi^{-1}(a_0, \theta_0)$.

For $T = nh$ and $n \in \mathbb{N}$, the corresponding Riemann sum reads

$$\langle A \rangle_{\varphi,h}^{\mathrm{Rie}}(T) := \frac{\sum_{j=0}^{n-1} \varphi(j/n) A(p_j, q_j)}{\sum_{j=0}^{n-1} \varphi(j/n)}. \tag{3.26}$$

Theorem 4.4 and 4.7 of Chapter X in [9], which strongly rely on the theory developed by Kolmogorov, Arnold and Moser [20,24,25,28,29] and on results from the backward analysis (see [9] pp. 288 and references therein), yield the following result :

Theorem 3.5.1 (Hairer, Lubich, Wanner [9]) *Let $a^* \in \mathbb{T}^d$ such that $\omega(a^*)$ satisfies the diophantine condition* (3.12) *with constants γ and ν, and suppose that $H(p,q)$ is analytic on a neighborhood of the torus $\{(p,q) = \psi(a^*, \theta) \,|\, \theta \in \mathbb{T}^d\}$. Then there exists positive constants ρ, c_0, c, C_0 and h_0 such that for all $h \leq h_0$ and $\mu \leq \min(r, \alpha)$ where $\alpha = \nu + d + 1$, the following holds : there exists a symplectic change of coordinates $\psi_h : (a, \theta) \mapsto (b, \chi)$ analytic for*

$$\|a - a^*\| \leq c_0 h^{2\mu} \quad \text{and} \quad \theta \in U_\rho = \{\theta \in \mathbb{T}^d + i\mathbb{R}^d \,;\, |\mathrm{Im}\,\theta| < \rho\}$$

and h^r-close to the identity in the sense that

$$\|(a, \theta) - \psi_h(a, \theta)\| \leq C_0 h^r \quad \text{for} \quad \|a - a^*\| \leq c_0 h^{2\mu} \quad \text{and} \quad \theta \in U_{\rho/2},$$

such that in coordinates (b, χ), the numerical trajectory $(b_n, \chi_n) = \psi_h^{-1} \circ \psi^{-1}(p_n, q_n)$ satisfies

$$\begin{aligned} b_n &= b_0 + \mathcal{O}(\exp(-ch^{-\mu/\alpha})), \\ \chi_n &= nh\omega_h(b_0) + \mathcal{O}(h^{-2\mu/\alpha} \exp(-ch^{-\mu/\alpha})), \end{aligned} \tag{3.27}$$

for $nh \leq \exp(ch^{-\mu/\alpha})$, where $\omega_h(b) = \omega(b) + \mathcal{O}(h^r)$ uniformly in b.

Using this result, we get the following Theorem :

Theorem 3.5.2 *Under the conditions and notations of Theorem 3.5.1, if the numerical trajectory starts with*

$$\|a_0 - a^*\| \leq c_0 h^{2\mu} \tag{3.28}$$

where $(a_0, \theta_0) = \psi^{-1}(p_0, q_0)$, then we have :

1. *If φ is C^{k+1} with $\varphi^{(j)}(0) = \varphi^{(j)}(1) = 0$ for all $j = 0, \ldots, k-1$ and if A is real analytic on $\mathbb{R}^d$, then there exist constants c_1 and C depending on A, γ, ν, d, k, φ, such that*

$$\forall h \leq h_0 \quad \forall\, T = nh \leq \exp(c_1 h^{-\mu/\alpha}),$$
$$|\langle A \rangle_{\varphi,h}^{\mathrm{Rie}}(T) - \langle A \rangle| \leq C \left(\frac{1}{T^{k+1}} + h^r \right). \tag{3.29}$$

2. *If ξ is taken as the filter function, if A is real analytic, then there exist constants c_1 and C, depending on A, γ, ν and d such that*

$$\forall h \leq h_0 \quad \forall\, T = nh \leq \exp(c_1 h^{-\mu/\alpha}),$$
$$|\langle A\rangle^{\mathrm{Rie}}_{\xi,h}(T) - \langle A\rangle| \leq C\left(e^{-\kappa T^{1/\rho}} + h^r\right). \quad (3.30)$$

Proof. With the notation $S_n = \sum_{j=0}^{n-1}(1/n)\varphi(j/n)$, we have using Theorem 3.5.1 that

$$\langle A\rangle^{\mathrm{Rie}}_{\varphi,h}(T) := \frac{1}{nS_n}\sum_{j=0}^{n-1}\varphi\left(\frac{j}{n}\right) A\circ\psi\circ\psi_h(b_j,\chi_j).$$

Using the Fourier expansion of $A\circ\psi\circ\psi_h$ and (3.27), we obtain

$$\langle A\rangle^{\mathrm{Rie}}_{\varphi,h}(T) := \frac{1}{nS_n}\sum_{\alpha\in\mathbb{Z}^d} \widehat{A\circ\psi\circ\psi_h}(b_0,\alpha)e^{i\alpha\cdot\varphi_0}\sum_{j=0}^{n-1}\varphi\Big(\frac{j}{n}\Big)e^{i\alpha\cdot jh\omega_h}$$
$$+\,\mathcal{O}(\exp(-ch^{-\mu/\alpha})) \quad (3.31)$$

for $nh \leq \mathcal{O}(\exp(ch^{-\mu/\alpha}))$ (we write c for a generic constant in the exponential), where $\omega_h = \omega_h(b)$.
As ψ_h is an analytic function $\mathcal{O}(h^r)$-close to the identity, we have

$$\widehat{A\circ\psi\circ\psi_h}(b_0,0) = \langle A\rangle + \mathcal{O}(h^r),$$

and the Fourier coefficients $\widehat{A\circ\psi\circ\psi_h}(b_0,\alpha)$ decay exponentially with respect to α, uniformly with respect to h. Now similarly to Lemma 3.2.1 we get that

$$\forall\, h\leq h_0 \quad \forall\,\alpha\in\mathbb{Z}^d, \quad |\alpha\cdot(h\omega_h))| \leq \frac{\gamma}{2}|\alpha|^{-\nu} \Longrightarrow |\alpha| \geq c\,h^{-\frac{r}{\nu+1}}.$$

And we conclude as in the proof of Theorem 3.4.1 using Lemma 3.4.1 and a splitting similar to (3.25). ∎

3.6 Remarks on the implementation and numerical experiments

Though optimal with respect to the rate of convergence, the filter function ξ does not seem to allow for the derivation of an error estimate : given that the values of the constant C in (3.30) is out of reach, the value of n for which

$$R_n^{\varphi} := \frac{\sum_{j=0}^{n}\varphi(j/n)A_j}{n\|\varphi\|_{\mathrm{L}^1}}$$

becomes sufficiently close (up to user's tolerance) to its limit as n goes to infinity cannot be determined in advance. An update formula for R_n^{φ} from n to $n+1$ thus

appears of much use and this should guide the choice of φ. In order to get such a formula, we study the dependence on T of

$$a(T) = \int_0^T \varphi\Big(\frac{t}{T}\Big) A(p(t), q(t)) dt.$$

Differentiating with respect to T leads to

$$\frac{da(T)}{dT} \quad = \quad \varphi(1) A(p(T), q(T)) - \frac{1}{T} \int_0^T \frac{t}{T} \varphi'\Big(\frac{t}{T}\Big) A(p(t), q(t)) dt. \qquad (3.32)$$

To be of practical use, it is thus necessary that $x\varphi'(x)$ is of the form $\alpha\varphi(x)$ (where α is an arbitrary constant) so that (3.32) becomes an ordinary differential equation for $a(T)$. The only admissible solutions are thus monomials in x. We thus consider the following polynomial filter functions

$$\varphi_p(x) \quad = \quad x^p (1-x)^p, \ p \in \mathbb{N}. \qquad (3.33)$$

Denoting for p and n in $\mathbb{N}$ the *elementary* Riemann sums

$$S_n^p = \sum_{j=0}^{n} \Big(\frac{j}{n}\Big)^p A_j,$$

it is easy to get the desired update formula

$$S_0^p = 0 \ \text{ and } \ S_n^p = A_n + (1 - 1/n)^p S_{n-1}^p, \quad n \geq 1.$$

Now, since

$$\varphi_p(x) = \sum_{k=0}^{p} (-1)^k \left(\begin{array}{c} p \\ k \end{array} \right) x^{p+k} \text{ and } \|\varphi_p\|_{\mathrm{L}^1} = \frac{(p!)^2}{(2p+1)!},$$

the approximation we seek for can be obtained as the linear combination

$$R_n^{\varphi_p} = \frac{(2p+1)!}{n(p!)^2} \sum_{k=0}^{p} (-1)^k \left(\begin{array}{c} p \\ k \end{array} \right) S_n^{p+k}.$$

We now consider the application of our method to the modified 2-dimensional Kepler problem with Hamiltonian

$$H(p,q) = p_1^2 + p_2^2 - \frac{1}{\sqrt{q_1^2 + q_2^2}} - \frac{\mu}{(\sqrt{q_1^2 + q_2^2})^3}.$$

Besides the Hamiltonian, this system has one other invariant, the angular momentum

$$L = q_2 p_1 - q_1 p_2.$$

Our goal is here to estimate the average over the manifold

$$S = \{(p,q) \in \mathbb{R}^4;\ L(p,q) = L(p_0,q_0),\ H(p,q) = H(p_0,q_0)\}$$

For $\mu = 0.2$, $p_0 = (0, 1.1)^T$ and $q_0 = (1, 0)^T$ this leads to $\langle r \rangle = 1.021466044527120$.

To this aim, we consider the Verlet method as the basic step and use the 8^{th}-order 15-stage composition of [38]. In Figures 3.1 and 3.2 are represented the errors $|\langle r \rangle_{\varphi_p}(T) - \langle r \rangle|$ in logarithmic scale for two different step-sizes. On Figure 3.1, the three curves all reach a plateau corresponding to the h^r-error term. Refining the step-size removes this plateau (or at least shifts it to the right, see Figure 3.2). In both cases, the predicted rate of convergence in $1/T^{p+1}$ is clearly observed (it corresponds to a slope of $p+1$ for φ_p).

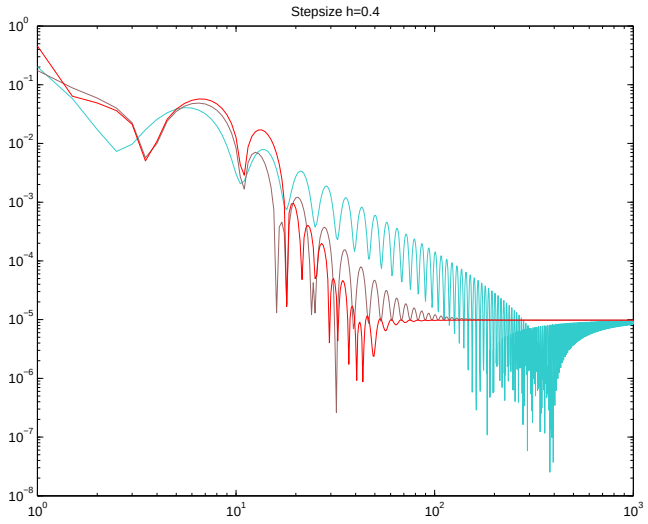

FIG. 3.1 – Error in the averages for $p = 1, 3, 5$ and $h = 0.4$ (2D-Kepler problem).

3.7 Appendix : some technical results

In this Appendix, we collect a few technical results used in the paper.

Lemma 3.7.1 *Let ξ be the function defined on $[0,1]$ by $\xi(x) = e^{-\frac{1}{x(1-x)}}$. There exist strictly positive constants $\mu \leq 1$, $\beta \leq (2\sqrt{3}+6)/e^2$ and $\delta \leq 3$ such that the following estimates hold for all $k \in \mathbb{N}^*$:*

$$\begin{aligned} \|\xi^{(k)}\|_{\mathrm{L}^1} := \int_0^1 |\xi^{(k)}(x)|dx &\leq \mu\beta^k k^{\delta k}, \\ \|\xi^{(k)}\|_{\mathrm{L}^\infty} = \sup_{x\in[0,1]} |\xi^{(k)}(x)| &\leq \mu\beta^k k^{\delta k}. \end{aligned}$$

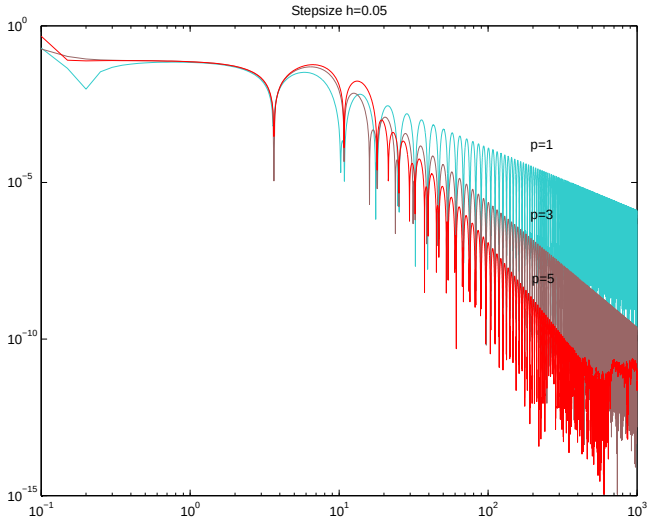

FIG. 3.2 – Error in the averages for $p = 1, 3, 5$ and $h = 0.05$ (2D-Kepler problem).

Proof. Looking for an expression of $\xi^{(k)}(x)$ of the form

$$\xi^{(k)}(x) = \frac{P_k(x)}{[\Pi(x)]^{2k}} e^{-\frac{1}{x(1-x)}},$$

where $\Pi(x) = x(1-x)$ and where P_k is a polynomial, we easily find the recurrence relation :

$$P_0 \equiv 1 \text{ and } P_{k+1} = \Pi' \left(1 - 2k\Pi\right) P_k + P_k' \, \Pi^2, \ k \geq 0. \tag{3.34}$$

We now look for bounds on balls B_r of radius $r > 0$ and center $z = 1/2 + 0\,i \in \mathbb{C}$. The bounds for Π and Π' read

$$\sup_{z \in B_r} |\Pi(z)| \leq (r^2 + 1/4), \ \sup_{z \in B_r} |\Pi'(z)| \leq r,$$

and the Cauchy integral representation of P_k' leads to

$$\forall\, \varepsilon > 0, \ \sup_{z \in B_r} |P_k'(z)| \leq \frac{r+\varepsilon}{\varepsilon} \sup_{z \in B_{r+\varepsilon}} |P_k(z)|.$$

Inserting these bounds in (3.34) we get :

$$\begin{aligned}
\sup_{z \in B_r} |P_{k+1}(z)| &\leq r[k(2r^2 - 1/2) + 1] \sup_{z \in B_r} |P_k(z)| \\
&\quad + (r^2 + 1/4)^2 \frac{r+\varepsilon}{\varepsilon} \sup_{z \in B_{r+\varepsilon}} |P_k(z)|, \\
&\leq \left(r[k(2r^2 - 1/2) + 1] + \frac{r+\varepsilon}{\varepsilon} (r^2 + 1/4)^2 \right) \sup_{z \in B_{r+\varepsilon}} |P_k(z)|.
\end{aligned}$$

Denoting $C(r,k,\varepsilon) := r[k(2r^2-1/2)+1] + \dfrac{r+\varepsilon}{\varepsilon}(r^2+1/4)^2$, we finally get

$$\begin{aligned}\sup_{z\in B_r}|P_{k+1}(z)| &\leq C(r,k,\varepsilon)\sup_{z\in B_{r+\varepsilon}}|P_k(z)|,\\ &\leq C(r,k,\varepsilon)C(r+\varepsilon,k-1,\varepsilon)\sup_{z\in B_{r+2\varepsilon}}|P_{k-1}(z)|,\\ &\leq \left(\prod_{i=0}^{k}C(r+i\varepsilon,k-i,\varepsilon)\right)\sup_{z\in B_{r+(k+1)\varepsilon}}|P_0(z)|.\end{aligned}$$

A bound can then be obtained as follows : let $\varepsilon_0 = \dfrac{-1+\sqrt{3}}{2}$, $\varepsilon = \dfrac{\varepsilon_0}{k}$ and $r = 1/2$. Then it is easy to check that for all $0 \leq i \leq k$, we have

$$\begin{aligned}C\left(\frac{1}{2}+i\frac{\varepsilon_0}{k},k-i,\frac{\varepsilon_0}{k}\right) &\leq \frac{\sqrt{3}}{2}[k-i+1]+\frac{1}{\sqrt{3}-1}k+i+1,\\ &\leq \frac{\sqrt{3}+3}{2}(k+1),\end{aligned}$$

and hence,

$$\left(\prod_{i=0}^{k}C(r+i\varepsilon,k-i,\varepsilon)\right) \leq \left[\frac{\sqrt{3}+3}{2}(k+1)\right]^{k+1}.$$

Taking into account that $P_0 \equiv 1$, we obtain

$$\forall\, k \in \mathbb{N}^*,\ \sup_{z\in B_{1/2}}|P_k(z)| \leq \left[\frac{\sqrt{3}+3}{2}k\right]^k.$$

It remains to bound $\dfrac{1}{[\Pi(x)]^{2k}}e^{-\frac{1}{x(1-x)}}$. Denoting $Y = \dfrac{1}{x(1-x)}$, we have :

$$\begin{aligned}\sup_{x\in[0,1]}\frac{1}{[\Pi(x)]^{2k}}e^{-\frac{1}{x(1-x)}} &= \sup_{Y\geq 4}e^{-Y}Y^{2k}\\ &\leq e^{-2k}(2k)! \leq \left(\frac{4}{e^2}\right)^k k^{2k}.\end{aligned}$$

■

Proof of lemma 3.4.1. Let us denote by ∇ the operator of *backward divided differences* defined by :

$$\begin{aligned}\forall\, j \in \{0,\dots,n\},\ \nabla^0\varphi_j &= \varphi_j,\\ \forall\, j \in \{m+1,\dots,n\},\ \nabla^{m+1}\varphi_j &= \nabla^m\varphi_j - \nabla^m\varphi_{j-1}.\end{aligned}$$

The sum in the statement can then be written as

$$\begin{aligned}\sum_{j=0}^{n-1}\varphi_j b^j &= \sum_{j=1}^{n-1} b^j \sum_{i=1}^{j}\nabla\varphi_i + \sum_{j=0}^{n-1}\varphi_0 b^j,\\ &= \frac{1-b^n}{1-b}\varphi_0 + \sum_{i=1}^{n-1}\nabla\varphi_i\frac{b^i-b^n}{1-b},\\ &= \frac{\varphi_0 - b^n\varphi_{n-1}}{1-b} + \frac{1}{1-b}\sum_{j=1}^{n-1}(\nabla\varphi_j)b^j = \dots,\\ &= \sum_{m=0}^{k}\frac{b^m\nabla^m\varphi_m - b^n\nabla^m\varphi_{n-1}}{(1-b)^{m+1}} + \frac{1}{(1-b)^{k+1}}\sum_{j=k+1}^{n-1}(\nabla^{k+1}\varphi_j)b^j.\end{aligned}$$

Denoting $h = 1/n$, it is well-known that, for all $n-1 \leq j \geq k+1$, there exists $\zeta_{j,k+1} \in [(j-k-1)h, jh] \subset [0,1]$ such that we have :

$$\nabla^{k+1}\varphi_j = \varphi^{(k+1)}(\zeta_{j,k+1})h^{k+1}$$

Hence, we can bound the second term in (3.35) as follows :

$$\left|\sum_{j=k+1}^{n-1}(\nabla^{k+1}\varphi_j)b^j\right| \leq \|\varphi^{(k+1)}\|_{\mathrm{L}^\infty} h^{k+1}(n-k-2).$$

In order to estimate the first sum, we notice that, for $0 \leq m \leq k \leq n-2$,

$$\nabla^m\varphi_m = \varphi^{(m)}(\zeta_{m,m})h^m$$

for some $\zeta_{m,m} \in [0, mh]$ and a Taylor-Lagrange expansion of $\varphi^{(m)}(\zeta_{m,m})$ at order $k+1-m$ gives

$$\nabla^m\varphi_m = \frac{\zeta_{m,m}^k h^k}{(k-m)!}\varphi^{(k)}(0) + \frac{\zeta_{m,m}^{k+1}h^{k+1}}{(k+1-m)!}\varphi^{(k+1)}(\eta_m)$$

for some $\eta_m \in [0, mh] \subset [0,1]$. Hence, we have :

$$\begin{aligned}\left|\sum_{m=0}^{k}\frac{b^m}{(1-b)^m}\nabla^m\varphi_m\right| &\leq |\varphi^{(k)}(0)|\frac{k^k h^k}{|1-b|^{k+1}}\sum_{m=0}^{k}\frac{|1-b|^m}{(m)!}\\ &\quad + \|\varphi^{(k+1)}\|_{\mathrm{L}^\infty}\frac{k^k h^{k+1}}{|1-b|}\sum_{m=0}^{k}\frac{|1-b|^m}{(m+1)!},\\ &\leq \frac{e^2 k^k h^k}{|1-b|^{k+1}}\left(|\varphi^{(k)}(0)| + h\|\varphi^{(k+1)}\|_{\mathrm{L}^\infty}\right).\end{aligned}$$

Similarly we have :

$$\nabla^m\varphi_{n-1} = \varphi^{(m)}(\zeta_{n-1,m})h^m$$

for some $\zeta_{n-1,m} \in [1-(m+1)h, 1-h] \subset [0,1]$, so that

$$\left| \sum_{m=0}^{k} \frac{b^n}{(1-b)^m} \nabla^m \varphi_{n-1} \right| \leq \frac{2e^2 k^k h^k}{|1-b|^{k+1}} \left(|\varphi^{(k)}(1)| + h \|\varphi^{(k+1)}\|_{L^\infty} \right).$$

Gathering the contributions of all terms then gives the result. ■

Acknowledgments The authors are glad to thank Christian Lubich for stimulating discussions on the subject of this paper, particularly for suggesting the use of general filtered averages rather than just iterated averages. We also gratefully acknowledge the financial support of INRIA through the contract grant “Action de Recherche Coopérative" PRESTISSIMO.

Chapitre 4

Construction d'algorithmes préservant une mesure et application à la dynamique moléculaire

Les résultats présentés dans cet article, obtenus en collaboration avec Régis Monneau, ont fait l'objet d'une publication dans *Journal of Chemical Physics* [P3].

La dynamique moléculaire repose sur l'équivalence entre moyennes dans l'espace des phases d'une part et moyennes temporelles le long d'une solution d'une équation différentielle ordinaire d'autre part, et permet d'approcher les premières en calculant les secondes. Lorsqu'on souhaite calculer des moyennes dans l'espace des phases pour des systèmes dont l'énergie est connue, les équations différentielles mentionnées ci-dessus sont les équations de Newton. Lorsqu'on étudie des systèmes dont l'énergie n'est pas connue (c'est le cas lorsqu'on travaille à température constante), on ne peut plus utiliser les équations de Newton, et d'autres systèmes dynamiques ont été proposés. Ces systèmes ne sont en général pas Hamiltoniens, mais ils conservent néanmoins une mesure. Dans ce chapitre, nous nous intéressons à la construction d'algorithmes préservant cette mesure et appliquons les nouveaux algorithmes proposés à un exemple simple.

Designing reversible measure invariant algorithms with applications to molecular dynamics

Frédéric Legoll[a,b] and Régis Monneau[a]

[a] *CERMICS, Ecole Nationale des Ponts et Chaussées, 6 et 8 avenue Blaise Pascal, 77455 Marne-la-Vallée Cedex 2*
[b] *EDF R & D, Analyse et Modèles Numériques, 1, avenue du Général de Gaulle, 92140 Clamart*

{legoll,monneau}@cermics.enpc.fr

A new method for generating measure invariant algorithms is presented. This method is based on a reformulation of the equations of molecular dynamics. These new equations are non-Hamiltonian but have a *normal form* which guarantees that the invariant measure is the canonical one for the new variables. Furthermore, from this normal form, one can easily build algorithms to integrate these equations. Using a Trotter-type factorization of the classical Liouville propagator, we build (time) reversible measure invariant integrators as successive direct translations. We apply this method to propose new algorithms to generate the Nosé-Hoover chain dynamics and the isothermal-isobaric dynamics. We also give a measure invariant integrator for the generalized Gaussian moment thermostatting dynamics recently introduced by Liu and Tuckerman. Finally, we present numerical results which show comparable performances with previously proposed algorithms.

4.1 Introduction

Continuous dynamical methods for generating statistical ensembles are, by now, standard. In this approach, we consider a single trajectory which generates a given sampling of the phase space. So, the integration over the trajectory of some physical quantities provides an estimate of some thermodynamic properties of a material. To calculate these estimates, we need to numerically simulate the trajectory. In general, these continuous dynamics preserve at least an energy and a measure. One may want to find algorithms that exactly preserve the energy, or the measure, or both. It has been shown [23] that for some dynamical systems, under some hypotheses, one cannot have algorithms that exactly and at the same time preserve the measure, the energy, and the other quantities preserved by the dynamics. In our case, preserving at the same time the energy and the measure seems therefore difficult. Recently, it has been shown that measure invariant algorithms play a key role to make a good

sampling of the phase space. So exactly preserving the measure is more interesting than exactly preserving the energy.

Many measure invariant algorithms have been proposed in the literature. We refer to the beautiful paper by Tuckerman and Martyna [57] for a survey of recent progress in molecular dynamics. The first molecular dynamics equations proposed to simulate the canonical ensemble were given by Nosé [69] and Hoover [62]. An improvement of this dynamics, called the Nosé-Hoover chain dynamics, has been proposed by Martyna, Klein and Tuckerman [67]. Tuckerman, Berne and Martyna [100] have proposed a first reversible measure invariant algorithm in the case of two thermostats. A generalization of this algorithm to the case of M thermostats has been given by Martyna *et al.* [48, 68]. Recently, Liu and Tuckerman [65] have proposed a new dynamics to simulate the canonical ensemble, called the generalized Gaussian moment thermostatting (GGMT) dynamics.

There are also many works on constant pressure molecular dynamics and on path integral molecular dynamics [66, 91, 121, 122]. Let us also note that multiple time scale methods have been used to improve the previous algorithms [84, 95, 97–99, 102, 103, 123].

As underlined previously, preserving at the same time the measure and the energy exactly is difficult, if not impossible. Thus, it is not surprising that the proposed measure invariant algorithms do not exactly preserve the energy of the system. This energy is only approximately conserved [15,63,64,70,96,101]. More generally, the molecular dynamics of non-microcanonical ensembles is refered to as non-Hamiltonian dynamics [71]. In practice, even if the energy is not exactly conserved by the measure invariant algorithms, there is no constant draft on it. This can be considered as very surprising. For instance, the well-known fourth-order Runge-Kutta method applied on the Kepler two-body problem gives a quadratic long term error growth [52]. The surprisingly good conservation of the energy is actually due to the fact that the algorithms are measure invariant. More precisely, let us suppose that we work with a Hamiltonian dynamics, and we use a p-order measure invariant algorithm to integrate it. The energy at the beginning is H_0, and the numerical energy at time step k is $H_{num}(k\,\Delta t)$. Then, under some regularity conditions on the Hamiltonian, and for small enough time steps Δt, one can prove [8,9] that the numerical error on the energy satisfiesab

$$\forall k \leq \frac{1}{\Delta t}\, e^{c_1/\Delta t}\,, \quad \mid H_{num}(k\,\Delta t) - H_0 \mid \leq \; c_2\, \Delta t^p,$$

where c_1 and c_2 are two constants which do not depend on the time step Δt (they depend on the Hamiltonian function which defines the dynamical system and on the measure invariant algorithm chosen). This means that the longtime error remains bounded for exponentially large times.

In this article we go one step further in the research of measure invariant algorithms. We present a *systematic method* to build reversible measure invariant

algorithms (see Sec. 4.2). Our approach allows us to generate many algorithms. In order to proceed pedagogically, we first apply our method on a simple example (see Sec. 4.3). Actually, this example is a simplified case of the equations used to simulate the isothermal-isobaric ensemble [68]. We show in appendix (see Sec. 4.5.1) how to get measure invariant algorithms for the whole set of equations. Then, in the main part of the article, we focus on the GGMT dynamics [65]. First, we show how to get new algorithms, then give some numerical results (see Sec. 4.4). In the appendix, we take a close look at measure invariance for the GGMT dynamics. For this dynamics, the algorithm proposed by Liu and Tuckerman [65] is numerically efficient. However, it is actually only approximately measure invariant (see Sec. 4.5.4 and 4.5.5). Our method gives us an algorithm which is *exactly measure invariant.*

4.2 A method to generate measure invariant algorithms

Our method simply consists in rewriting (when it is possible) systems of ordinary differential equations

$$\dot{X}_j = F_j(X_1, ..., X_n), \quad j = 1, \ldots, n, \tag{4.1}$$

in the *normal form*

$$\dot{Y}_i = G_i\left(\overset{\vee}{Y}_i\right), \quad i = 1, \ldots, n, \tag{4.2}$$

where $\overset{\vee}{Y}_i = (Y_1, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_n)$ is the whole set of variables *except* Y_i. The Y_i are obtained from the X_j by a change of variables. So, the normal form is characterized by the fact that $\dot{Y}_i$ does not depend on Y_i. Following the usual method [57], one can check that system (4.2) preserves the following measure

$$\tilde{m} = dY_1 \ \ldots \ dY_n.$$

Now, we want to build measure invariant algorithms. One can notice that the system (4.2) is a divergence-free dynamical system. Many ways to build volume-preserving algorithms for this kind of dynamics are known [22]. In this article, we take advantage of the fact that the system (4.2) is more than divergence-free, it is in a normal form. We follow a usual method [57], and build algorithms by Trotter factorizations of Liouville propagators. This gives simple algorithms by successive translations :

$$e^{\Delta t\, L_{eff}} = \prod_k e^{\Delta t_{p(k)}\, G_{p(k)}\, \partial_{Y_{p(k)}}}.$$

The k^{th} operation is a translation on the variable $Y_{p(k)}$, where $p(k)$ is a subscript in $[1, n]$. This translation preserves the measure $\tilde{m}$, since it reads $Y_{p(k)} \rightarrow Y_{p(k)} + \Delta t_{p(k)}\, G_{p(k)}$. So the complete operator $e^{\Delta t\, L_{eff}}$ also preserves the measure $\tilde{m}$.

Let us now briefly discuss the existence of a transformation such as the one considered at the beginning of this section. Is it always possible to find variables Y_i so that the dynamics on these new variables is in a normal form ? In general, it is not possible to have an explicit expression for a good change of variables. So, in general, it is not possible to explicitly transform a dynamics such as Eq. (4.1) into a normal form dynamics such as Eq. (4.2). However, for some particular cases, such a transformation exists and can be explicitly written. In the following parts of this article, we consider specific examples of dynamics (the ones usually used to generate NVT and NPT ensembles), and we explicitly transform them into normal form dynamics.

4.3 A simple application

To illustrate our purpose, we consider Nosé-Hoover equations. Actually, we are going to work on a simplified example of these equations. So we simulate a one-dimensional particle coupled with a thermostat of Nosé-Hoover [57] :

$$\begin{aligned} \dot{q} &= \frac{p}{m}, \\ \dot{p} &= F(q) - \frac{p_\xi}{Q}\, p, \\ \dot{p}_\xi &= \frac{p^2}{m} - k_B T, \\ \dot{\xi} &= \frac{p_\xi}{Q}. \end{aligned} \tag{4.3}$$

Here, the particle position is q, its impulsion is p, and its mass is m. The temperature is T, and k_B is the Boltzmann constant. The forces are $F(q) = -V'(q)$. The thermostat state is defined by two variables, ξ and p_ξ. The particle is coupled to the thermostat by the variable p_ξ, and Q is a coupling parameter (it can be considered as the thermostat mass). It is known that system (4.3) preserves the following energy and measure

$$H' = \frac{p^2}{2m} + V(q) + \frac{p_\xi^2}{2Q} + \xi\, k_B T,$$

$$m_0 = e^\xi\, dq\; dp\; dp_\xi\; d\xi.$$

To be efficient, an algorithm in molecular dynamics applied to Eqs. (4.3) has to exactly preserve the measure m_0. Actually, for many dynamical systems, the same approach has been used. The continuous dynamics preserves a measure, one has to find an algorithm that exactly preserves this measure. The discrete conservation of a measure similar to m_0 can be checked analytically or numerically. In this article we propose a unified approach to design measure invariant reversible integrators. So no *a posteriori* check is required anymore.

Let us apply our method to the system (4.3). We introduce

$$\tilde{p} = e^{\xi}\, p. \tag{4.4}$$

Then the variables $(q, \tilde{p}, p_\xi, \xi)$ satisfy

$$\begin{aligned} \dot{q} &= \frac{\tilde{p}}{m}\, e^{-\xi}, \\ \dot{\tilde{p}} &= e^{\xi}\, F(q), \\ \dot{p}_\xi &= \frac{\tilde{p}^2}{m}\, e^{-2\xi} - k_B T, \\ \dot{\xi} &= \frac{p_\xi}{Q}. \end{aligned} \tag{4.5}$$

Now, we have a dynamical system in a normal form. So, it is clear that the measure

$$\tilde{m} = dq\; d\tilde{p}\; dp_\xi\; d\xi$$

is preserved. When writing $\tilde{m}$ in the original variables, one can check that $\tilde{m} = m_0$. Furthermore, from the normal form, building a reversible measure invariant algorithm is quite easy. Using the Trotter formula :

$$e^{\Delta t(A+B)} = e^{\frac{\Delta t}{2}B}\; e^{\Delta t A}\; e^{\frac{\Delta t}{2}B} + O(\Delta t^3), \tag{4.6}$$

and using the fact that $[\dot{q}\,\partial_q, \dot{p}_\xi\,\partial_{p_\xi}] = 0$, we get

$$L = L_1 + L_2 + L_3$$

with

$$\begin{aligned} L_1 &= \dot{\xi}\,\partial_\xi, \\ L_2 &= \dot{\tilde{p}}\,\partial_{\tilde{p}}, \\ L_3 &= \dot{q}\,\partial_q + \dot{p}_\xi\,\partial_{p_\xi}. \end{aligned}$$

We can now generate 3! different algorithms of the form

$$e^{\Delta t\, L_{eff}} = e^{\frac{\Delta t}{2}L_a}\; e^{\frac{\Delta t}{2}L_b}\; e^{\Delta t L_c}\; e^{\frac{\Delta t}{2}L_b}\; e^{\frac{\Delta t}{2}L_a} \tag{4.7}$$

with $\{a, b, c\} = \{1, 2, 3\}$. All these algorithms are measure invariant. From Eq. (4.6), we deduce

$$e^{\Delta t\, L_{eff}} = e^{\Delta t\, L} + O(\Delta t^3). \tag{4.8}$$

With variables $(q, \tilde{p}, p_\xi, \xi)$, the energy reads

$$\tilde{H}' = \frac{\tilde{p}^2}{2m}\, e^{-2\xi} + V(q) + \frac{p_\xi^2}{2Q} + \xi k_B T.$$

It is preserved up to the order 2 in Δt. For instance, the particular choice $(a, b, c) = (1, 2, 3)$ in Eq. (4.7) gives the following algorithm :

$$(1) \quad \xi \to \xi + \frac{1}{2}\Delta t \, \frac{p_\xi}{Q},$$

$$(2) \quad \tilde{p} \to \tilde{p} + \frac{1}{2}\Delta t \, e^{\xi} F(q),$$

$$(3) \quad \begin{cases} q & \to & q + \Delta t \, \dfrac{\tilde{p}}{m} e^{-\xi}, \\ p_\xi & \to & p_\xi + \Delta t \left(\dfrac{\tilde{p}^2}{m} e^{-2\xi} - k_B T \right), \end{cases}$$

$$(2) \quad \tilde{p} \to \tilde{p} + \frac{1}{2}\Delta t \, e^{\xi} F(q),$$

$$(1) \quad \xi \to \xi + \frac{1}{2}\Delta t \, \frac{p_\xi}{Q}.$$

So, the dynamics given by Eqs. (4.5) is exactly the same as the dynamics given by Eqs. (4.3), provided that p and $\tilde{p}$ are linked according to Eq. (4.4). However, working with new variables allows us to derive measure invariant algorithms in a quite simple way. The numerical properties of these algorithms are studied in the following parts on some examples.

Finally, let us notice that more sophisticated algorithms than the one given in Eq. (4.7) are also possible. For instance, we can write

$$e^{\Delta t \, L_{eff}} = e^{\frac{\Delta t}{4} L_a} \, e^{\frac{\Delta t}{2} L_b} \, e^{\frac{\Delta t}{4} L_a} \, e^{\Delta t L_c} \, e^{\frac{\Delta t}{4} L_a} \, e^{\frac{\Delta t}{2} L_b} \, e^{\frac{\Delta t}{4} L_a}.$$

It is also possible to use a Yoshida-Suzuki like decomposition [37, 60, 68]. Thanks to these more complex decompositions, energy conservation properties are probably better.

4.4 Generalized Gaussian Moment Thermostatting

4.4.1 Normal form for the GGMT dynamics

Let us now take the example of the Generalized Gaussian Moment Thermostatting (GGMT) equations. In this case, transforming the system into its normal form is more complex than in the simple example described in the beginning of the article.

We simulate N particules in d dimensions. The temperature is T, and the forces are $\mathbf{F}_i(\mathbf{q}) = -\dfrac{\partial V}{\partial \mathbf{q}_i}(\mathbf{q})$. The GGMT dynamics is :

$$\begin{aligned}
\dot{\mathbf{q}}_i &= \frac{\mathbf{p}_i}{m_i}, & 1 \leq i \leq N, \\
\dot{\mathbf{p}}_i &= \mathbf{F}_i(\mathbf{q}) - \left(\sum_{k=1}^{M} R_k \right) \mathbf{p}_i, & 1 \leq i \leq N, \\
\dot{p}_{\xi_k} &= \frac{S^k}{C_{k-1}} - dN \left(k_B T\right)^k, & 1 \leq k \leq M.
\end{aligned} \tag{4.9}$$

The constants $(C_k)_{k \geq -1}$ are given by

$$C_{-1} = \frac{1}{dN} \quad , \qquad C_0 = 1 \quad , \qquad C_k = \prod_{j=1}^{k} (dN + 2j).$$

We also set

$$\begin{aligned}
S &= \sum_{i=1}^{N} \frac{\mathbf{p}_i^2}{m_i}, \\
R_k &= \frac{S^{k-1}}{C_{k-1}} \sum_{j=k}^{M} \frac{p_{\xi_j}}{Q_j} \left(k_B T\right)^{j-k}, & 1 \leq k \leq M.
\end{aligned}$$

With the help of some M additionnal variables η_k, one can find [65] a preserved energy and a preserved measure m_0.

The previous dynamical system is clearly not a system in a normal form, since $\dot{\mathbf{p}}_i$ depends on $\mathbf{p}_i$. In order to transform it into a normal form system, we add M variables ξ_k, whose dynamics are given by

$$\dot{\xi}_k = \mathcal{F}_k \left[(\mathbf{q}_i)_{i=1,N} , (\mathbf{p}_i)_{i=1,N} , (p_{\xi_k})_{k=1,M} , (\xi_k)_{k=1,M} \right],$$

where $\mathcal{F}_k$ is a function we are going to make precise later on. We also need to take some initial conditions on these additional variables ξ_k. With all these data, we have a new dynamical system, written with variables $(\mathbf{q}_i, \mathbf{p}_i, p_{\xi_k}, \xi_k)$. This system is not in a normal form. We are going to transform it, finding new variables, in order for the new system to be in a normal form. Actually, the new variables will be $(\mathbf{q}_i, \tilde{\mathbf{p}}_i, p_{\xi_k}, \xi_k)$. So $\mathbf{p}_i$ is going to be replaced by $\tilde{\mathbf{p}}_i$.

Let us choose the following dynamics for ξ_k :

$$\begin{aligned}
\dot{\xi}_1 &= \tilde{R}_1, \\
\dot{\xi}_k &= e^{-2(k-1)\left(\xi_1 + \sum_{j \neq k, j=2,\ldots,M} \frac{1}{2(j-1)} \ln \xi_j\right)} \tilde{R}_k, & 2 \leq k \leq M.
\end{aligned} \tag{4.10}$$

The expressions of the $\tilde{R}_k$ are similar to the expressions of the R_k :

$$\begin{aligned}\tilde{R}_1 &= \sum_{j=1}^{M} \frac{p_{\xi_j}}{Q_j} (k_B T)^{j-1}, \\ \tilde{R}_k &= 2(k-1) \frac{\tilde{S}^{k-1}}{C_{k-1}} \sum_{j=k}^{M} \frac{p_{\xi_j}}{Q_j} (k_B T)^{j-k}, \qquad 2 \leq k \leq M,\end{aligned}$$

where $\tilde{S} = \sum_{i=1}^{N} \frac{\tilde{\mathbf{p}}_i^2}{m_i}$. So the quantity $\tilde{S}$ is the same as S, except that it is defined from variables $\tilde{\mathbf{p}}_i$ instead of variables $\mathbf{p}_i$. The new impulsions are linked to the old ones by

$$\tilde{\mathbf{p}}_i = e^{\chi} \mathbf{p}_i,$$

with

$$\begin{aligned}\chi_1 &= \xi_1, \\ \chi_k &= \frac{1}{2(k-1)} \ln \xi_k, \qquad 2 \leq k \leq M, \\ \chi &= \textstyle\sum_{k=1}^{M} \chi_k.\end{aligned} \tag{4.11}$$

Let us notice that it is possible to rewrite the dynamics on ξ_k for $k \geq 2$ as

$$\dot{\xi}_k = e^{-2(k-1)\left(\sum_{j \neq k, j=1,\ldots,M} \chi_j\right)} \tilde{R}_k.$$

From this equation, we can deduce that the dynamics of the M variables χ_k reads $\dot{\chi}_k = R_k$.

Let us now explain the transformation. In the initial dynamical system (4.9), $\dot{\mathbf{p}}_i$ depends on $\mathbf{p}_i$. It is thus natural to add the M variables χ_k defined by $\dot{\chi}_k = R_k$, and to transform $\mathbf{p}_i$ into $\tilde{\mathbf{p}}_i$ according to $\tilde{\mathbf{p}}_i = e^{\chi_1 + \ldots + \chi_M} \mathbf{p}_i$. We want now to write the complete dynamical system only using variables

$$\left((\mathbf{q}_i)_{i=1,N}, (\tilde{\mathbf{p}}_i)_{i=1,N}, (p_{\xi_k})_{k=1,M}, (\chi_k)_{k=1,M}\right).$$

Reminding χ is the sum of the χ_k, we get

$$\begin{aligned}\dot{\mathbf{q}}_i &= \frac{\tilde{\mathbf{p}}_i}{m_i} e^{-\chi}, && 1 \leq i \leq N, \\ \dot{\tilde{\mathbf{p}}}_i &= e^{\chi} F_i(\mathbf{q}), && 1 \leq i \leq N, \\ \dot{p}_{\xi_k} &= \frac{\tilde{S}^k}{C_{k-1}} e^{-2k\chi} - dN (k_B T)^k, && 1 \leq k \leq M,\end{aligned} \tag{4.12}$$

and

$$\dot{\chi}_k = \frac{\tilde{S}^{k-1}}{C_{k-1}} e^{-2(k-1)\chi} \sum_{j=k}^{M} \frac{p_{\xi_j}}{Q_j} (k_B T)^{j-k}, \qquad 1 \leq k \leq M.$$

So the system is not yet in a normal form, since $\dot{\chi}_k$ depends on χ_k (for $k \geq 2$). That is why we need to go from the χ_k to the ξ_k, defined as $\xi_k = e^{2(k-1)\,\chi_k}$ (for $k \geq 2$). From this definition, it is possible to write the dynamics on ξ_k, which is exactly what we announced in Eq. (4.10). Then the dynamics for $(\mathbf{q}_i, \tilde{\mathbf{p}}_i, p_{\xi_k}, \xi_k)$ is given by the system (4.10)-(4.11)-(4.12). One can check that this set of differential equations is now a normal form system. It clearly preserves the measure

$$\tilde{m} = d^N\tilde{\mathbf{p}}\; d^N\mathbf{q}\; d^M p_\xi\; d^M\xi.$$

If we write the measure $\tilde{m}$ with the original variables, we can check that $\tilde{m} = m_0$, where m_0 is the conserved measure given by Liu and Tuckerman [65]. In addition, the energy

$$\begin{aligned}\tilde{H}' = &\sum_{i=1}^{N} \frac{\tilde{\mathbf{p}}_i^2}{2m_i}\, e^{-2\chi} + V(\mathbf{q}) + \sum_{k=1}^{M} \frac{p_{\xi_k}^2}{2Q_k} \\ &+ k_B T \left(\frac{C_0}{C_{-1}}\,\xi_1 + \frac{C_1}{C_0}\,\frac{\ln \xi_2}{2} + \frac{C_2}{C_1}\,\frac{\ln \xi_3}{2\times 2} + \ldots + \frac{C_{M-1}}{C_{M-2}}\,\frac{\ln \xi_M}{2(M-1)} \right) \end{aligned} \quad (4.13)$$

is preserved. A proof of this conservation is given in appendix (see Sec. 4.5.3).

4.4.2 Numerical results

On their web site [1], Frenkel and Smit propose many molecular dynamics codes, including sources. To get the results we present in this article, we worked from one of these codes.

We now want to compare the algorithms given by our method with algorithms given by Liu and Tuckerman [65]. We choose to simulate one particle ($N = 1$), in one dimension ($d = 1$), in the quartic double well potential given by :

$$V(q) = D_0\,(a^2 - q^2)^2.$$

Usually, molecular dynamics algorithms are first tested with the harmonic potential. However, when using this potential, the invariant measure is a Gaussian function. We want to test algorithms given by our method on more demanding potentials. That is why we choose a double quartic well potential. We work with $D_0 = 1$, $a = 1.5$, and $k_B T = 1$. So, the barrier height is close to $5\,k_B T$. Initial conditions are $q(0) = 0$ and $p(0) = 1$. Thus, at the beginning of the simulation, the particle is in a non-equilibrium position, and goes toward the right well. In order to control the temperature, we use two thermostats ($M = 2$). Equations of motion in this case are given in appendix (see Sec. 4.5.2). We set the masses Q_1 and Q_2 of the thermostats according to adviced values [65], so $Q_1 = 1$ and $Q_2 = 8/3$.

To generate trajectories, two algorithms have been used. The first one is a reversible measure invariant algorithm. It is given in appendix (see Sec. 4.5.2). Generally speaking, we have focused on measure invariance, and not on energy conservation.

So this algorithm is a simple one rather than a sophisticated one (see the end of Sec. 4.3 for more details on this distinction).

The other algorithm is the one given by Liu and Tuckerman [65]. The authors present a general algorithm using a Yoshida-Suzuki decomposition [37, 60]. We work with $n_c = n_{sy} = 1$. We have made this choice in order to compare our algorithm, which is not really improved in term of energy conservation, with an algorithm having the same feature. Of course, it is also possible to use a Yoshida-Suzuki decomposition on both algorithms.

For the Liu and Tuckerman algorithm, initial conditions for thermostat variables are $\eta_1(0) = \eta_2(0) = 0$ and $p_{\xi_1}(0) = -p_{\xi_2}(0) = 1$. For the measure invariant algorithm, initial conditions are $\xi_1(0) = 0$, $\xi_2(0) = 1$ and $p_{\xi_1}(0) = -p_{\xi_2}(0) = 1$. Thus, the energy has the same initial value in both simulations. With both algorithms, we generate trajectories of length $2.5\,10^6$ steps, using a time step of $\Delta t = 0.001$.

With this dynamics, analytical position and impulsion distribution functions are known. As the potential is symmetric, the particle spends equal amounts of time into both wells (let us notice that the particle has enough energy to cross the barrier). So the analytical position distribution function is symmetric. From a numerical point of view, getting the proper function is a challenge.

Results on position distribution functions are given in Fig. 4.1. One can see the functions generated by the algorithms, as well as the analytical solution. No algorithm gives a perfectly symmetrical function. However, the asymmetry is lower for the measure invariant algorithm.

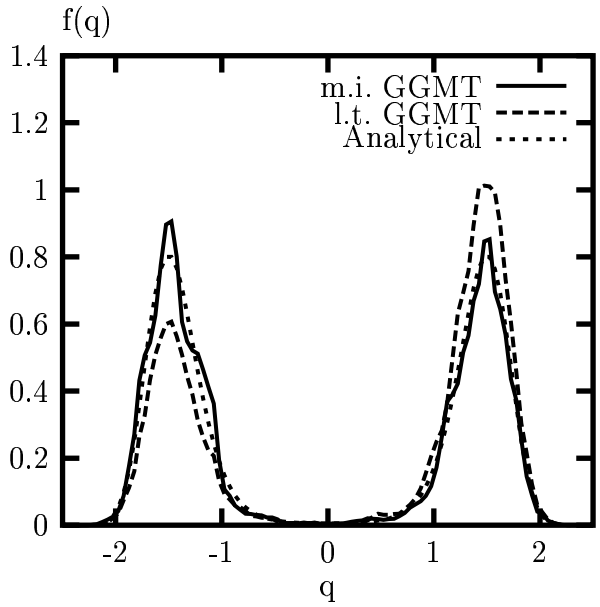

FIG. 4.1 – Position distribution functions for the double quartic well potential generated by measure invariant GGMT algorithm (solid line) and by Liu and Tuckerman GGMT algorithm (long dashed line), compared with the analytical result (short dashed line).

In Fig. 4.2, we plot the quantity $< f(q) - f_{exact}(q) > (t)$ as a function of time. This quantity is an estimator of the difference between the analytical distribution function and the calculated function, and it is defined [65] by

$$< f(q) - f_{exact}(q) > (t) = \frac{1}{\mathcal{N}(t)} \sum_{i=1}^{\mathcal{N}(t)} \left| \overline{f}_t(q_i) - f_{exact}(q_i) \right| .$$

At time t, $\mathcal{N}(t)$ bins have been generated. With these bins, it is possible to calculate a distribution function, which is $\overline{f}_t$. We can see that the distribution function generated by the measure invariant algorithm converges more quickly to the proper one. So, in term of position distribution function, the measure invariant algorithm gives better results.

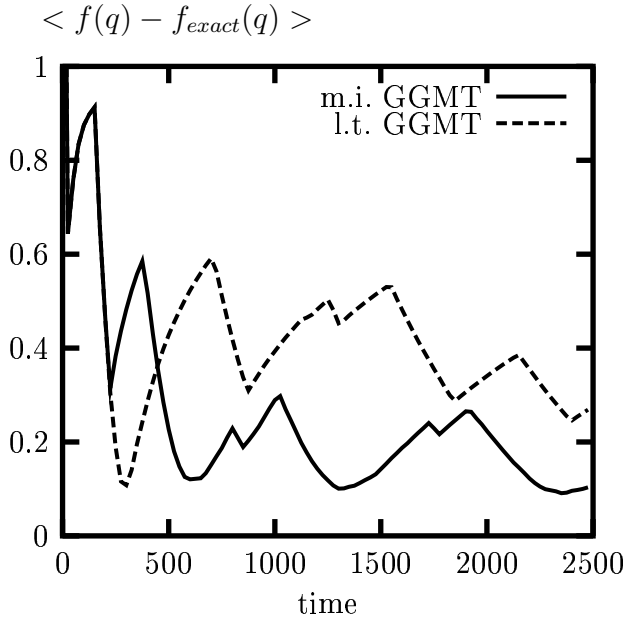

FIG. 4.2 – Convergence of the position distribution functions for measure invariant GGMT algorithm (solid line) and for Liu and Tuckerman GGMT algorithm (dashed line) for the double quartic well potential.

We show on Fig. 4.3 the impulsion distribution functions generated by the algorithms, as well as the analytical solution.

The Liu and Tuckerman algorithm function is closer to the analytical solution. However, when one looks at the convergence of the calculated distribution function to the exact one, algorithms performances are similar. The convergence can be estimated by many ways. We can look at the quantity $< f(p) - f_{exact}(p) >$ (cf. Fig. 4.4), but also at the moments of the distribution function. The second moment (which is linked to the temperature) is given on Fig. 4.5, whereas the fourth moment is given on Fig. 4.6. Obviously, both algorithms have the same performances.

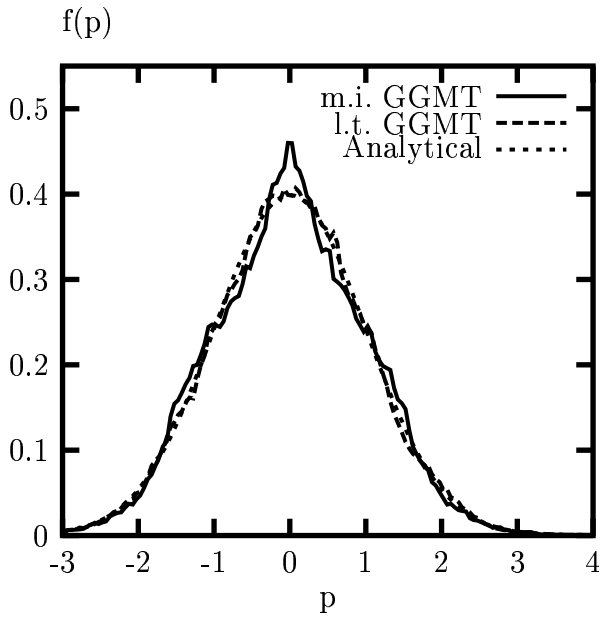

FIG. 4.3 – Impulsion distribution functions for the double quartic well potential generated by measure invariant GGMT algorithm (solid line) and by Liu and Tuckerman GGMT algorithm (long dashed line), compared with the analytical result (short dashed line).

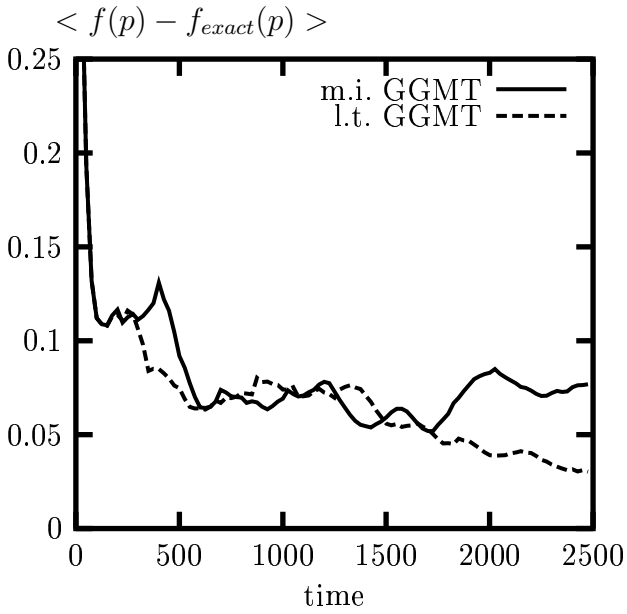

FIG. 4.4 – Convergence of the impulsion distribution functions for measure invariant GGMT algorithm (solid line) and for Liu and Tuckerman GGMT algorithm (dashed line) for the double quartic well potential.

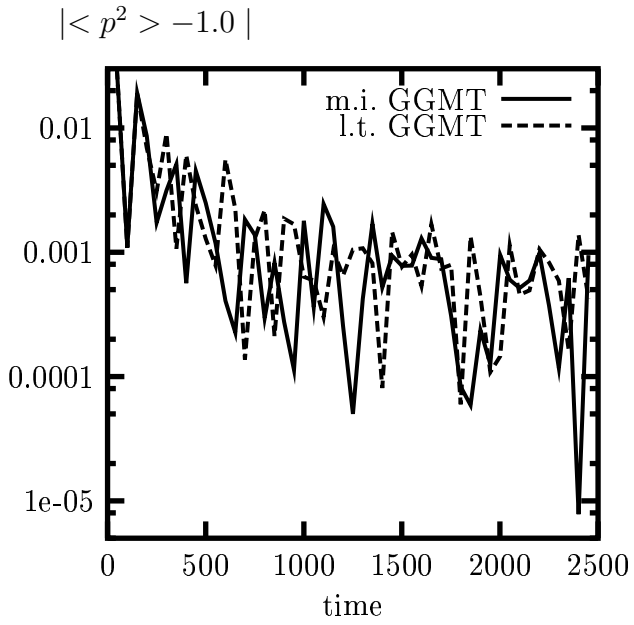

FIG. 4.5 – Convergence of the second moment of the impulsion distribution functions for measure invariant GGMT algorithm (solid line) and for Liu and Tuckerman GGMT algorithm (dashed line) for the double quartic well potential.

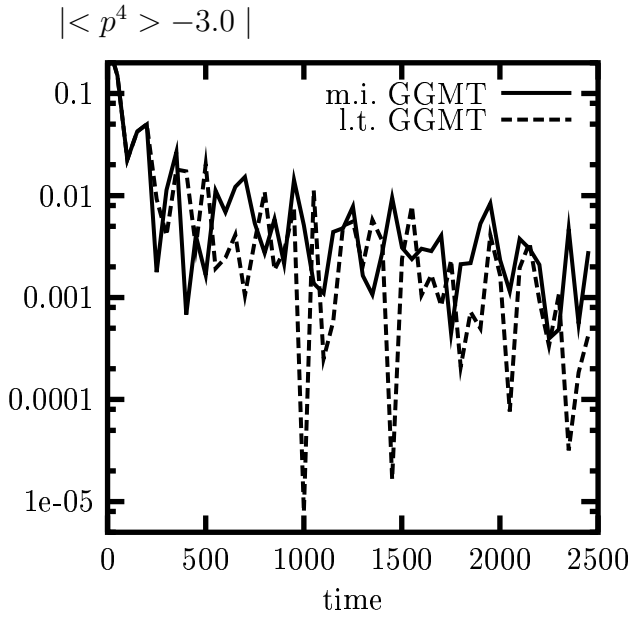

FIG. 4.6 – Convergence of the fourth moment of the impulsion distribution functions for measure invariant GGMT algorithm (solid line) and for Liu and Tuckerman GGMT algorithm (dashed line) for the double quartic well potential.

Finally, let us give results on the conservation of the energy. The expression of the conserved energy is given in Eq. (4.13) when using the new variables. Let us underline the fact that we have chosen initial conditions for both simulations so that the initial values of the energy are the same : $\tilde{H}'_{mi}(t=0) = H'_{lt}(t=0) = 6.25$. Results are given in Fig. 4.7. The Liu and Tuckerman algorithm better preserves the energy. At all times, its numerical energy is close to the correct value. Our algorithm quite well preserves the energy for times $t \leq 1800$. At that moment, there is a sudden change. For times $t \geq 1800$, the numerical energy keeps constant, but at another value. However, one can notice that this change has no consequence on the quality of the distribution functions generated by our algorithm. We can see this kind of "shock" neither on the position distribution function (see Fig. 4.2) nor on the impulsion functions (see Figs. 4.4, 4.5 and 4.6). We think that it can be possible to improve our algorithm by changing the Trotter decomposition order.

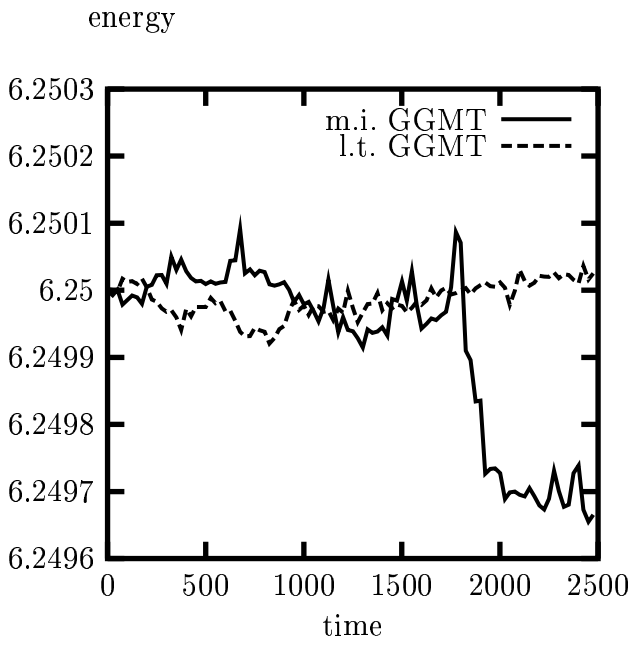

FIG. 4.7 – Evolution of the numerical energy for measure invariant GGMT algorithm (solid line) and for Liu and Tuckerman GGMT algorithm (dashed line) for the double quartic well potential.

Acknowledgments

We would like to thank Eric Cancès and Philippe Chartier for stimulating discussions and useful indications of the literature. We would also like to thank the referee for his suggestions.

4.5 Appendix

4.5.1 Nosé-Hoover chains for NVT and NPT ensembles

In this part, we want to show how to proceed to find equations in the normal form from the NPT equations, that simulate particles at constant temperature and constant pressure. So, we study N particles in d dimensions, that are coupled to M thermostats and one barostat. Let N_f be the number of degrees of freedom of the system that we have to thermostate. For the NPT equations, given below, $N_f = dN + 1$.

This formalism also includes the NVT ensemble. To go back to it, we first need to set $N_f = dN$. We need also to set $1/W = 0$, so V becomes a constant. The NVT ensemble equations only involve $(\mathbf{q}_i, \mathbf{p}_i, p_{\xi_k}, \xi_k)$.

We consider the following NPT equations [68]. The particles positions are $\mathbf{q}_i$, their impulsion are $\mathbf{p}_i$. They are coupled to a single barostat described by p_ε. The volume V is allowed to fluctuate, but we want the pressure to stay constant at P_{ext}. The barostat and the particles are coupled to the same chain of thermostats, described by p_{ξ_k}, in order for the simulation to run at constant temperature T (let us notice that slightly different dynamics are possible, that also generate the NPT ensemble).

$$\begin{aligned}
\dot{\mathbf{q}}_i &= \frac{\mathbf{p}_i}{m_i} + \frac{p_\varepsilon}{W}\,\mathbf{q}_i, && 1 \le i \le N,\\
\dot{\mathbf{p}}_i &= \mathbf{F}_i - \frac{p_{\xi_1}}{Q_1}\,\mathbf{p}_i - \left(1 + \frac{1}{N}\right)\frac{p_\varepsilon}{W}\,\mathbf{p}_i, && 1 \le i \le N,\\
\dot{p}_{\xi_1} &= \sum_{i=1}^{N} \frac{\mathbf{p}_i^2}{m_i} + \frac{p_\varepsilon^2}{W} - N_f k_B T - \frac{p_{\xi_2}}{Q_2}\,p_{\xi_1},\\
\dot{p}_{\xi_k} &= \frac{p_{\xi_{k-1}}^2}{Q_{k-1}} - k_B T - \frac{p_{\xi_{k+1}}}{Q_{k+1}}\,p_{\xi_k}, && 2 \le k \le M-1,\\
\dot{p}_{\xi_M} &= \frac{p_{\xi_{M-1}}^2}{Q_{M-1}} - k_B T,\\
\dot{\xi}_k &= \frac{p_{\xi_k}}{Q_k}, && 1 \le k \le M,\\
\dot{V} &= \frac{dV p_\varepsilon}{W},\\
\dot{p}_\varepsilon &= dV\,(P_{int} - P_{ext}) \;+\; \frac{1}{N}\sum_{i=1}^{N}\frac{\mathbf{p}_i^2}{m_i} - \frac{p_{\xi_1}}{Q_1}\,p_\varepsilon.
\end{aligned} \tag{4.14}$$

Internal pressure and forces are defined by

$$P_{int} = \frac{1}{dV}\left(\sum_{i=1}^{N}\frac{\mathbf{p}_i^2}{m_i} + \sum_{i=1}^{N}\mathbf{q}_i\cdot\mathbf{F}_i - dV\,\frac{\partial\phi(\mathbf{q},V)}{\partial V}\right) \tag{4.15}$$

and $\mathbf{F}_i = -\dfrac{\partial \phi(\mathbf{q}, V)}{\partial \mathbf{q}_i}$. The system (4.14) preserves [57] a measure m_0 and an energy H'.

The system (4.14), written in variables $(\mathbf{q}_i, \mathbf{p}_i, p_{\xi_k}, \xi_k, V, p_\varepsilon)$, is obviously not in a normal form. We are going to change variables. The new ones are $(\tilde{\mathbf{q}}_i, \tilde{\mathbf{p}}_i, \tilde{p}_{\xi_k}, \xi_k, \tilde{V}, \tilde{p}_\varepsilon)$. We set

$$\begin{aligned}
\tilde{\mathbf{q}}_i &= V^{-\frac{1}{d}}\ \mathbf{q}_i, && 1 \le i \le N,\\
\tilde{\mathbf{p}}_i &= V^{\frac{N+1}{dN}}\ e^{\xi_1}\ \mathbf{p}_i, && 1 \le i \le N,\\
\tilde{p}_{\xi_k} &= e^{\xi_{k+1}}\ p_{\xi_k}, && 1 \le k \le M-1,\\
\tilde{p}_{\xi_M} &= p_{\xi_M},\\
\tilde{V} &= \ln V,\\
\tilde{p}_\varepsilon &= e^{\xi_1}\ p_\varepsilon.
\end{aligned} \tag{4.16}$$

With these new variables, the dynamical system can be rewritten as

$$\begin{aligned}
\dot{\tilde{\mathbf{q}}}_i &= \frac{\tilde{\mathbf{p}}_i}{m_i}\ e^{-\xi_1}\ e^{-\tilde{V}\left(\frac{1}{dN}+\frac{2}{d}\right)}, && 1 \le i \le N,\\
\dot{\tilde{\mathbf{p}}}_i &= e^{\tilde{V}\frac{N+1}{dN}}\ e^{\xi_1}\ \mathbf{F}_i, && 1 \le i \le N,\\
\dot{\tilde{p}}_{\xi_1} &= e^{\xi_2}\left[e^{-2\xi_1}\ e^{-2\tilde{V}\frac{N+1}{dN}}\left(\sum_{i=1}^{N}\frac{\tilde{\mathbf{p}}_i^2}{m_i}\right) + e^{-2\xi_1}\ \frac{\tilde{p}_\varepsilon^2}{W} - N_f k_B T\right],\\
\dot{\tilde{p}}_{\xi_k} &= e^{\xi_{k+1}}\left(\frac{\tilde{p}_{\xi_{k-1}}^2}{Q_{k-1}}\ e^{-2\xi_k} - k_B T\right), && 2 \le k \le M-1,\\
\dot{\tilde{p}}_{\xi_M} &= \frac{\tilde{p}_{\xi_{M-1}}^2}{Q_{M-1}}\ e^{-2\xi_M} - k_B T,\\
\dot{\xi}_k &= \frac{\tilde{p}_{\xi_k}}{Q_k}\ e^{-\xi_{k+1}}, && 1 \le k \le M-1,\\
\dot{\xi}_M &= \frac{\tilde{p}_{\xi_M}}{Q_M},\\
\dot{\tilde{V}} &= \frac{d}{W}\ e^{-\xi_1}\tilde{p}_\varepsilon,\\
\dot{\tilde{p}}_\varepsilon &= e^{\xi_1}\left[d\ e^{\tilde{V}}\left(P_{int} - P_{ext}\right) + \frac{1}{N}\ e^{-2\xi_1}\ e^{-2\tilde{V}\frac{N+1}{dN}}\sum_{i=1}^{N}\frac{\tilde{\mathbf{p}}_i^2}{m_i}\right].
\end{aligned} \tag{4.17}$$

In the last equation, P_{int} is defined as in (4.15), but has to be written with variables introduced in (4.16). The important thing is that we can check by (4.15) that P_{int} does not depend on $\tilde{p}_\varepsilon$.

It is straightforward to notice that this last system of equations is in a normal

form. So the measure

$$\tilde{m} = d^N\tilde{\mathbf{p}}\; d^N\tilde{\mathbf{q}}\; d^M\tilde{p}_\xi\; d^M\xi\; d\tilde{p}_\varepsilon\; d\tilde{V}$$

is conserved. Thanks to (4.16), we can express $\tilde{m}$ with the original variables, and check that we find the same measure as the one already known [57]. With the new variables, the energy reads :

$$\begin{aligned}\tilde{H}' &= e^{-2\xi_1}\, e^{-2\tilde{V}\frac{N+1}{dN}} \sum_{i=1}^{N} \frac{\tilde{\mathbf{p}}_i^2}{2m_i} \;+\; \phi(e^{\tilde{V}/d}\,\tilde{\mathbf{q}}, e^{\tilde{V}}) \;+\; \sum_{k=1}^{M-1} \frac{\tilde{p}_{\xi_k}^2}{2Q_k}\, e^{-2\xi_{k+1}} \\ &+ \frac{\tilde{p}_{\xi_M}^2}{2Q_M} + e^{-2\xi_1}\,\frac{\tilde{p}_\varepsilon^2}{2W} \;+\; N_f\, k_B T\, \xi_1 + k_B T \sum_{k=2}^{M} \xi_k \;+\; P_{ext}\, e^{\tilde{V}}.\end{aligned}$$

From the system (4.17), it is straightforward to generate an algorithm. In the particular case $M = 2$, the classical Liouville propagator can be factorized as

$$L = L_1 + L_2 + L_3 + L_4 + L_5 + L_6 + L_7 + L_8$$

with

$$\begin{aligned} L_1 &= \dot{\xi}_1\, \partial_{\xi_1}, & L_2 &= \dot{\xi}_2\, \partial_{\xi_2}, & L_3 &= \dot{\tilde{p}}_{\xi_1}\, \partial_{\tilde{p}_{\xi_1}}, \\ L_4 &= \dot{\tilde{p}}_{\xi_2}\, \partial_{\tilde{p}_{\xi_2}}, & L_5 &= \textstyle\sum_{i=1}^{N} \dot{\tilde{\mathbf{p}}}_i\, \partial_{\tilde{\mathbf{p}}_i}, & L_6 &= \textstyle\sum_{i=1}^{N} \dot{\tilde{\mathbf{q}}}_i\, \partial_{\tilde{\mathbf{q}}_i}, \\ L_7 &= \dot{\tilde{V}}\, \partial_{\tilde{V}}, & L_8 &= \dot{\tilde{p}}_\varepsilon\, \partial_{\tilde{p}_\varepsilon}. \end{aligned}$$

We took advantage of the fact that the operators $\dot{\tilde{\mathbf{p}}}_i\, \partial_{\tilde{\mathbf{p}}_i}$ commute one with each other, as well as the operators $\dot{\tilde{\mathbf{q}}}_i\, \partial_{\tilde{\mathbf{q}}_i}$. From then on, we go on as in the first example. We expand $e^{\Delta t\, L}$ as in equations (4.7) - (4.8), or in a more sophisticated way to better preserve the energy.

4.5.2 GGMT dynamics, Case $M = 2, N = 1, d = 1$

In section 4.4, we present numerical results for the simulation of one particle coupled to a chain of two thermostats. In the following lines, we are going to detail the algorithm we implemented. Let us first rewrite the system (4.10)-(4.12) in this

particular case. We have $e^{\chi} = \sqrt{\xi_2}\, e^{\xi_1}$. The dynamics for $(q, \tilde{p}, p_{\xi_1}, p_{\xi_2}, \xi_1, \xi_2)$ is

$$\begin{aligned}
\dot{q} &= \frac{\tilde{p}}{m}\,\frac{e^{-\xi_1}}{\sqrt{\xi_2}}, \\
\dot{\tilde{p}} &= \sqrt{\xi_2}\, e^{\xi_1}\, F(q), \\
\dot{p_{\xi_1}} &= \frac{\tilde{p}^2}{m}\,\frac{e^{-2\xi_1}}{\xi_2} - k_B T, \\
\dot{p_{\xi_2}} &= \frac{\tilde{p}^4}{3m^2}\,\frac{e^{-4\xi_1}}{\xi_2^2} - (k_B T)^2, \\
\dot{\xi}_1 &= \frac{p_{\xi_1}}{Q_1} + \frac{p_{\xi_2}}{Q_2}\, k_B T, \\
\dot{\xi}_2 &= \frac{2}{3}\, e^{-2\xi_1}\,\frac{\tilde{p}^2}{m}\,\frac{p_{\xi_2}}{Q_2},
\end{aligned} \tag{4.18}$$

and the energy reads

$$\tilde{H}' = \frac{\tilde{p}^2}{2m}\,\frac{e^{-2\xi_1}}{\xi_2} + V(q) + \frac{p_{\xi_1}^2}{2Q_1} + \frac{p_{\xi_2}^2}{2Q_2} + k_B T\left(\xi_1 + \frac{3}{2}\,\ln \xi_2\right).$$

The preserved measure is

$$\tilde{m} = dq\; dp\; dp_{\xi_1}\; dp_{\xi_2}\; d\xi_1 d\xi_2.$$

Let

$$\begin{aligned}
L_1 &= \dot{\xi}_1\, \partial_{\xi_1}, \qquad & L_2 &= \dot{\xi}_2\, \partial_{\xi_2} \\
L_3 &= \dot{\tilde{p}}\, \partial_{\tilde{p}}, \qquad & L_4 &= \dot{q}\, \partial_q + \dot{p}_{\xi_1}\, \partial_{p_{\xi_1}} + \dot{p}_{\xi_2}\, \partial_{p_{\xi_2}}.
\end{aligned}$$

We can generate 4! algorithms of the form

$$e^{\frac{\Delta t}{2} L_a}\; e^{\frac{\Delta t}{2} L_b}\; e^{\frac{\Delta t}{2} L_c}\; e^{\Delta t L_d}\; e^{\frac{\Delta t}{2} L_c}\; e^{\frac{\Delta t}{2} L_b}\; e^{\frac{\Delta t}{2} L_a} \;=\; e^{\Delta t L_{eff}}$$

with $\{a, b, c, d\} = \{1, 2, 3, 4\}$. We implemented this algorithm with $(a, b, c, d) = (1, 2, 3, 4)$.

4.5.3 Proof of the conservation of the energy for the GGMT dynamics

We want to prove that the energy written in (4.13) is conserved by the dynamics (4.10)-(4.11)-(4.12). Thanks to the relations $\xi_1 = \chi_1$, $\xi_k = e^{2(k-1)\chi_k}$ for $k \geq 2$, we

can first rewrite the energy as

$$\begin{aligned}\tilde{H}' = & \sum_{i=1}^{N} \frac{\tilde{\mathbf{p}}_i^2}{2m_i} e^{-2\chi} + V(\mathbf{q}) + \sum_{k=1}^{M} \frac{p_{\xi_k}^2}{2Q_k} \\ & + k_B T \left(\frac{C_0}{C_{-1}} \chi_1 + \frac{C_1}{C_0} \chi_2 + \frac{C_2}{C_1} \chi_3 + \ldots + \frac{C_{M-1}}{C_{M-2}} \chi_M \right).\end{aligned}$$

We know [65] that the following energy is preserved :

$$H' = \sum_{i=1}^{N} \frac{\tilde{\mathbf{p}}_i^2}{2m_i} e^{-2\chi} + V(\mathbf{q}) + \sum_{k=1}^{M} \frac{p_{\xi_k}^2}{2Q_k} + dN\, k_B T \sum_{k=1}^{M} \eta_k,$$

where the dynamics on η_k is

$$\dot{\eta}_k = \frac{p_{\xi_k}}{Q_k} \frac{1}{dN} \sum_{j=1}^{k} (k_B T)^{k-j} \frac{S^{j-1}}{C_{j-2}}.$$

We then only have to check that :

$$dN \sum_{k=1}^{M} \dot{\eta}_k = \left(\frac{C_0}{C_{-1}} \dot{\chi}_1 + \frac{C_1}{C_0} \dot{\chi}_2 + \frac{C_2}{C_1} \dot{\chi}_3 + \ldots + \frac{C_{M-1}}{C_{M-2}} \dot{\chi}_M \right). \tag{4.19}$$

We have $\dot{\chi}_k = \sum_{j=k}^{M} \frac{p_{\xi_j}}{Q_j} (k_B T)^{j-k} \frac{S^{k-1}}{C_{k-1}}$. So

$$\begin{aligned}\dot{\chi}_1 &= \frac{p_{\xi_1}}{Q_1} + \frac{p_{\xi_2}}{Q_2}(k_B T) + \frac{p_{\xi_3}}{Q_3}(k_B T)^2 + \ldots + \frac{p_{\xi_M}}{Q_M}(k_B T)^{M-1}, \\ \dot{\chi}_2 &= \frac{p_{\xi_2}}{Q_2}\frac{S}{C_1} + \frac{p_{\xi_3}}{Q_3}\frac{S}{C_1}(k_B T) + \ldots + \frac{p_{\xi_M}}{Q_M}\frac{S}{C_1}(k_B T)^{M-2}, \\ \dot{\chi}_M &= \frac{p_{\xi_M}}{Q_M}\frac{S^{M-1}}{C_{M-1}},\end{aligned}$$

whereas

$$
\begin{aligned}
dN\,\dot{\eta}_1 &= \frac{1}{C_{-1}}\frac{p_{\xi_1}}{Q_1}, \\
dN\,\dot{\eta}_2 &= \frac{1}{C_{-1}}\frac{p_{\xi_2}}{Q_2}\,(k_BT) &&+ \frac{p_{\xi_2}}{Q_2}\frac{S}{C_0}, \\
dN\,\dot{\eta}_3 &= \frac{1}{C_{-1}}\frac{p_{\xi_3}}{Q_3}\,(k_BT)^2 &&+ \frac{p_{\xi_3}}{Q_3}\frac{S}{C_0}\,(k_BT) &&+ \frac{p_{\xi_3}}{Q_3}\frac{S^2}{C_1}, \\
dN\,\dot{\eta}_M &= \frac{1}{C_{-1}}\frac{p_{\xi_M}}{Q_M}\,(k_BT)^{M-1} &&+ \frac{p_{\xi_M}}{Q_M}\frac{S}{C_0}\,(k_BT)^{M-2} &&+ \frac{p_{\xi_M}}{Q_M}\frac{S^2}{C_1}\,(k_BT)^{M-3}, \\
& &&+ \ldots &&+ \frac{p_{\xi_M}}{Q_M}\frac{S^{M-1}}{C_{M-2}}.
\end{aligned}
$$

So the equation (4.19) is true.

4.5.4 Non-exact preservation of the measure for the algorithm Liu *et al.* [65] proposed for GGMT dynamics, case of the free particle

We show here that the algorithm proposed by Liu and Tuckerman [65] does not exactly preserve the measure analytically, in the very special case of a free particle (a more general case is studied in the next section). However, we will see that, for some initial conditions, the preservation, if not exact, is very good. This may explain the good numerical properties that Liu and Tuckerman noticed.

We work in the case $M = 2$ (two thermostats), $N = 1$ and $d = 1$ (one particle in a one-dimensionnal space), and with the variables used by the authors, i.e. $X = (q, p, p_{\xi_1}, p_{\xi_2}, \eta_1, \eta_2)$. We use MAPLE to get explicit formulas when needed.

The GGMT equations :

For the present moment, we recall the GGMT equations and the algorithm proposed by Liu and Tuckerman, without any assumption on the force. The equations

are

$$\begin{aligned}
\dot{q} &= \frac{p}{m}, \\
\dot{p} &= F(q) - \frac{p_{\xi_1}}{Q_1}\, p - \frac{p_{\xi_2}}{Q_2}\left((k_B T)p + \frac{p^3}{3m}\right), \\
\dot{\eta}_1 &= \frac{p_{\xi_1}}{Q_1}, \\
\dot{\eta}_2 &= \left(k_B T + \frac{p^2}{m}\right)\frac{p_{\xi_2}}{Q_2}, \\
\dot{p}_{\xi_1} &= \frac{p^2}{m} - k_B T, \\
\dot{p}_{\xi_2} &= \frac{p^4}{3m^2} - (k_B T)^2.
\end{aligned}$$

They preserve the following measure

$$m_0 = e^{\eta_1+\eta_2}\; dp\; dq\; dp_{\xi_1}\; dp_{\xi_2}\; d\eta_1\; d\eta_2. \tag{4.20}$$

The algorithm proposed by Liu and Tuckerman is the following

$$e^{\Delta t\, L_{eff}} = e^{\frac{\Delta t}{2} L_{GGMT}}\; e^{\frac{\Delta t}{2} F(q)\,\partial_p}\; e^{\Delta t \frac{p}{m}\partial_q}\; e^{\frac{\Delta t}{2} F(q)\,\partial_p}\; e^{\frac{\Delta t}{2} L_{GGMT}}. \tag{4.21}$$

The central part of this decomposition corresponds to the simple Velocity Verlet algorithm. The external operator is

$$\exp\left(\frac{\Delta t}{2}\, L_{GGMT}\right) = A(\Delta t)\; B(\Delta t)\; C(\Delta t)\; B(\Delta t)\; A(\Delta t).$$

The operators $A(\Delta t)$, $B(\Delta t)$ and $C(\Delta t)$ are defined by

$$\begin{aligned}
A(\Delta t) &= \exp\left(\frac{\Delta t}{4}\, G_1(p)\,\partial_{p_{\xi_1}}\right)\; \exp\left(\frac{\Delta t}{4}\, G_2(p)\,\partial_{p_{\xi_2}}\right), \\
B(\Delta t) &= \exp\left(-\frac{\Delta t}{8}\lambda\, p\,\partial_p\right)\; \exp\left(-\frac{\Delta t}{4}\,\frac{p_{\xi_2}}{Q_2}\,\frac{p^3}{3m}\,\partial_p\right)\; \exp\left(-\frac{\Delta t}{8}\lambda\, p\,\partial_p\right), \\
C(\Delta t) &= \exp\left(\frac{\Delta t}{2}\,\frac{p_{\xi_1}}{Q_1}\,\partial_{\eta_1}\right)\; \exp\left(\frac{\Delta t}{2}\, g(p)\,\frac{p_{\xi_2}}{Q_2}\,\partial_{\eta_2}\right).
\end{aligned}$$

We set

$$\begin{aligned}
\lambda(X) &= \frac{p_{\xi_1}}{Q_1} + k_B T \frac{p_{\xi_2}}{Q_2}, \\
g(p) &= k_B T + \frac{p^2}{m}, \\
G_1(p) &= \frac{p^2}{m} - k_B T, \\
G_2(p) &= \frac{p^4}{3m^2} - (k_B T)^2, \\
\mathcal{G}(p) &= \frac{G_1(p)}{Q_1} + k_B T \frac{G_2(p)}{Q_2}, \\
U(X) &= \eta_1 + \eta_2.
\end{aligned}$$

If the vector X' is a function of the vector X, the Jacobian matrix is noted $\dfrac{\partial X'}{\partial X}$. The Jacobian of the transformation is Jac $\left(\dfrac{\partial X'}{\partial X}\right)$.

If the complete algorithm (4.21) preserves the measure m_0 defined by Eq. (4.20), we have

$$m_0\left[X'(X)\right] = m_0\left[X\right].$$

Using the Jacobian of the function $X'(X)$, and the function $U(X)$ already defined, we get

$$\text{Jac}\left(\frac{\partial X'}{\partial X}\right) = e^{U(X)-U(X')}. \tag{4.22}$$

Study in the case $F(q) = 0$:

The equation (4.22) must be true for all parameters m, Q_1, Q_2, T, Δt, and for all initial conditions X. Let us now choose some specific values for some parameters :

$$m = 1.0, \qquad Q_1 = 1.0, \qquad Q_2 = 8/3, \qquad k_B T = 1.0.$$

Then $\mathcal{G}(p) = \dfrac{p^4}{8} + p^2 - \dfrac{11}{8}$. Let μ be the real positive root of $\mathcal{G}$: $\mu = \sqrt{-4 + 3\sqrt{3}}$. We also choose some specific initial conditions $X = (q, p, p_{\xi_1}(p), p_{\xi_2}(p), \eta_1, \eta_2)$. So, in

X, there are only 4 free variables, and we choose p_{ξ_1} and p_{ξ_2} according to

$$\begin{aligned} p_{\xi_2}(p) &= 3\,m\,Q_2\,\frac{p^2-\mu^2}{\Delta t\ \mu^2 p^2} - \frac{\Delta t}{4} G_2(p) \\ &= 8\,\frac{p^2-\mu^2}{\Delta t\ \mu^2 p^2} - \frac{\Delta t}{4}\left(\frac{p^4}{3} - 1.0\right), \\ \text{and} \qquad p_{\xi_1}(p) &= -Q_1\left(\frac{k_B T}{Q_2} p_{\xi_2}(p) + \frac{\Delta t}{4}\mathcal{G}(p)\right) \\ &= -\left(\frac{3}{8} p_{\xi_2}(p) + \frac{\Delta t}{4}\mathcal{G}(p)\right), \end{aligned} \tag{4.23}$$

where μ is the real positive root of $\mathcal{G}(p)$ (we assumed that $\Delta t > 0$). Reasons to make this choice can be found in the next section, in which we give a general proof of the non preservation of the measure.

Using MAPLE, we compute $\eta_1(\Delta t)$ and $\eta_2(\Delta t)$. Working only with $p \geq 0$, we have

$$\eta_1(\Delta t) = -\frac{1}{4}\,\frac{16\,\eta_1\,p^2 - 12\,\eta_1\,p^2\,\sqrt{3} + 12\,p^2 + 48 - 36\,\sqrt{3} - 47\,\Delta t^2\,p^2 + 27\,\Delta t^2\,p^2\,\sqrt{3}}{(-4+3\sqrt{3})\ p^2}$$

and

$$\eta_1 + \eta_2 - \eta_1(\Delta t) - \eta_2(\Delta t) =$$
$$6\,\frac{p^2\ \left(-36\,p^2\,\sqrt{3} + 249\,\Delta t^2\,p^2\,\sqrt{3} - 288\,\sqrt{3} - 431\,\Delta t^2\,p^2 + 48\,p^2 + 516\right)}{(-4+3\,\sqrt{3})\ (-4+3\,\sqrt{3}+p^2)\ (18\,p^2 + 24 - 18\,\sqrt{3} - 47\,\Delta t^2\,p^2 + 27\,\Delta t^2\,p^2\,\sqrt{3})}$$

where η_1 stands for $\eta_1(0)$ and η_2 for $\eta_2(0)$.

We can also compute

$$\text{Jac}\left(\frac{\partial X'}{\partial X}\right) = 6\,\frac{\left(-4+3\,\sqrt{3}\right)^{3/2}\ \sqrt{3}\ \sqrt{2}}{\left(12\,p^2 + 24 - 18\,\sqrt{3} - 47\,\Delta t^2\,p^2 + 27\,\Delta t^2\,p^2\,\sqrt{3}\right)^{3/2}}.$$

If $\Delta t = 0$, we choose $p = \mu$, so initial values for p_{ξ_1} and p_{ξ_2} are well defined. With these values, one can check that $\eta_1+\eta_2-\eta_1(\Delta t)-\eta_2(\Delta t) = 0$ and that Jac $\left(\frac{\partial X'}{\partial X}\right) = 1$.

Now we have to check whether Jac $\left(\frac{\partial X'}{\partial X}\right)$ is equal to $e^{\eta_1+\eta_2-\eta_1(\Delta t)-\eta_2(\Delta t)}$. We notice that these two functions depend only on p and Δt, and not on q, η_1 or η_2. In Fig. 4.8, we plot the ratio $\text{Jac}\left(\frac{\partial X'}{\partial X}\right) / e^{\eta_1+\eta_2-\eta_1(\Delta t)-\eta_2(\Delta t)}$ for $\Delta t = 0.001$ and for $p \in [2.0; 5.0]$, and we compare it with 1.

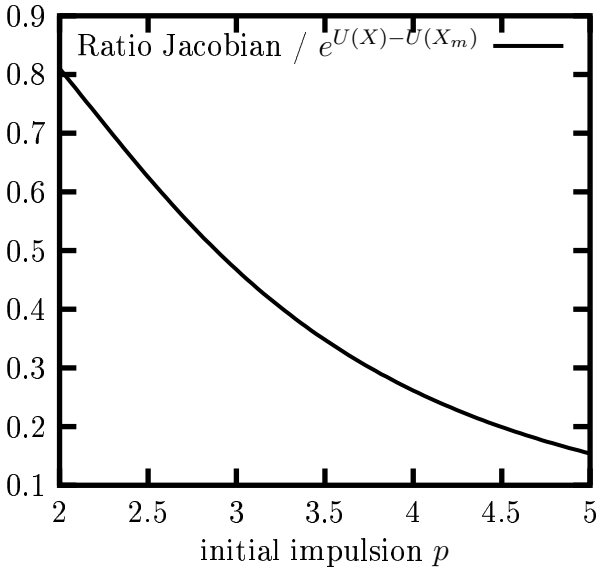

FIG. 4.8 – Checking the preservation of the measure in the case of the free particle, for $p \in [2.0; 5.0]$: we compare $\mathrm{Jac}\left(\frac{\partial X'}{\partial X}\right)(p)$ with $e^{\eta_1+\eta_2-\eta_1(\Delta t)-\eta_2(\Delta t)}(p)$ by plotting their ratio.

We clearly see that, for some values of the impulsion p, the ratio is different from 1. However, in Fig. 4.9, we plot the same ratio for $p \in [0.9; 1.1]$ (with the same value for Δt) :

Functions are close to each other. We can check that their values for $p = 1$ are really close :

$$\mathrm{Jac}\left(\frac{\partial X'}{\partial X}\right)(p=1) \;=\; 1.81517,$$

$$e^{\eta_1+\eta_2-\eta_1(\Delta t)-\eta_2(\Delta t)}(p=1) \;=\; 1.81153.$$

When $p = 1$, the initial condition is $X = (q, 1, 491.96, -1311.89, \eta_1, \eta_2)$.

So, for some initial conditions, the measure is very well preserved. However, for some other initial conditions, it is not at all preserved.

4.5.5 Non-exact preservation of the measure for the algorithm Liu *et al.* [65] proposed for GGMT dynamics, a more general case

In the previous section, we study the very special case of a free particle, $F(q) = 0$. Now, we have a look at a more general case : we suppose that the force $F(q)$ is a polynomial function of q, whose degree is odd (for instance, this is the force given by

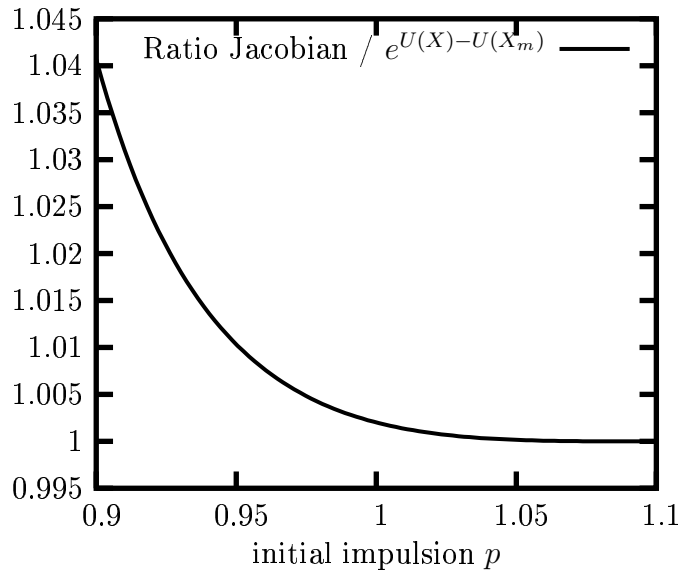

FIG. 4.9 – Checking the preservation of the measure in the case of the free particle, for $p \in [0.9; 1.1]$: we compare $\mathrm{Jac}\left(\frac{\partial X'}{\partial X}\right)(p)$ with $e^{\eta_1+\eta_2-\eta_1(\Delta t)-\eta_2(\Delta t)}(p)$ by plotting their ratio.

the harmonic oscillator). We still work with $M = 2$ (two thermostats), $N = 1$ and $d = 1$ (one particle in a one-dimensionnal space), and with the variables used by the authors, i.e. $X = (q, p, p_{\xi_1}, p_{\xi_2}, \eta_1, \eta_2)$. The GGMT equations and the algorithm to integrate them are the same as in the previous section.

The main idea :

Previously, thanks to the simple choice made for the force, we used MAPLE to get explicit formulas. Here, we do not use MAPLE, but we rather give a more general proof. It will enlighten the choice of the functions $p_{\xi_1}(p)$ and $p_{\xi_2}(p)$ made in Eq. (4.23).

Our proof is based on the concept of algebraicity [12]. We will use the following facts. A polynomial function is algebraic. Any function obtained by addition, multiplication or composition of a finite number of algebraic functions is algebraic. Any root of a polynomial function is an algebraic function of the coefficients. However, the function $x \mapsto e^x$, with x real, is not algebraic. We suppose that the algorithm (4.21) is measure invariant, and we look for a contradiction.

Let us set $X^a = A(\Delta t)\, X$. It can be checked that the vector X^a is a polynomial function of X, and thus a real algebraical function of X. Let us suppose for a moment that $X' = e^{\Delta t\, L_{eff}}\, X$ is an algebraical function of X. We assume that the complete algorithm (4.21) preserves the measure m_0 defined by Eq. (4.20). So we have

$$\mathrm{Jac}\left(\frac{\partial X'}{\partial X}\right) = e^{U(X)-U(X')}.$$

We know that $X'(X)$ and $U(X)$ are algebraical functions of X. Furthermore, they are not constant. So we conclude that the exponential of an algebraical function is an algebraical function, which is not true. We reach the contradiction.

The difficulty and how to solve it :

The previous proof is incorrect, since X' is actually not an algebraical function of X. The issue is on the operator $B(\Delta t)$, which only modifies the value of the impulsion. Let us set $X^b = B(\Delta t)X$. The impulsion in X and X^b are respectively p and p^b, and the second thermostat impulsion in X is p_{ξ_2}. Setting $\alpha = \frac{p_{\xi_2}}{3mQ_2}$ and $\lambda = \lambda(X)$, we have

$$\begin{aligned} p^b &= \frac{p\, e^{-\lambda \frac{\Delta t}{4}}}{\sqrt{1 + \alpha \frac{\Delta t}{2} p^2 e^{-\lambda \frac{\Delta t}{4}}}}; \\ \mathrm{Jac}\left(\frac{\partial X^b}{\partial X}\right)(X) &= \frac{\partial p^b}{\partial p}(p) = \frac{e^{-\lambda \frac{\Delta t}{4}}}{\left(1 + \alpha \frac{\Delta t}{2} p^2 e^{-\lambda \frac{\Delta t}{4}}\right)^{3/2}}. \end{aligned}$$

Because p^b depends on $\lambda(X)$ by an exponential function, the vector X^b is not an algebraical function of X.

Now, let us suppose that, instead of working with X, we work with $\tilde{X} = (q, p, p_{\xi_2})$. We define X as a function of $\tilde{X}$ by

$$X(\tilde{X}) = (q, p, -Q_1\, k_B T\, p_{\xi_2}/Q_2, p_{\xi_2}, 0, 0).$$

Let us now set $X^b = B(\Delta t)\, X(\tilde{X})$. As $\lambda(X) = 0$, we check, thanks to the formulas previously written, that

– X^b is an algebraical function of $\tilde{X}$

– $\mathrm{Jac}\left(\frac{\partial X^b}{\partial X}\right)(X(\tilde{X}))$ is also an algebraical function of $\tilde{X}$.

So we solved our issue.

The solution on the whole algorithm :

In the algorithm (4.21), the operator $B(\Delta t)$ appears four times. However, between the first and the second time, and between the third and the fourth time, the value of λ is not modified. All we need to ensure is that $\lambda = 0$ just before the first application of $B(\Delta t)$, and also just before the third application. We set

$$X \xrightarrow{A(\Delta t)} X^a \xrightarrow{B(\Delta t)} X^b \xrightarrow{C(\Delta t)} X^c \xrightarrow{B(\Delta t)} X^d \xrightarrow{A(\Delta t)} X^e,$$

$$X^e \xrightarrow{L_p(\Delta t/2)} X^f \xrightarrow{L_q(\Delta t)} X^g \xrightarrow{L_p(\Delta t/2)} X^h,$$

$$X^h \xrightarrow{A(\Delta t)} X^i \xrightarrow{B(\Delta t)} X^j \xrightarrow{C(\Delta t)} X^k \xrightarrow{B(\Delta t)} X^l \xrightarrow{A(\Delta t)} X^m,$$

with $X = (q, p, p_{\xi_1}, p_{\xi_2}, \eta_1, \eta_2)$, $X^a = (q^a, p^a, p^a_{\xi_1}, p^a_{\xi_2}, \eta^a_1, \eta^a_2)$, and so on. So we need to ensure that $\lambda(X^a) = \lambda(X^i) = 0$.

Expression of the constraints :
We compute

$$\lambda(X^a) = \frac{1}{Q_1}\left(p_{\xi_1} + \frac{\Delta t}{4}\,G_1(p)\right) + \frac{k_B T}{Q_2}\left(p_{\xi_2} + \frac{\Delta t}{4}\,G_2(p)\right)$$

and

$$\begin{aligned}
\lambda(X^i) &= \lambda(X^h) + \frac{\Delta t}{4}\left(\frac{G_1(p^h)}{Q_1} + k_B T\,\frac{G_2(p^h)}{Q_2}\right) \\
&= \lambda(X^e) + \frac{\Delta t}{4}\,\mathcal{G}(p^h) \\
&= \lambda(X^d) + \frac{\Delta t}{4}\,\mathcal{G}(p^d) + \frac{\Delta t}{4}\,\mathcal{G}(p^h) \\
&= \lambda(X^a) + \frac{\Delta t}{4}\,\mathcal{G}(p^d) + \frac{\Delta t}{4}\,\mathcal{G}(p^h).
\end{aligned}$$

We set $\Phi(p^d, q) = p^d + \dfrac{\Delta t}{2}\left[F(q) + F\left(q + \Delta t\, p^d + \dfrac{\Delta t^2}{2}\,F(q)\right)\right]$, so that $p^h = \Phi(p^d, q)$.
We want to choose X so that

$$\frac{1}{Q_1}\left(p_{\xi_1} + \frac{\Delta t}{4}G_1(p)\right) + \frac{k_B T}{Q_2}\left(p_{\xi_2} + \frac{\Delta t}{4}G_2(p)\right) = 0, \tag{4.24}$$

$$\mathcal{G}\left(\Phi(p^d, q)\right) + \mathcal{G}(p^d) = 0. \tag{4.25}$$

Choosing X :
The goal of this part is to show how to choose X in order for the two previous constraints to be fulfilled. The variable q will be free, and we will define X by $X = (q, p = \psi(q, \Delta t), p_{\xi_1} = \psi_1(q, \Delta t), p_{\xi_2} = \psi_2(q, \Delta t), \eta_1 = 0, \eta_2 = 0)$.

Let us first show that there exists a value of p^d, which is $\theta(q, \Delta t)$, so that the second constraint (4.25) is fullfilled.

Let us suppose that $q = 0$. Since F is a odd degree polynomial function, we see that $\lim_{|p^d|\to\infty} \mathcal{G}\left(\Phi(p^d, q = 0)\right) = +\infty$. So the function $p^d \mapsto \mathcal{G}\left(\Phi(p^d, q = 0)\right) + \mathcal{G}\left(p^d\right)$ goes to $+\infty$ when $\mid p^d \mid$ goes to $+\infty$, and has a negative value when $p^d = 0$. So there exists $\theta(q = 0, \Delta t) > 0$ so that $(p^d, q) = (\theta(0, \Delta t), 0)$ fulfills the second constraint. When q is small enough, we can do the same. So we define a function $\theta(q, \Delta t)$ which is algebraic in q, positive, and which satisfies $\mathcal{G}\left(\Phi(\theta(q, \Delta t), q)\right) + \mathcal{G}\left(\theta(q, \Delta t)\right) = 0$. Let us call μ the strictly positive root of $\mathcal{G}$. We can check that $\lim_{\Delta t\to 0} \theta(q, \Delta t) = \mu > 0$, where μ is independent of q.
Let us set

$$\tau(q, \Delta t) = \frac{3\,m\,Q_2}{\Delta t + q^2 + 1}\;\frac{1}{\left[\theta(q, \Delta t)\right]^2 + 1}. \tag{4.26}$$

We suppose that we can choose X so that $p^a_{\xi_2} = p^c_{\xi_2} = \tau(q, \Delta t)$ and $\lambda(X_a) = 0$. So

$$p^d = \frac{p^c}{\sqrt{1 + \tau(q,\Delta t)\,\frac{\Delta t}{6mQ_2}\,(p^c)^2}}, \quad p^c = p^b, \quad p^b = \frac{p}{\sqrt{1 + \tau(q,\Delta t)\,\frac{\Delta t}{6mQ_2}\,p^2}}.$$

The function $p \mapsto p^d(p, q, \Delta t)$ is an increasing function, and its limit when $p \to \pm\infty$ is $\pm\sqrt{\frac{3\,m\,Q_2}{\Delta t\,\tau(q,\Delta t)}}$. Thanks to the choice of τ, the equation

$$p^d(p, q, \Delta t) = \theta(q, \Delta t),$$

where p is the unknown, has a unique solution, $p = \psi(q, \Delta t)$. This function is algebraic in q. Using the limit of θ, one can check that $\lim_{\Delta t \to 0} \psi(q, \Delta t) = \mu > 0$. Let us notice that many other expressions for the function τ are possible. The main constraint is to ensure that the previous equation (where p is the unknown) has a unique solution, and that $\lim_{\Delta t \to 0} \tau(q, \Delta t)$ exists and depends on q.

Now we define $\psi_2(q, \Delta t) = \tau(q, \Delta t) - \frac{\Delta t}{4}\,G_2\,(\psi(q, \Delta t))$, and we choose in X the value $p_{\xi_2} = \psi_2(q, \Delta t)$. Once again, this function is algebraic in q, and $\lim_{\Delta t \to 0} \psi_2(q, \Delta t) = \tau(q, 0)$.

Let us sum up what we have done until now. We show that it is possible to choose p and p_{ξ_2} as algebraic functions of q, so that the second contraint is fulfilled. Furthermore, we identified the limit when $\Delta t \to 0$ of these functions.

Now, we use the first constraint (4.24) and define

$$\psi_1(q, \Delta t) = -\frac{\Delta t}{4}\,Q_1\,G_1\,(\psi(q, \Delta t)) - k_B T\,\frac{Q_1}{Q_2}\,\left[\psi_2(q, \Delta t) + \frac{\Delta t}{4}\,G_2\,(\psi(q, \Delta t))\right],$$

and we choose in X the value $p_{\xi_1} = \psi_1(q, \Delta t)$. Once again, this function is algebraic in q, and $\lim_{\Delta t \to 0} \psi_1(q, \Delta t) = -k_B T\,\frac{Q_1}{Q_2}\,\tau(q, 0)$.

We proved what we announced at the beginning of this part.

Conclusion :

In the previous part, we show it was possible to define an algebraic function $q \mapsto X(q)$ so that :

- $X'(X(q))$ is an algebraical function of q
- Jac $\left(\frac{\partial X'}{\partial X}\right)(X(q))$ is also an algebraical function of q.

where $X' = e^{\Delta t\,L_{eff}}\,X$. If the algorithm (4.21) is measure invariant, then we have the relation (4.22). We just need to prove that $U(X'(X(q))) - U(X(q))$ is not a constant function to reach a contradiction. In order to do so, we make a Taylor expansion on the variable Δt, and we check that the first term really depends on q. We can do so because we have checked that all the functions we defined to build the function

$q \mapsto X(q)$ have a finite limit when $\Delta t \to 0$.

$$\begin{aligned} U(X'(X(q))) &- U(X(q)) = \eta_1' + \eta_2' - \eta_1 - \eta_2 \\ &= \frac{\Delta t}{2}\left(\frac{1}{Q_1}p^b_{\xi_1} + \frac{1}{Q_1}p^j_{\xi_1} + \frac{1}{Q_2}g(p^b)\,p^b_{\xi_2} + \frac{1}{Q_2}g(p^j)\,p^j_{\xi_2}\right) \\ &= \Delta t\left(\frac{1}{Q_1}\psi_1(q,0) + \frac{1}{Q_2}g\left(\psi(q,0)\right)\psi_2(q,0) + o(1)\right) \\ &= \Delta t\left(-\frac{k_BT}{Q_2}\,\tau(q,0) + \frac{1}{Q_2}g(\mu)\tau(q,0) + o(1)\right) \\ &= \Delta t\left(\frac{\mu^2}{m\,Q_2}\tau(q,0) + o(1)\right). \end{aligned}$$

One can check that $\tau(q,0)$ depends on q (thanks to the particular choice of τ in Eq. (4.26)). The real μ is a strictly positive constant. So we reach the contradiction.

Chapitre 5

Algorithmes pour la résolution de problèmes de mécanique moléculaire de grande taille

Ce chapitre reprend des résultats obtenus en collaboration avec Véronique Duwig, et qui ont donné lieu à un rapport technique interne à EDF. Ce travail a été fait dans le cas unidimensionnel, et nous présentons ci-dessous ses limitations.

On considère un système de $N+1$ atomes en une dimension. Le problème qu'on cherche à résoudre est le problème de minimisation (dit problème de *mécanique moléculaire*)

$$\inf_{\mathbf{u}\in\mathbb{R}^{1+N}} \left\{ \sum_{i=0}^{N-1} W(u_{i+1}-u_i) - \sum_{i=0}^{M-1} f_i u_i;\ u_0 = 0,\ u_N = a \right\}, \tag{5.1}$$

où W est le potentiel interatomique d'interaction (nous faisons l'hypothèse d'interaction à plus proches voisins) et f sont les forces de volume. Le vecteur $\mathbf{u} \in \mathbb{R}^{N+1}$ représente la position, dans la configuration déformée, des $N+1$ atomes. On suppose que les forces n'agissent pas dans tout le domaine, mais simplement sur les atomes $i \in [0, M-1]$, avec $1 \ll M \ll N$. Nous supposons que N est trop grand pour que le problème (5.1) soit tractable numériquement, mais que M est assez petit pour que le même problème, posé sur un système contenant M atomes, soit tractable. Le but de l'étude est de construire des algorithmes fondés sur la résolution d'un problème de taille M pour approcher $\mathbf{u}^r_{|[0,M]}$, où $\mathbf{u}^r$ est la solution de (5.1) (nous revenons ci-dessous sur les problèmes d'existence et d'unicité de solution).

Nous expliquons maintenant notre démarche dans un cadre EDP (description du matériau par un continuum plutôt que par un ensemble discret d'atomes), et en dimension 2. Soit donc W une densité d'énergie élastique, $\Omega \subset \mathbb{R}^2$ le domaine occupé par le matériau, et Ω_{int} le domaine intérieur, dans lequel des forces de volume agissent. Soit $\Omega_{ext} = \Omega \setminus \Omega_{int}$ le domaine extérieur. On note γ l'interface entre Ω_{int} et Ω_{ext}, et Γ le bord du domaine extérieur Ω_{ext} (cf. la figure 5.1).

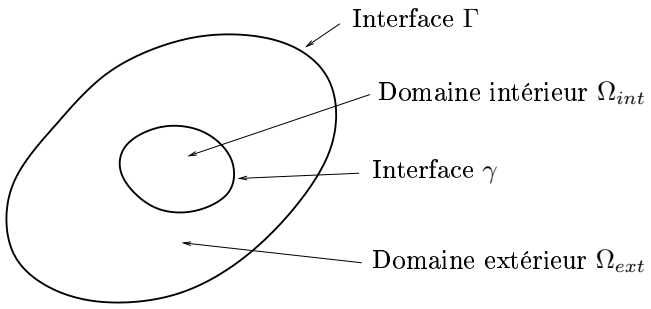

FIG. 5.1 – Décomposition du domaine Ω en un domaine intérieur Ω_{int} et un domaine extérieur Ω_{ext}, séparés par une interface γ. Il n'y a pas de recouvrement entre les deux domaines.

Le problème de référence est

$$\inf \left\{ \int_\Omega W(\nabla u(x))\, dx - \int_{\Omega_{int}} f(x)u(x)dx,\ u \in \left(W^{1,p}(\Omega)\right)^2,\ u_{|\Gamma} = u_0 \right\}, \qquad (5.2)$$

où $p \in \mathbb{N}^*$ est tel que l'énergie est bien définie dès que $u \in (W^{1,p}(\Omega))^2$, et u_0 est une condition aux limites définie sur Γ. La plupart des méthodes que nous étudions (en particulier l'algorithme d'Uzawa 5.2.4) sont des méthodes itératives : partant d'une condition aux limites u^k définie sur γ,

1. on résout le problème de minimisation

$$\inf \left\{ \int_{\Omega_{int}} W(\nabla u(x)) - f(x)u(x)dx,\ u \in \left(W^{1,p}(\Omega)\right)^2, u_{|\gamma} = u^k \right\}, \qquad (5.3)$$

   qui est un problème posé sur le petit domaine Ω_{int}, on note u^k_{int} sa solution ;

2. on résout le problème de minimisation

$$\inf \left\{ \int_{\Omega_{ext}} W(\nabla u(x))dx,\ u \in \left(W^{1,p}(\Omega)\right)^2, u_{|\gamma} = u^k, u_{|\Gamma} = u_0 \right\}, \qquad (5.4)$$

   et on note u^k_{ext} sa solution ;

3. les fonctions u^k_{int} et u^k_{ext} sont utilisées pour mettre à jour la condition d'interface et définir u^{k+1} sur γ.

Le problème de cet algorithme (et les limitations de notre approche) provient de l'étape (5.4), qui est un problème de minimisation sur un grand domaine Ω_{ext}, donc *a priori* aussi difficile à résoudre que le problème de référence (5.2).

Dans certains cas particuliers, on connaît néanmoins la solution de (5.4). Supposons que $F \mapsto W(F)$ soit une fonction strictement convexe sur les matrices 2×2,

alors les problèmes (5.2), (5.3) et (5.4) ont une unique solution. On s'intéresse maintenant à la résolution de (5.4). Supposons que, à l'itération k, il existe une matrice constante A^k telle que les fonctions u^k et u_0 qui apparaissent dans (5.4) vérifient

$$\forall x \in \gamma, u^k(x) = A^k x \quad \text{et} \quad \forall x \in \Gamma, u_0(x) = A^k x. \tag{5.5}$$

Alors, comme W est convexe donc quasi-convexe, on sait que $u(x) = A^k x$ est une solution de (5.4), et c'est la seule car W est strictement convexe. Donc, dans ce cas très particulier, on connaît la solution analytique de (5.4).

On affaiblit maintenant l'hypothèse de convexité sur W, et on suppose simplement que W est quasi-convexe. Alors les problèmes (5.2), (5.3) et (5.4) ont au moins une solution. Sous l'hypothèse (5.5), on sait ici encore que $u(x) = A^k x$ est une solution de (5.4).

Si maintenant W n'est pas quasiconvexe, ou bien si (5.5) n'est pas vérifiée, la résolution de (5.4) est difficile.

Dans l'étude qui suit, nous supposons que W est strictement convexe, et nous considérons le problème dans le cas unidimensionnel. Par conséquent, l'hypothèse (5.5) est toujours vérifiée (en 1D, A^k est un scalaire et γ est la réunion de deux singletons, de même que Γ), donc on connaît analytiquement la solution de (5.4), ce qui permet donc d'utiliser l'algorithme fondé sur (5.3)-(5.4), dans sa version atomistique, pour résoudre le problème de référence (5.1).

Analysis of some domain decomposition algorithms for lattice statics computations

Véronique Duwig[a] and Frédéric Legoll[a,b]

[a] *EDF R & D, Analyse et Modèles Numériques, 1, avenue du Général de Gaulle, 92140 Clamart*
[b] *CERMICS, Ecole Nationale des Ponts et Chaussées, 6 et 8 avenue Blaise Pascal, 77455 Marne-la-Vallée Cedex 2*
and
MICMAC, INRIA Rocquencourt, Domaine de Voluceau, 78153 Le Chesnay Cedex

veronique.duwig@edf.fr, legoll@cermics.enpc.fr

We study some methods to solve minimization problems set on atomistic systems of large size. We propose several algorithms based on the domain decomposition paradigm. Convergence properties of these algorithms are studied, and implementation issues are discussed. Many numerical examples are provided.

5.1 Introduction

In many practical situations, the strength of materials is governed by the presence of singularities in the atomistic lattice. For instance, stacking faults, dislocations or grain boundaries are weak points of a crystalline solid. In order to accurately compute the macroscopic properties of the material, one needs to account for these microscopic singularities, and, for this purpose, the use of an atomistic model is appropriate. However, the size of the material that can be simulated by only resorting to such a fine scale description is very small in comparison with the size of the sample that one is interested in. The same issues arise when one wants to compute the atomistic positions of a dislocation core. The question we address in the present work is the computation of the equilibrium configuration of such large atomistic systems, when the singularities (the grain boundaries, the dislocation core, ...), that call for a fine scale description, are localized in the materials, that is, the deformation is smooth in the main part of the sample.

Many methods have been proposed to address this question. The idea we follow here is to simulate a smaller system, centered around the singularity of interest, and to impose adequate conditions on its boundary. Such methods have been used to compute dislocation properties in [201] (see also [202,203,205] for other applications).

In a dynamical setting, a common problem with such an approach is the unphysical reflection of waves on the interface. In [172, 173], boundary conditions have been developed to minimize these artefacts.

The methods that we will study stem from the domain decomposition and from the integral representation techniques that have been successfully developed in the PDE literature. Let us fix an immersed interface in the system so that the whole system is partitioned into a small *interior* domain, which contains the singularity, and an *exterior* domain (see Fig. 5.2).

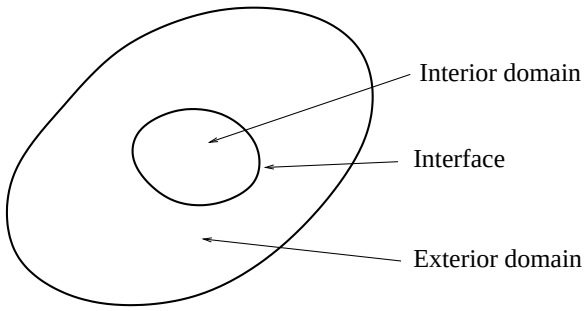

FIG. 5.2 – Decomposition of the whole domain into an interior domain and an exterior domain, separated by an interface.

Both domains are described by the same atomistic model, and the *interior* domain is supposed to be small enough so that, if conditions are applied on its boundary, one ends up with a tractable numerical problem. A natural idea is to choose some arbitrary boundary conditions on the interface, then to solve the corresponding minimization problem in the *interior* domain, and finally to update the boundary condition. The displacement must be continuous on the interface between the two domains, and the sum of the forces acting on the interface atoms must be equal to zero. Thus, updating the boundary condition can be done either by cancelling the residual force on the interface, or by cancelling the discontinuity of the displacement (that is, atoms on the interface are considered to belong to two systems, the *interior* one and the *exterior* one, and the global equilibrium has been reached when the displacement according to the *interior* problem is equal to the displacement according to the *exterior* problem).

The aim of the article is to compare several different methods, both on the numerical analysis standpoint and regarding implementation issues. For this purpose, we assume that we have two softwares at hand :

(i) the first one can compute the equilibrium configuration of a not too large[1] atomistic system ;

(ii) the second one can handle large atomistic systems described by *quadratic* interaction potentials.

For each of the algorithms we shall present below, we will discuss its implementation

[1] A more precise definition will be given in the next sections.

issues, and point out whether it is possible to use the two softwares as black boxes that are simply coupled, or whether one has to modify them.

The article is organized as follows. The numerical analysis of the methods is presented in Section 5.2. To simplify the analysis, we only consider one dimensional problems. First, a reference problem is set in Section 5.2.1. In Section 5.2.2, three methods based on the cancellation of the interface forces are presented. The first one has been proposed in [201], and we propose in Sections 5.2.2.2 and 5.2.2.3 some modifications of this method. In Section 5.2.3, the problem is treated with an adaptation of the Uzawa algorithm. The relation between the latter and the algorithm proposed in Section 5.2.2.2 is outlined. Next, methods based on the continuity of the displacement on the interface are presented in Section 5.2.4. We first develop in Section 5.2.4.1 an iterative method, in the spirit of domain decomposition techniques. Next, in Section 5.2.4.2, we develop a direct technique which is close to integral representation techniques.

Many numerical examples are provided in Section 5.3, to illustrate the results obtained by numerical analysis. A conclusion of the present work is given in Section 5.4.

5.2 Analysis of several algorithms

5.2.1 Reference problem and main setting

Let us consider a one-dimensional atomistic chain composed of $N+1$ atoms, and let us denote by $\{u_i\}_{0\leq i\leq N}$ the positions of the atoms in the current configuration. We assume that only the atoms of index $i \in [0, M-1]$, with $1 \ll M \ll N$, are subjected to body forces, and we denote by f_i the force acting on atom i. The energy of the system is given by

$$E^r(\mathbf{u}) = \sum_{i=0}^{N-1} W(u_{i+1} - u_i) - \sum_{i=0}^{M-1} f_i u_i, \tag{5.6}$$

where $\mathbf{u} \in \mathbb{R}^{1+N}$ stands for the atomistic position vector $(u_0, \ldots, u_N)$. In this equation, W is the interaction potential between atoms, which is such that its minimum is attained at 1. For the sake of simplicity, we will assume only nearest neighbour interactions throughout this article.

Definition 5.2.1 (Reference problem) *The reference problem is the minimization problem*

$$I^r = \inf_{\boldsymbol{u}\in\mathbb{R}^{1+N}} \{E^r(\boldsymbol{u});\ u_0 = 0,\ u_N = a\}\,. \tag{5.7}$$

Lemma 5.2.1 *Let us assume that the interaction potential $W \in \mathcal{C}^2(\mathbb{R})$ satisfies*

$$\begin{aligned} &\exists \alpha > 0, \quad \forall x \in \mathbb{R}, \quad \alpha \leq W''(x), \\ &\exists \beta > 0, \quad \forall x \in \mathbb{R}, \quad W''(x) \leq \beta \text{ and } |W'(x)| \leq \beta\,|x-1|\,. \end{aligned} \tag{5.8}$$

Then problem (5.7) has a unique minimizer $\boldsymbol{u}^r$.

We skip the proof since it is based on standard arguments. The Euler-Lagrange equations of (5.7) read

$$\forall i \in [1, N-1],\ -W'(u^r_{i+1} - u^r_i) + W'(u^r_i - u^r_{i-1}) = \left\{ \begin{array}{l} f_i \ \text{ if } \ i \leq M-1, \\ 0 \ \text{ if } \ i \geq M. \end{array} \right. \tag{5.9}$$

We assume that N is too large for the minimization problem (5.7) to be solved in practice, but that M is small enough so that it is tractable to solve a minimization problem of type (5.7) set on a system of M atoms. In the sequel, as pointed out in the Introduction, we also assume that we can solve the quadratic problem

$$\inf_{\mathbf{u} \in \mathbb{R}^{1+N}} \left\{ \frac{1}{2} \sum_{i=0}^{N-1} (u_{i+1} - u_i - 1)^2 - f_j u_j;\ u_0 = 0,\ u_N = 0 \right\} \tag{5.10}$$

for any $j \in [1, N-1]$, that is to compute the response of the whole system (in the harmonic approximation) when it is subjected to a body force localized on the atom j. More comments on how to make this computation are given in Section 5.4.

5.2.2 Methods based on the cancellation of the interface forces

5.2.2.1 A first algorithm

For any $\mathbf{v} = (v_0, \ldots, v_M) \in \mathbb{R}^{1+M}$, let us set

$$E^{int}(\mathbf{v}) = \sum_{i=0}^{M-1} W(v_{i+1} - v_i) - \sum_{i=0}^{M-1} f_i v_i, \tag{5.11}$$

and for any $\mathbf{u} \in \mathbb{R}^{1+N}$, let us set

$$E^{har}(\mathbf{u}) = \frac{1}{2} \sum_{i=0}^{N-1} (u_{i+1} - u_i - 1)^2. \tag{5.12}$$

The energy E^{int} is the energy of the interior domain system, whereas E^{har} is the elastic energy of the whole domain in the harmonic approximation. In order to approximate the reference solution $\mathbf{u}^r$, the following iterative algorithm has been proposed in [201] (see also [204]) :

Algorithm 5.2.1 (A first method based on interface forces) *Let* E^{int} *and* E^{har} *be defined by (5.11) and (5.12), and let us denote by* $\boldsymbol{u}^k \in \mathbb{R}^{1+N}$ *the configuration at iteration* k.

Let $\boldsymbol{v}^k \in \mathbb{R}^{1+M}$ *be the minimizer of*

$$I^{int}(u_M^k) = \inf_{\boldsymbol{v}\in\mathbb{R}^{1+M}} \left\{ E^{int}(\boldsymbol{v});\ v_0 = 0,\ v_M = u_M^k \right\}. \tag{5.13}$$

Let us set

$$g_M^k = W'\left(v_M^k - v_{M-1}^k\right) - W'\left(u_{M+1}^k - v_M^k\right), \tag{5.14}$$

and let $\boldsymbol{d}^k \in \mathbb{R}^{1+N}$ *be the minimizer of*

$$I^{har}(g_M^k) = \inf_{\boldsymbol{d}\in\mathbb{R}^{1+N}} \left\{ E^{har}(\boldsymbol{d}) + g_M^k d_M;\ d_0 = 0,\ d_N = 0 \right\}. \tag{5.15}$$

The configuration at iteration $k+1$ *is defined by*

$$\boldsymbol{u}^{k+1} = \left(v_0^k, \ldots, v_{M-1}^k, u_M^k + d_M^k, \ldots, u_N^k + d_N^k\right). \tag{5.16}$$

One can see that problems (5.13) and (5.15) are well posed. In the sequel, the problem (5.13) will be referred to as the *interior* problem. Let us now define the following function that will be needed to analyze the algorithm 5.2.1 :

Definition 5.2.2 *Let* E^{int} *be defined by (5.11). We define*

$$h\ :\ \tilde{u} \in \mathbb{R}\ \mapsto\ \tilde{v}_{M-1} \in \mathbb{R},\quad \textit{where } \tilde{\boldsymbol{v}} \in \mathbb{R}^{1+M} \textit{ is defined by}$$

$$E^{int}(\tilde{\boldsymbol{v}}) = \inf_{\boldsymbol{y}\in\mathbb{R}^{1+M}} \left\{ E^{int}(\boldsymbol{y});\ y_0 = 0,\ y_M = \tilde{u} \right\}. \tag{5.17}$$

We now turn to the study of algorithm 5.2.1. Let $\mathbf{u}^0$ be the initial configuration, and let us set

$$Q(x) = W'\left(x - h(x)\right) - W'\left(\frac{x}{c} - \tau_0(\mathbf{u}^0) - x\right), \tag{5.18}$$

where

$$c = \frac{1 - \dfrac{M}{N}}{1 - \dfrac{M+1}{N}}, \quad \tau_0(\mathbf{u}^0) = \frac{u_M^0}{c} - u_{M+1}^0, \tag{5.19}$$

and where h is the function defined by Definition 5.2.2. We define

$$f_F(x) = x - M\left(1 - \frac{M}{N}\right) Q(x). \tag{5.20}$$

Lemma 5.2.2 (Interpretation of algorithm 5.2.1 as a fixed point algorithm) *The sequence* $\left\{u_M^k\right\}_k$ *provided by the algorithm 5.2.1 satisfies*

$$u_M^{k+1} = f_F(u_M^k). \tag{5.21}$$

Theorem 5.2.1 (Convergence of the algorithm 5.2.1) *We assume that the initial configuration* $\boldsymbol{u}^0 \in \mathbb{R}^{1+N}$ *satisfies*

$$u_0^0 = 0,\ u_N^0 = a \text{ and } \forall i \in [M+1, N-1],\quad 2u_i^0 - u_{i-1}^0 - u_{i+1}^0 = 0. \tag{5.22}$$

Let $\left\{\boldsymbol{u}^k\right\}_k$ *be the sequence provided by the algorithm 5.2.1. We assume that (5.8) is satisfied as well as the inequality*

$$\beta\left(\frac{\beta}{M\alpha} + \frac{1}{N-M}\right) \leq \frac{2}{M\left(1 - \dfrac{M}{N}\right)}. \tag{5.23}$$

Then the sequence $\left(v_0^k, \ldots, v_M^k = u_M^k, \ldots, u_N^k\right)$ *converges, as* $k \to +\infty$*, to the reference solution* $\boldsymbol{u}^r$*.*

An admissible initial guess is $u_i^0 = a\ i/N$. Before proving Lemma 5.2.2 and Theorem 5.2.1, let us make a few comments.

Remark 5.2.1 (Properties of the algorithm 5.2.1) *The algorithm 5.2.1 can be implemented with the two codes we have at hand, without modifying them. If* α*,* β*,* M *and* N *do not satisfy (5.23), the convergence of the algorithm is not guaranteed, and it is indeed possible to build some test cases such that the algorithm does not converge (see Section 5.3).*

The following lemmae will be useful in the sequel :

Lemma 5.2.3 *Let* Δ *be the matrix of size* $(N-1) \times (N-1)$ *of the Laplacien operator, with Dirichlet boundary conditions equal to zero. Its inverse* Δ^{-1} *is given by*

$$\forall i \leq j,\quad (\Delta^{-1})_{i,j} = i\left(1 - \frac{j}{N}\right),$$

$$\forall i \geq j+1,\quad (\Delta^{-1})_{i,j} = j\left(1 - \frac{i}{N}\right).$$

Lemma 5.2.4 *Let* δ *be the matrix of size* $(M-1) \times (M-1)$ *of the Laplacien operator, with Dirichlet boundary conditions equal to zero, and let* δ^{-1} *be its inverse. The matrices* Δ^{-1} *and* δ^{-1} *satisfy*

$$-\Delta^{-1}_{M-1,M} + \Delta^{-1}_{M,M}\ \delta^{-1}_{M-1,M-1} = 0.$$

We skip the proof of the two lemmae, which relies on simple matrix analysis. We now prove first Lemma 5.2.2, next Theorem 5.2.1.

Proof of Lemma 5.2.2: In view of (5.15), the vector $\mathbf{d}^k$ satisfies

$$\Delta\ \left(d_1^k, d_2^k, \ldots, d_{N-1}^k\right)^T = -g_M^k\ \mathbf{e}_M, \tag{5.24}$$

where $\mathbf{e}_M = (0, \ldots, 1, \ldots, 0)^T \in \mathbb{R}^{N-1}$ is the M-th unitary vector of $\mathbb{R}^{N-1}$, and where Δ is the matrix of size $(N-1) \times (N-1)$ of the Laplacian operator. With (5.16) and (5.24), we obtain

$$\begin{aligned} u_M^{k+1} &= u_M^k - g_M^k \left(\Delta^{-1}\right)_{M,M}, \\ u_{M+1}^{k+1} &= u_{M+1}^k - g_M^k \left(\Delta^{-1}\right)_{M+1,M}. \end{aligned} \tag{5.25}$$

Let us set $c = \left(\Delta^{-1}\right)_{M,M} / \left(\Delta^{-1}\right)_{M+1,M}$ and $\tau_0 = u_M^0/c - u_{M+1}^0$. We have $u_M^{k+1} - cu_{M+1}^{k+1} = u_M^k - cu_{M+1}^k$, hence

$$u_{M+1}^k = \frac{u_M^k}{c} - \tau_0. \tag{5.26}$$

Making use of the function h defined by Definition 5.2.2, we see that (5.14) reads

$$g_M^k = W'\left(u_M^k - h(u_M^k)\right) - W'\left(u_{M+1}^k - u_M^k\right).$$

With (5.26), we see that $g_M^k = Q(u_M^k)$, where Q is defined by (5.18). The first line of (5.25) then yields

$$u_M^{k+1} = u_M^k - Q(u_M^k)\left(\Delta^{-1}\right)_{M,M}.$$

With Lemma 5.2.4 above, we obtain the expression of $\left(\Delta^{-1}\right)_{M,M}$, that allows us to compute c. We thus obtain (5.21), with f_F defined by (5.20). □

Proof of Theorem 5.2.1: We show that f_F defined by (5.20) is a contracting mapping. First, because of (5.22), we see that τ_0 defined by (5.19) does not depend on $\mathbf{u}^0$. We have

$$Q'(x) = W''\left(x - h(x)\right)\left(1 - h'(x)\right) - W''\left(\frac{x}{c} - \tau_0 - x\right)\left(\frac{1}{c} - 1\right), \tag{5.27}$$

where Q is defined by (5.18) and

$$1 - \frac{1}{c} = \frac{1}{N-M}.$$

With (5.20), we obtain

$$f_F'(x) = 1 - M\left(1 - \frac{M}{N}\right)Q'(x). \tag{5.28}$$

We now need to estimate $Q'(x)$, thus $h'(\tilde{u})$ for any $\tilde{u}$. By definition, $h(\tilde{u}) = \tilde{v}_{M-1}$, where $\tilde{\mathbf{v}}$ is the minimizer of (5.17). The Euler-Lagrange equation of (5.17) reads

$$\forall i \in [1, M-1], \quad -W'(\tilde{v}_{i+1} - \tilde{v}_i) + W'(\tilde{v}_i - \tilde{v}_{i-1}) = f_i. \tag{5.29}$$

In addition, $\tilde{v}_0 = 0$ and $\tilde{v}_M = \tilde{u}$. Let us differentiate the above equation with respect to the boundary condition $\tilde{u}$: we obtain

$$\forall i \in [1, M], \quad a_i - a_{i-1} = \frac{W''(\tilde{v}_1 - \tilde{v}_0)}{W''(\tilde{v}_i - \tilde{v}_{i-1})} (a_1 - a_0), \tag{5.30}$$

where we set

$$a_i = \frac{\partial \tilde{v}_i}{\partial \tilde{u}}, \quad i \in [0, M].$$

Summing the equations (5.30) over $i \in [1, M-1]$ yields

$$h'(\tilde{u}) = a_{M-1} = (a_1 - a_0) \sum_{i=1}^{M-1} \frac{W''(\tilde{v}_1 - \tilde{v}_0)}{W''(\tilde{v}_i - \tilde{v}_{i-1})}.$$

So we have

$$h'(\tilde{u}) = 1 - \frac{1}{1 + \sum_{i=1}^{M-1} \frac{W''(\tilde{v}_M - \tilde{v}_{M-1})}{W''(\tilde{v}_i - \tilde{v}_{i-1})}}, \tag{5.31}$$

where $\tilde{\mathbf{v}} \in \mathbb{R}^{1+M}$ depends on $\tilde{u} \in \mathbb{R}$ (see (5.17)). With the convexity assumption (5.8), we see that, for any $\tilde{u} \in \mathbb{R}$,

$$0 < h'(\tilde{u}) < 1 \quad \text{and} \quad -\frac{\beta}{M\alpha} \leq h'(\tilde{u}) - 1 \leq -\frac{\alpha}{M\beta}. \tag{5.32}$$

Inserting (5.32) in (5.27) and (5.28), we see that

$$\forall x, \ 0 < \kappa_1 \leq Q'(x) \leq \kappa_2 \ \text{ and } \ f_F'(x) < 1, \tag{5.33}$$

with $\kappa_1 = \alpha \left(\frac{\alpha}{M\beta} + \frac{1}{N-M} \right)$ and $\kappa_2 = \beta \left(\frac{\beta}{M\alpha} + \frac{1}{N-M} \right)$. In addition, (5.23) is a sufficient condition to ensure $f_F' > -1$. Thus f_F is a contracting mapping, and the sequence $\left\{u_M^k\right\}_k$ converges. By construction, we have, at any iteration k,

$$\forall i \in [M+1, N-1], \quad 2u_i^k - u_{i-1}^k - u_{i+1}^k = 0. \tag{5.34}$$

So the set $\left\{u_i^k\right\}_{M \leq i \leq N}$ is completely determined by u_M^k and $u_N^k = a$. As $\left\{u_M^k\right\}_k$ converges, one can see that, for all $i \in [M+1, N-1]$, the sequence $\left\{u_i^k\right\}_k$ converges. By definition (see (5.13)), the sequence $\left\{\mathbf{v}^k\right\}_k$ also converges. Let us now consider the configuration $\mathbf{w}^k$ defined by

$$\mathbf{w}^k = \left(v_0^k, \ldots, v_M^k = u_M^k, \ldots, u_N^k\right).$$

We have proved that the sequence $\left\{\mathbf{w}^k\right\}_k$ converges, and one can show that its limit $\mathbf{w}^\infty$ satisfies the Euler-Lagrange equations (5.9) of the reference problem (5.7). So $\mathbf{w}^\infty = \mathbf{u}^r$. □

Remark 5.2.2 *Let us show that, when the algorithm 5.2.1 converges to some* $\boldsymbol{u}^\infty$*, the assumption (5.22) is a necessary condition for* $\boldsymbol{u}^\infty$ *to be the reference solution. We argue by contradiction. With (5.24), we see that*

$$\forall i \in [M+1, N-1], \quad 2d_i^k - d_{i-1}^k - d_{i+1}^k = 0.$$

By construction (see (5.16)), for all $i \in [M+1, N-1]$*, we have* $u_i^k = u_i^0 + \sum_{p=0}^{k-1} d_i^p$ *for all* $k \geq 0$*. Thus we see that*

$$\forall k, \ \forall i \in [M+1, N-1], \quad 2u_i^k - u_{i-1}^k - u_{i+1}^k = 2u_i^0 - u_{i-1}^0 - u_{i+1}^0. \tag{5.35}$$

If (5.22) is not satisfied, and if the algorithm converges, then $\lim_{k \to +\infty} u_i^k = u_i^\infty$ *and we infer from (5.35) that*

$$\forall i \in [M+1, N-1], \quad 2u_i^\infty - u_{i-1}^\infty - u_{i+1}^\infty \neq 0.$$

On the other hand, with (5.9), we see that

$$\forall i \in [M+1, N-1], \quad 2u_i^r - u_{i-1}^r - u_{i+1}^r = 0.$$

Thus $\boldsymbol{u}^\infty \neq \boldsymbol{u}^r$*.*

5.2.2.2 Another algorithm

In the algorithm 5.2.1, the current position u_M^k of the atom M at the iteration k is updated according to $u_M^{k+1} = u_M^k + d_M^k$. We now propose a different algorithm, based on the updating relation $u_M^{k+1} = u_M^k + \mu\, d_M^k$, where $\mu > 0$ is a user-chosen parameter that is kept fixed along the iterations. The algorithm that we propose reads :

Algorithm 5.2.2 (Another method based on interface forces) *Let* E^{int} *and* E^{har} *be defined by (5.11) and (5.12), and let us choose* $\mu > 0$*. Let us denote by* $\boldsymbol{u}^k \in \mathbb{R}^{1+N}$ *the configuration at iteration* k*.*

Let $\boldsymbol{v}^k \in \mathbb{R}^{1+M}$ *be the minimizer of (5.13), let* g_M^k *be defined by (5.14), and let* $\boldsymbol{d}^k \in \mathbb{R}^{1+N}$ *be the minimizer of (5.15). The configuration at iteration* $k+1$ *is defined by*

$$\boldsymbol{u}^{k+1} = \left(v_0^k, \ldots, v_{M-1}^k, u_M^k + \mu d_M^k, \ldots, u_N^k + \mu d_N^k\right). \tag{5.36}$$

The algorithm 5.2.1 is a special case of the algorithm 5.2.2 (it corresponds to the case $\mu = 1$). Let $\mathbf{u}^0$ be the initial configuration, and let us set

$$f_{F,\mu}(x) = x - \mu M \left(1 - \frac{M}{N}\right) Q(x), \tag{5.37}$$

with Q given by (5.18).

Theorem 5.2.2 (Convergence of the algorithm 5.2.2) *We assume that the initial configuration* $\boldsymbol{u}^0 \in \mathbb{R}^{1+N}$ *satisfies (5.22). Let* $\left\{\boldsymbol{u}^k\right\}_k$ *be the sequence provided by the algorithm 5.2.2. The sequence* $\left\{u_M^k\right\}_k$ *satisfies*

$$u_M^{k+1} = f_{F,\mu}(u_M^k). \tag{5.38}$$

Let us now assume that (5.8) is satisfied and that α, β, M, N *and* $\mu > 0$ *satisfy*

$$\beta\left(\frac{\beta}{M\alpha} + \frac{1}{N-M}\right) \leq \frac{2}{\mu M\left(1-\frac{M}{N}\right)}. \tag{5.39}$$

Then the sequence $\left(v_0^k, \ldots, v_M^k = u_M^k, \ldots, u_N^k\right)$ *converges to the reference solution* $\boldsymbol{u}^r$.

An admissible initial guess is $u_i^0 = a\ i/N$. The proof of Theorem 5.2.2 is skipped since it follows the same pattern as the proofs of Lemma 5.2.2 and Theorem 5.2.1.

Let us now study the influence of μ on the convergence rate of the algorithm 5.2.2. With (5.37), we see that

$$f'_{F,\mu}(x) = 1 - \mu M\left(1 - \frac{M}{N}\right) Q'(x), \tag{5.40}$$

and we also have $Q'(x) > 0$ (see (5.33)).

Let us assume that the algorithm 5.2.1 does not converge : then the sequence $\{u_M^k\}_k$ provided by the algorithm does not converge. Since this sequence satisfies (5.21), it implies that f_F is not a contracting mapping. With (5.33), we know that $\sup f'_F < 1$, so we have $\inf f'_F = \inf f'_{F,\mu=1} \leq -1$. In view of (5.40), we see that, when μ decreases, $f'_{F,\mu}$ increases. Since $Q'(x)$ is bounded with respect to x (see (5.33)), it is possible to find a value $\mu \in (0,1)$ such that, for all $x \in \mathbb{R}$, $f'_{F,\mu}(x) > -1$, and for such a value, the algorithm 5.2.2 converges.

Note that, if $\inf f'_F > 0$, the algorithm 5.2.1 converges, and the algorithm 5.2.2 with $\mu < 1$ also converges but with a slower convergence rate.

5.2.2.3 A Newton algorithm

As the algorithm 5.2.2 is a fixed point algorithm on u_M defined by the contracting mapping $f_{F,\mu}$, one can choose to use, at iteration k, the value μ^k such that $f'_{F,\mu^k}(u_M^k) = 0$. With (5.27) and (5.40), we see that this "optimal" value reads

$$\frac{1}{\mu^k} = M\left(1-\frac{M}{N}\right)\Big(W''\left(u_M^k - v_{M-1}^k\right)\left(1-h'(u_M^k)\right) - W''\left(u_{M+1}^k - u_M^k\right)\left(\frac{1}{c}-1\right)\Big), \tag{5.41}$$

where $\mathbf{v}^k$ is the minimizer of (5.13) and where $h'(u_M^k)$ can be computed from (5.31). The equation (5.41) allows one to guess, on the fly, a good value for the parameter μ. So the algorithm with optimal value of μ reads :

Algorithm 5.2.3 (A Newton algorithm based on interface forces) *Let E^{int} and E^{har} be defined by (5.11) and (5.12). Let us denote by $\boldsymbol{u}^k \in \mathbb{R}^{1+N}$ the configuration at iteration k.*

Let $\boldsymbol{v}^k \in \mathbb{R}^{1+M}$ be the minimizer of (5.13), let g_M^k be defined by (5.14), and let $\boldsymbol{d}^k \in \mathbb{R}^{1+N}$ be the minimizer of (5.15). The configuration at iteration $k+1$ is defined by

$$\boldsymbol{u}^{k+1} = \left(v_0^k, \ldots, v_{M-1}^k, u_M^k + \mu^k d_M^k, \ldots, u_N^k + \mu^k d_N^k\right), \tag{5.42}$$

where μ^k is given by (5.41).

The algorithm 5.2.3 can be recast as a Newton algorithm. Indeed, with (5.40), we see that the equation $f'_{F,\mu^k}(u_M^k) = 0$ implies

$$\mu^k = \frac{1}{M\left(1 - \frac{M}{N}\right) Q'(u_M^k)}.$$

At each step, we have $u_M^{k+1} = f_{F,\mu^k}(u_M^k)$ (see (5.38)), which reads

$$u_M^{k+1} = u_M^k - \frac{Q(u_M^k)}{Q'(u_M^k)}.$$

So the algorithm 5.2.3 is exactly a Newton algorithm for computing a zero of function Q. Note that the algorithm 5.2.2 has been presented as an algorithm to find a fixed point of $f_{F,\mu}$, but, in view of (5.37), x is a fixed point for $f_{F,\mu}$ if and only if $Q(x) = 0$.

Theorem 5.2.3 (Convergence of the algorithm 5.2.3) *We assume that the initial configuration $\boldsymbol{u}^0 \in \mathbb{R}^{1+N}$ satisfies (5.22), and that α, β, M and N satisfy*

$$\beta\left(\frac{\beta}{M\alpha} + \frac{1}{N-M}\right) \leq 2\alpha\left(\frac{\alpha}{M\beta} + \frac{1}{N-M}\right). \tag{5.43}$$

Then the sequence $\left(v_0^k, \ldots, v_M^k = u_M^k, \ldots, u_N^k\right)$ provided by the algorithm 5.2.3 converges to the reference solution $\boldsymbol{u}^r$.

Note that, in the limit $1 \ll M \ll N$, the condition (5.43) reads $(\beta/\alpha)^3 < 2$.

Proof of Theorem 5.2.3: We know that $u_{M-1}^r = h(u_M^r)$, where h is defined by (5.17) and $\mathbf{u}^r$ is the reference solution. Because of the choice (5.22), we have

$$\frac{u_M^r}{c} - \tau_0(\mathbf{u}^0) - u_M^r = u_{M+1}^r - u_M^r.$$

Inserting this information in (5.18), we obtain

$$Q(u_M^r) = W'(u_M^r - u_{M-1}^r) - W'(u_{M+1}^r - u_M^r).$$

With (5.9), we infer that $Q(u_M^r) = 0$. Since Q is an increasing function (see (5.33)), the equation $Q(x) = 0$ has a unique solution. We now prove that the sequence $\{u_M^k\}_k$ provided by the algorithm converges to u_M^r. Let us denote by

$$e^k = u_M^k - u_M^r$$

the error : we have $e^{k+1} = e^k - Q(u_M^k)/Q'(u_M^k)$. Let us assume for instance that $u_M^k \geq u_M^r$: since $Q(u_M^r) = 0$ and Q' is bounded from below by $\kappa_1 > 0$ and from above by κ_2 (see (5.33)), we have

$$(u_M^k - u_M^r)\frac{\kappa_1}{\kappa_2} \leq \frac{Q(u_M^k)}{Q'(u_M^k)} \leq (u_M^k - u_M^r)\frac{\kappa_2}{\kappa_1},$$

thus

$$\left(1 - \frac{\kappa_2}{\kappa_1}\right) e^k \leq e^{k+1} \leq \left(1 - \frac{\kappa_1}{\kappa_2}\right) e^k.$$

We have $\kappa_2 \geq \kappa_1 > 0$, thus $(\kappa_2/\kappa_1 - 1) \geq (1 - \kappa_1/\kappa_2) > 0$, so

$$\left|\frac{e^{k+1}}{e^k}\right| \leq \left(\frac{\kappa_2}{\kappa_1} - 1\right)$$

and the condition (5.43) ensures that the right hand side of the above inequality is lower than 1. Thus the sequence $\{e^k\}_k$ converges to 0 as $k \to +\infty$ and therefore u_M^k converges to u_M^r. □

5.2.3 Uzawa algorithms

The algorithms 5.2.1, 5.2.2 and 5.2.3 rely on the successive minimization of, first an energy in the interior domain, next another energy in the whole domain. We now adapt the well-known Uzawa algorithm, which often leads to such alternate minimizations, to our problem.

The Uzawa algorithm is an algorithm to solve constrained problems. In the sequel, we associate to the reference problem (5.7) two different constrained problems, and thus we obtain two different algorithms. The first one will be shown to be, in the setting we work in, identical to the algorithm 5.2.2 (see Section 5.2.3.1), whereas the second one, obtained with more natural ideas, will be shown to be very difficult to implement (see Section 5.2.3.2).

5.2.3.1 An algorithm based on the stress at the interface

Let us consider the minimization problem

$$I^r = \inf_{\mathbf{u} \in \mathbb{R}^{1+N}} \{E^r(\mathbf{u});\ u_0 = \tilde{a},\ u_N = a\}, \tag{5.44}$$

where E^r is the energy defined by (5.6), and let U be the Legendre transform of the interaction potential W :

$$U(s) = \max_x \left(sx - W(x)\right). \tag{5.45}$$

Let us consider the problem

$$\tilde{I}^r = \inf_{\sigma \in \mathbb{R}^N} \left\{ \tilde{G}^r(\sigma);\ \forall i \in [0, N-1],\ \sigma_i - \sigma_{i-1} + f_i = 0 \right\}, \tag{5.46}$$

where $f_i = 0$ for all $i \geq M$, and where the energy $\tilde{G}^r$ reads

$$\forall \sigma = (\sigma_0, \ldots, \sigma_{N-1}) \in \mathbb{R}^N, \quad \tilde{G}^r(\sigma) = \sum_{i=0}^{N-1} U(\sigma_i) - \sigma_{N-1} a + \sigma_0 \tilde{a}, \tag{5.47}$$

where $\tilde{a}$ and a are the Dirichlet boundary conditions of (5.44). We recall the following theorem :

Theorem 5.2.4 *Let us assume that (5.8) is satisfied. Then U is well defined by (5.45), problems (5.44) and (5.46) are well posed, and*

$$I^r = \tilde{I}^r. \tag{5.48}$$

In addition, let $\tilde{\boldsymbol{u}}^r$ be the minimizer of (5.44) and let σ^r be the minimizer of (5.46) : then

$$\sigma_i^r = W'\left(\tilde{u}_{i+1}^r - \tilde{u}_i^r\right). \tag{5.49}$$

Thus, in view of this theorem, we can associate to (5.7) the problem

$$\inf_{\sigma \in \mathbb{R}^N} \left\{ G^r(\sigma);\ \forall i \in [0, N-1],\ \sigma_i - \sigma_{i-1} + f_i = 0 \right\}, \tag{5.50}$$

where $f_i = 0$ for all $i \geq M$, and where the energy G^r reads

$$\forall \sigma = (\sigma_0, \ldots, \sigma_{N-1}) \in \mathbb{R}^N, \quad G^r(\sigma) = \sum_{i=0}^{N-1} U(\sigma_i) - \sigma_{N-1} a, \tag{5.51}$$

where a is the boundary condition of (5.7). As $f_M = 0$, we have $\sigma_M = \sigma_{M-1}$ in the variational space of the problem (5.50). Instead of working with $\sigma = (\sigma_0, \ldots, \sigma_{N-1}) \in \mathbb{R}^N$, we now work with

$$\sigma^{int} = \left(\sigma_0^{int}, \ldots, \sigma_{M-1}^{int}\right) \in \mathbb{R}^M \ \text{ and } \ \sigma^{ext} = \left(\sigma_M^{ext}, \ldots, \sigma_{N-1}^{ext}\right) \in \mathbb{R}^{N-M}.$$

The problem (5.50) is equivalent to a contrained minimization problem set on variables $(\sigma^{int}, \sigma^{ext})$ where the constraint reads $\sigma_M^{ext} = \sigma_{M-1}^{int}$. For this minimization

problem, where the unknown are the stresses, it is possible to write down an Uzawa algorithm. We now change of variables, and work with

$$\mathbf{v} = (v_0, \ldots, v_M) \in \mathbb{R}^{M+1} \text{ and } \mathbf{x} = (x_M, \ldots, x_N) \in \mathbb{R}^{N-M+1}, \tag{5.52}$$

which are homogeneous to position variables. Let us introduce the energy

$$E_U^{ext}(\mathbf{x}) = \sum_{i=M}^{N-1} W(x_{i+1} - x_i). \tag{5.53}$$

Expressed in position variables, the preceeding Uzawa algorithm reads :

Algorithm 5.2.4 *Let E^{int} and E_U^{ext} be defined by (5.11) and (5.53). Let $\boldsymbol{v}^k \in \mathbb{R}^{M+1}$, $\boldsymbol{x}^k \in \mathbb{R}^{N-M+1}$ and $\lambda^k \in \mathbb{R}$ be the positions and the Lagrange multiplier at iteration k. We now define them at iteration $k+1$.*

Let $\boldsymbol{v}^{k+1}$ be the minimizer of

$$\inf_{\boldsymbol{v}} \left\{ E^{int}(\boldsymbol{v}),\ v_0 = 0,\ v_M = \lambda^k \right\}, \tag{5.54}$$

let $\boldsymbol{x}^{k+1}$ be the minimizer of

$$\inf_{\boldsymbol{x}} \left\{ E_U^{ext}(\boldsymbol{x}),\ x_M = \lambda^k,\ x_N = a \right\} \tag{5.55}$$

and let λ^{k+1} be given by

$$\lambda^{k+1} = \lambda^k + \mu \left(W'\left(x_{M+1}^{k+1} - x_M^{k+1}\right) - W'\left(v_M^{k+1} - v_{M-1}^{k+1}\right) \right)$$

where $\mu > 0$ is a user-chosen parameter.

Remark 5.2.3 (Properties of the algorithm 5.2.4) *The first step (5.54) consists in the resolution of a nonlinear problem set on the interior domain $[0, M]$ with Dirichlet boundary conditions, we can thus use the first code that we have at hand.*

The second step (5.55) is a nonlinear minimization problem with Dirichlet boundary conditions and without body force term. As W is convex and since we work in a one dimensional setting, the minimizer of (5.55) is known, it is the linear deformation

$$x_i^{k+1} = \frac{N-i}{N-M}\,\lambda^k + \frac{i-M}{N-M}\,a. \tag{5.56}$$

We now compare the algorithms 5.2.2 and 5.2.4. The first step (5.13) of the algorithm 5.2.2 is exactly the same problem as (5.54). We have seen that the solution of (5.55) satisfies (5.56), so, on the exterior domain, the deformation is always linear. In view of (5.34), the same property holds for the algorithm 5.2.2. So, if the initial guesses are such that $\lambda^0 = u_M^0$, and if the initial guess of the algorithm 5.2.2 satisfies (5.22), then algorithms 5.2.2 and 5.2.4 yield the same sequence.

5.2.3.2 An algorithm based on the position at the interface

For the sake of completeness, we study in this section a constrained problem which is naturally associated to the reference problem (5.7). Then, we write down the Uzawa algorithm to solve this constrained problem, and we show that the so-obtained numerical problems are as difficult to solve as the reference problem.

The idea is to use two degrees of freedom to describe the position of atom M, one corresponding to the interior system, the other one corresponding to the exterior system (in a PDE setting, this method would correspond to a discontinuous FE discretization of the displacement field). Let us define $\mathbf{v} \in \mathbb{R}^{M+1}$ and $\mathbf{x} \in \mathbb{R}^{N-M+1}$ by (5.52), and

$$Q(\mathbf{v}, \mathbf{x}) = x_M - v_M. \tag{5.57}$$

The reference problem (5.7) is equivalent to the constrained minimization problem

$$\inf \left\{ E_U(\mathbf{v}, \mathbf{x});\ v_0 = 0,\ x_N = a,\ Q(\mathbf{v}, \mathbf{x}) = 0 \right\}, \tag{5.58}$$

where the energy E_U is given by

$$E_U(\mathbf{v}, \mathbf{x}) = E^{int}(\mathbf{v}) + E_U^{ext}(\mathbf{x}), \tag{5.59}$$

where E^{int} and E_U^{ext} are defined by (5.11) and (5.53). For any $\lambda \in \mathbb{R}$, let us set

$$\mathcal{L}^{int}(\mathbf{v}, \lambda) = E^{int}(\mathbf{v}) - \lambda v_M, \tag{5.60}$$

$$\mathcal{L}^{ext}(\mathbf{x}, \lambda) = E_U^{ext}(\mathbf{x}) + \lambda x_M. \tag{5.61}$$

To solve (5.58), the Uzawa algorithm reads :

Algorithm 5.2.5 *Let $\boldsymbol{v}^k \in \mathbb{R}^{M+1}$, $\boldsymbol{x}^k \in \mathbb{R}^{N-M+1}$ and $\lambda^k \in \mathbb{R}$ be the positions and the Lagrange multiplier at iteration k. We now define them at iteration $k+1$.*

Let $\boldsymbol{v}^{k+1}$ be the minimizer of

$$\inf_{\boldsymbol{v}} \left\{ \mathcal{L}^{int}(\boldsymbol{v}, \lambda^k),\ v_0 = 0 \right\}, \tag{5.62}$$

let $\boldsymbol{x}^{k+1}$ be the minimizer of

$$\inf_{\boldsymbol{x}} \left\{ \mathcal{L}^{ext}(\boldsymbol{x}, \lambda^k),\ x_N = a \right\} \tag{5.63}$$

and let λ^{k+1} be given by

$$\lambda^{k+1} = \lambda^k + \mu Q(\boldsymbol{v}^{k+1}, \boldsymbol{x}^{k+1}),$$

where $\mu > 0$ is a user-chosen parameter.

The second step (5.63) of the algorithm is as difficult to solve as the reference problem, for it is a nonlinear problem set on the exterior domain $[M, N]$, with Dirichlet and Neumann boundary conditions. This is why the algorithm 5.2.5 cannot be used in practice for large scale problems.

Theorem 5.2.5 (Convergence of the algorithm 5.2.5) *Let us assume that (5.8) is satisfied, and that*

$$0 < \mu < \frac{2\alpha}{N}. \tag{5.64}$$

Then the algorithm 5.2.5 converges.

Proof: To simplify the notation, let us set $\psi = (W')^{-1}$. Then one can show that the sequence $\left\{\lambda^k\right\}_k$ provided by the algorithm 5.2.5 satisfies

$$\lambda^{k+1} = f_{d,\mu}\left(\lambda^k\right),$$

where $f_{d,\mu}$ is such that

$$f'_{d,\mu}\left(\lambda^k\right) = 1 - \mu\left((N-M)\psi'\left(\lambda^k\right) + \sum_{i=0}^{M-1} \frac{1}{W''\left(v_{i+1}^{k+1} - v_i^{k+1}\right)}\right), \tag{5.65}$$

where $\mathbf{v}^{k+1}$ is the minimizer of (5.62). The condition (5.64) ensures that $\left|f'_{d,\mu}\right| < 1$, thus the sequence $\left\{\lambda^k\right\}_k$ converges. □

Remark 5.2.4 (Algorithm with optimal value of μ) *In view of (5.65), one can propose a value μ^k such that, at the iteration k, we have $f'_{d,\mu^k}\left(\lambda^k\right) = 0$. However, to compute this value, one needs to know $\psi'\left(\lambda^k\right)$, which can be difficult to compute. The "optimal" value that we propose is based on the harmonic approximation $\psi'\left(\lambda^k\right) = 1/W''(\psi(\lambda^k)) \approx 1/W''(1)$ and reads*

$$\mu^k = \frac{1}{\dfrac{N-M}{W''(1)} + \displaystyle\sum_{i=0}^{M-1} \frac{1}{W''(v_{i+1}^{k+1} - v_i^{k+1})}}, \tag{5.66}$$

where $\boldsymbol{v}^{k+1}$ is the minimizer of (5.62). This value can be computed once (5.62) has been solved.

5.2.4 Methods based on the continuity of the displacement at the interface

We first present in this section a method related to domain decomposition ideas, then a method based on integral representation ideas.

5.2.4.1 An iterative algorithm

For any $\overline{\mathbf{v}} = (\overline{v}_{M-1}, \ldots, \overline{v}_N) \in \mathbb{R}^{N-M+2}$, we define the elastic energy (in the harmonic approximation) of the exterior domain by

$$\bar{E}^{har}(\overline{\mathbf{v}}) = \frac{1}{2} \sum_{i=M-1}^{N-1} (\overline{v}_{i+1} - \overline{v}_i - 1)^2 . \tag{5.67}$$

Let $\bar{\delta}$ be the matrix of size $(N-M)\times(N-M)$ of the Laplacien operator (with Dirichlet boundary conditions equal to zero). To compute the reference solution $\mathbf{u}^r \in \mathbb{R}^{1+N}$, the following algorithm can be used :

Algorithm 5.2.6 (A method based on the positions at the interface) *Let us denote by $\boldsymbol{u}^k \in \mathbb{R}^{1+N}$ the configuration at iteration k.*

Let $\boldsymbol{v}^k \in \mathbb{R}^{1+M}$ be the minimizer of (5.13), and let $\overline{\boldsymbol{v}}^{k+1} \in \mathbb{R}^{N-M+2}$ be the minimizer of

$$\bar{I}^{har} = \inf_{\overline{\boldsymbol{v}}\in\mathbb{R}^{N-M+2}} \left\{ \bar{E}^{har}(\overline{\boldsymbol{v}});\ \overline{v}_{M-1} = v^k_{M-1},\ \overline{v}_N = a \right\}, \tag{5.68}$$

where $\bar{E}^{har}$ is defined by (5.67). The configuration at the iteration $k+1$ is defined by

$$\boldsymbol{u}^{k+1} = (v^k_0, .., v^k_{M-1} = \overline{v}^{k+1}_{M-1}, \overline{v}^{k+1}_M, .., \overline{v}^{k+1}_N).$$

This algorithm is a domain decomposition algorithm, where the two domains overlap on a single atom (the atom M).

Remark 5.2.5 (Relation between $\bar{\delta}$ and Δ) *Let $\tilde{\boldsymbol{v}} = (\tilde{v}_{M-1}, \ldots, \tilde{v}_N) \in \mathbb{R}^{N-M+2}$ be the minimizer of*

$$\bar{E}^{har}(\tilde{\boldsymbol{v}}) = \inf_{\boldsymbol{y}\in\mathbb{R}^{N-M+2}} \left\{ \bar{E}^{har}(\boldsymbol{y});\ y_{M-1} = \tilde{b},\ y_N = \tilde{a} \right\}, \tag{5.69}$$

where $\tilde{b}$ and $\tilde{a}$ are two real numbers. Then the restriction $(\tilde{v}_M, \ldots, \tilde{v}_{N-1})$ of $\tilde{\boldsymbol{v}}$ is the solution of the following linear system of size $(N-M)$:

$$\bar{\delta}\,(\tilde{v}_M, \ldots, \tilde{v}_{N-1})^T = \left(\tilde{b}, 0, \ldots, 0, \tilde{a}\right)^T. \tag{5.70}$$

It is possible to compute $\tilde{\boldsymbol{v}}$ with the Green function of the whole problem, that is, in our case, the inverse of the matrix Δ. Indeed, we have the relation

$$\tilde{\boldsymbol{v}} = (\tilde{b}, \tilde{u}_M, \ldots, \tilde{u}_{N-1}, \tilde{a}),$$

where $\tilde{\boldsymbol{u}} = (\tilde{u}_1, \ldots, \tilde{u}_{N-1})$ is defined by

$$\tilde{\boldsymbol{u}}^T = \Delta^{-1}\,(0, \ldots, 0, F_{M-1}, 0, \ldots, 0, \tilde{a})^T,$$

with

$$F_{M-1} = \frac{\tilde{b} - (\Delta^{-1})_{M-1,N-1}\,\tilde{a}}{(\Delta^{-1})_{M-1,M-1}}.$$

The proof relies on the observation that $(\tilde{u}_M, \ldots, \tilde{u}_{N-1})$ satisfies (5.70).

So, when one knows the Green functions of the whole system (composed on $N+1$ atoms), one can directly compute the minimizer of the exterior problem (5.68). There is no need for computing the Green functions of the domain $[M-1, N]$.

Theorem 5.2.6 (Convergence of the algorithm 5.2.6) *Let $\boldsymbol{u}^0 \in \mathbb{R}^{1+N}$ be any initial guess, and let us define the sequence $\left\{\boldsymbol{u}^k\right\}_k$ by the algorithm 5.2.6. If (5.8) is satisfied, then the sequence $\left\{\boldsymbol{u}^k\right\}_k$ converges to the reference solution $\boldsymbol{u}^r$ of (5.7).*

Proof: Let us first assume that the algorithm converges to a configuration $\mathbf{u}^\infty$. Then, letting k go to $+\infty$ into the Euler-Lagrange of (5.13) and (5.68), it can be shown that $\mathbf{u}^\infty$ satisfies the Euler-Lagrange equations (5.9) of (5.7). So $\mathbf{u}^\infty = \mathbf{u}^r$.

We now prove that the algorithm converges. For this purpose, it is sufficient to prove that the sequence $\left\{v_{M-1}^k\right\}_k$ converges. The Euler-Lagrange equation of (5.68) reads

$$\left(\overline{v}_M^{k+1}, \ldots, \overline{v}_{N-1}^{k+1}\right)^T = \bar{\delta}^{-1}\left(v_{M-1}^k, 0, \ldots, 0, a\right)^T,$$

so

$$u_M^{k+1} = \overline{v}_M^{k+1} = \left(\bar{\delta}^{-1}\right)_{1,1} v_{M-1}^k + \left(\bar{\delta}^{-1}\right)_{1,N-M} a\,.$$

Hence, the sequence $\left\{v_{M-1}^k\right\}_k$ satisfies the recursion formula

$$v_{M-1}^{k+1} = h\left(u_M^{k+1}\right) = h\left(\left(\bar{\delta}^{-1}\right)_{1,1} v_{M-1}^k + \left(\bar{\delta}^{-1}\right)_{1,N-M} a\right),$$

where h is defined by Definition 5.2.2. So the algorithm (5.2.6) is a fixed point algorithm on v_{M-1}, namely $v_{M-1}^{k+1} = f_P\left(v_{M-1}^k\right)$, where

$$f_P \;:\; v \in \mathbb{R} \;\mapsto\; h\left(\left(\bar{\delta}^{-1}\right)_{1,1} v + \left(\bar{\delta}^{-1}\right)_{1,N-M} a\right). \tag{5.71}$$

Let us now show that f_P is a contracting mapping. Its derivative reads

$$f_P'(v) = h'\left(\left(\bar{\delta}^{-1}\right)_{1,1} v + \left(\bar{\delta}^{-1}\right)_{1,N-M} a\right)\left(\bar{\delta}^{-1}\right)_{1,1}. \tag{5.72}$$

With (5.32), we have $\sup|h'| < 1$. Since $\left(\bar{\delta}^{-1}\right)_{1,1} = 1 - \dfrac{1}{N-M+1}$, we obtain $\sup|f_P'| < 1$. □

Remark 5.2.6 (Convergence rate of the algorithm 5.2.6) *We work in the regime $1 \ll M \ll N$. So $\left(\bar{\delta}^{-1}\right)_{1,1} \approx 1$, and the convergence rate of the algorithm 5.2.6 directly depends on h'. In view of (5.72), we can see that, the closer h' to 0, the larger the convergence rate.*

If W is a quadratic potential, then (5.32) implies that

$$\forall u \in \mathbb{R},\; h'(u) = 1 - \frac{1}{M}\,.$$

As soon as M is large, h' is close to 1 and the convergence rate, which reads, with (5.72),

$$f_P'(u) = \left(1 - \frac{1}{M}\right)\left(1 - \frac{1}{N-M+1}\right), \tag{5.73}$$

is slowed down.

Remark 5.2.7 (Properties of the algorithm 5.2.6) *With Remark 5.2.5, we see that the algorithm 5.2.6 can be implemented with the two codes we have at hand. Unlike the algorithm 5.2.1, the algorithm 5.2.6 always converges. However, in general, the convergence rate is slow (see Remark 5.2.6).*

Remark 5.2.8 (Domain decomposition with a larger overlap) *The algorithm 5.2.6 can be generalized to an algorithm with a larger overlap ($R \in \mathbb{N}^*$ atoms instead of one, namely the atoms $M-R+1,\dots,M$ instead of the atom M). Then the sequence $\left\{u_M^k\right\}_k$ provided by the algorithm satisfies $u_M^{k+1} = f_{P,R}(u_M^k)$, where $f_{P,R}$ is some contracting mapping. If W is a quadratic potential, then, one obtains, in the regime $R \ll M$,*

$$f'_{P,R}(\tilde{u}) \approx \left(1 - \frac{R}{N-M+R}\right)\left(1 - \frac{R}{M}\right).$$

So $f'_{P,R} > 0$ is lower than f'_P (see (5.73)), hence the convergence rate increases when the overlap increases. However, the difference between $f'_{P,R}$ and f'_P is of order $1/M$, which is very small. So a larger overlap than one atom does not significantly improve the convergence rate in comparison to that of algorithm 5.2.6.

Remark 5.2.9 (Coupling of algorithms) *The algorithm 5.2.1, that has been analyzed in Section 5.2.2.1, is not sure to converge (see Remark 5.2.1). With (5.28), one can see that, the closer h' to 1, the larger its convergence rate. If h' is too small, then the algorithm may not converge anymore.*

On the contrary, the algorithm 5.2.6, that has been studied in Section 5.2.4.1, always converges. However, its convergence rate is in general slow (see Remarks 5.2.6 and 5.2.7), and we have mentioned that, the closer $h'(u_M^k)$ to 1, the slower the convergence.

We know that $0 < h' < 1$ (see (5.32)), and we have at hand two algorithms, one that fastly converges when h' is close to 1, and one that fastly converges when h' is close to 0. An idea is then to couple these two algorithms, that is, at each iteration, to use the one whose convergence rate is the largest. In other words, denoting by u_M the current position of the atom M, we can compute $h'(u_M)$ with (5.31) and compare the convergence rate of the algorithms 5.2.1 and 5.2.6.

This algorithm always converges but has not proved computationally very efficient.

5.2.4.2 A method inspired of integral representation

When one wants to solve a PDE on an unbounded domain, one can use a method based on integral representation, which amounts to defining an artificial boundary and to solve a PDE in the so-obtained bounded domain, with an adequate boundary condition. This method is a non iterative method. We adapt here such an idea to our case. The method consists in determining only $\mathbf{u}^r_{|[0,M-1]}$. For this purpose, we

look for an equivalent problem that makes use of the fact that u^r_M can be considered as a function of u^r_{M-1}. For any $\mathbf{w} = (w_0, \ldots, w_{M-1}) \in \mathbb{R}^M$, let us set

$$\begin{aligned} E^{int}_{direct}(\mathbf{w}) &= \sum_{i=0}^{M-2} W(w_{i+1} - w_i) - \sum_{i=0}^{M-1} f_i\, w_i \\ &\quad - \frac{1}{(\bar{\delta}^{-1})_{1,1} - 1}\, W\left((\bar{\delta}^{-1})_{1,1} w_{M-1} + (\bar{\delta}^{-1})_{1,N-M} a - w_{M-1}\right), \end{aligned} \tag{5.74}$$

where $\bar{\delta}$ be the matrix of size $(N-M)\times(N-M)$ of the Laplacien operator (with Dirichlet boundary conditions equal to zero).

Theorem 5.2.7 (A direct method) *We assume that (5.8) is satisfied.*

1. *Let $\boldsymbol{w} \in \mathbb{R}^M$ be the minimizer of*

$$\inf_{\tilde{\boldsymbol{w}} \in \mathbb{R}^M} \left\{ E^{int}_{direct}(\tilde{\boldsymbol{w}});\ \tilde{w}_0 = 0 \right\}. \tag{5.75}$$

   Then $(w_0, \ldots, w_{M-1}) = (u^r_0, \ldots, u^r_{M-1})$, where $\boldsymbol{u}^r$ is the solution of the reference problem (5.7).

2. *Let $\overline{\boldsymbol{v}} \in \mathbb{R}^{N-M+2}$ be the minimizer of*

$$\inf_{\boldsymbol{y} \in \mathbb{R}^{N-M+2}} \left\{ \bar{E}^{har}(\boldsymbol{y});\ y_{M-1} = w_{M-1},\ y_N = a \right\}. \tag{5.76}$$

   Then $\boldsymbol{u}^r = (w_0, \ldots, w_{M-1} = \overline{v}_{M-1}, \overline{v}_M, \ldots, \overline{v}_N)$.

Proof: As $(\bar{\delta}^{-1})_{1,1} = 1 - \dfrac{1}{N-M+1} < 1$, the function $\tilde{\mathbf{w}} \mapsto E^{int}_{direct}(\tilde{\mathbf{w}})$ is strictly convex, so the minimization problem

$$E^{int}_{direct}(\mathbf{w}) = \inf_{\tilde{\mathbf{w}} \in \mathbb{R}^M} \left\{ E^{int}_{direct}(\tilde{\mathbf{w}});\ \tilde{w}_0 = 0 \right\} \tag{5.77}$$

is well posed. We now prove assertion 2. The vector $\overline{\mathbf{v}}$ satisfies

$$(\overline{v}_M, \ldots, \overline{v}_{N-1})^T = \bar{\delta}^{-1}\, (w_{M-1}, 0, \ldots, 0, a)^T,$$

hence

$$(\bar{\delta}^{-1})_{1,1} w_{M-1} + (\bar{\delta}^{-1})_{1,N-M} a - w_{M-1} = \overline{v}_M - w_{M-1}. \tag{5.78}$$

The vector $\mathbf{w}$, which is a minimizer of (5.75), satisfies the Euler-Lagrange equations of (5.75). With (5.78), it can be shown that the configuration $(w_0, \ldots, w_{M-1} = \overline{v}_{M-1}, \overline{v}_M, \ldots, \overline{v}_N)$ satisfies the Euler-Lagrange equations (5.9) of the reference problem. This proves assertion 2, which implies assertion 1. □

This method is a non-iterative method, unlike all the other methods that we study in the present work. It is based on the minimization of the energy E^{int}_{direct} defined by (5.74), which is not a standard molecular mechanics energy. In addition, when iteractions beyond nearest neighbours are taken into account, it is possible to generalize the method but the expression of the energy E^{int}_{direct} becomes more and more complicated. Let us also mention that this method cannot be implemented just by interfacing the two codes we have at hand.

5.2.5 Conclusions

We collect here the results of the analysis conducted so far.
- Under the simplification assumptions we have adopted, the algorithms 5.2.2 and 5.2.4 are identical.
- The algorithms 5.2.2 and 5.2.5 depend on a parameter μ, they converge if μ is small enough. The algorithm 5.2.1 corresponds to the algorithm 5.2.2 with the choice $\mu = 1$, its convergence is not guaranteed.
- An optimal choice for the parameter μ at each iteration has been proposed for the algorithm 5.2.2, thus obtaining the algorithm 5.2.3, that reads as a Newton algorithm.
- The algorithms 5.2.6 always converges. However, we expect its convergence rate to be slow.

The algorithms 5.2.1, 5.2.2 and 5.2.6 can be implemented with the two codes we have at hand. The practical implementation of the algorithm 5.2.5 raises issues.

5.3 Numerical examples

We now turn in this section to numerical tests. We first study the different algorithms on small systems (see Section 5.3.1). Based on these first tests, the best algorithms are identified and then tested on larger systems (see Section 5.3.2).

5.3.1 Comparison of the methods on small systems

In this subsection, we consider a system of $N = 100$ atoms, and we set $M = 20$. The interatomic potential is

$$W(x) = \varepsilon \frac{(x-1)^4}{4} + \frac{(x-1)^2}{2}, \quad \varepsilon \geq 0, \tag{5.79}$$

where we choose $\varepsilon = 0.5$. The potential W attains its minimum at 1. The boundary condition on the last atom is $u_N = a = 125$. The system is subjected to body forces $\{f_i\}_{i=1,M}$ such that

$$\begin{array}{rcll} f_i & = & i\,f & \text{for } i \leq 12, \\ f_i & = & 0 & \text{otherwise,} \end{array} \tag{5.80}$$

where f is a parameter. In the sequel, we choose $f = 0.006$ or $f = 0.6$. For this small system, it is possible to solve (5.7). The positions $\{u_i^r\}_{i=0}^N$ and the gradient of the deformation, that is $\left\{u_{i+1}^r - u_i^r\right\}_{i=0}^{N-1}$, are displayed on Fig. 5.3.

5.3.1.1 "Easy to implement" algorithms

We study here the algorithms 5.2.1, 5.2.2, 5.2.3 and 5.2.6, that can all be implemented with the two codes we have at hand (the algorithm 5.2.5 will be studied in Section 5.3.1.2). The algorithms 5.2.1, 5.2.2 and 5.2.6 are based on fixed point

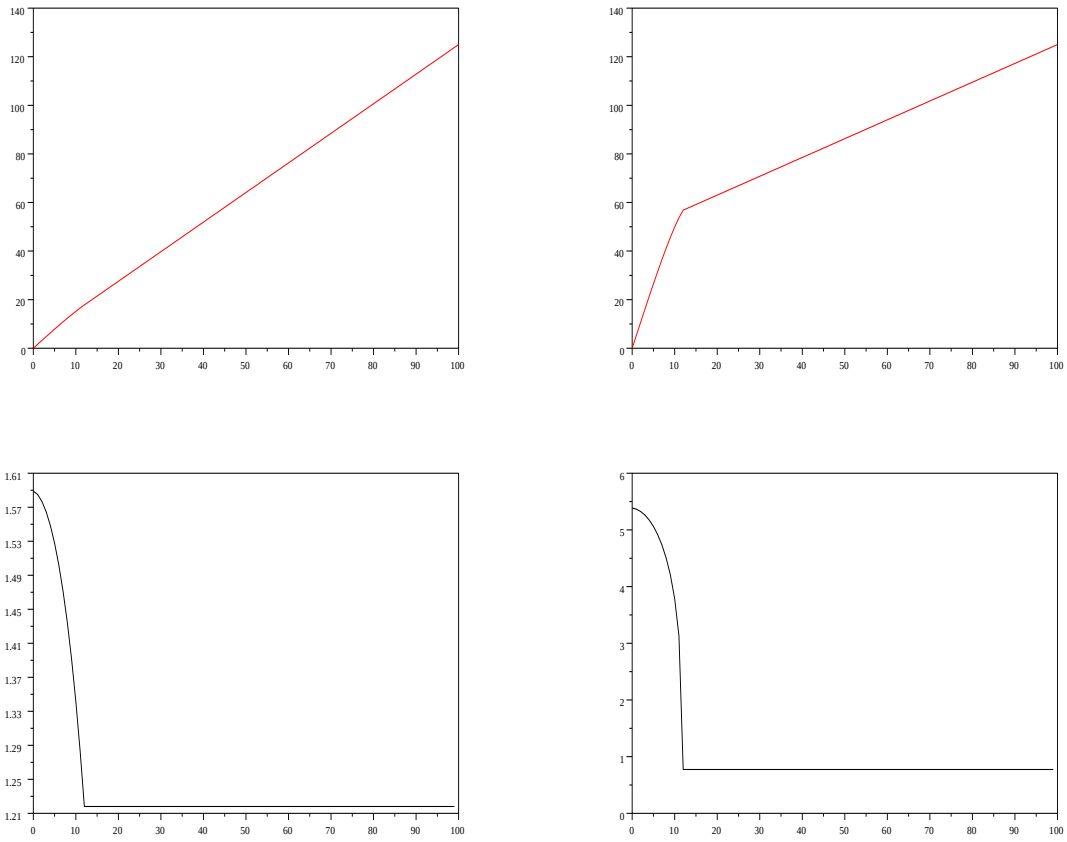

FIG. 5.3 – Equilibrium positions of the atoms (top) and gradient of deformation (bottom) of the reference solution $\mathbf{u}^r$ (small system problem ; on the left hand side, case f=0.006, and on the right hand side, case f=0.6).

iterations (the contracting mapping on which these algorithms are based is respectively f_F, $f_{F,\mu}$ and f_P, see (5.20), (5.37) and (5.71)), whereas the algorithm 5.2.3 is a Newton algorithm. The initial guess is always the linear configuration $u_i^0 = a\,i/N$.

In the case $f = 0.006$, all the algorithms converge (see Fig. 5.4). To reach a given accuracy, the algorithms 5.2.1, 5.2.2 (with $\mu = 0.8$) and 5.2.3 need the same number of iterations. The algorithm 5.2.6 converges at a slower rate than the three algorithms mentioned above.

We now turn to the case f=0.6 (see Fig. 5.5). The algorithm 5.2.1 does not converge, and the algorithm 5.2.2 converges neither with $\mu = 0.8$ nor with $\mu = 0.3$. The algorithm 5.2.6 converges very slowly. The best results are obtained with the algorithm 5.2.2 (with $\mu = 0.1$ or $\mu = 0.2$) and the algorithm 5.2.3.

This example shows that algorithms 5.2.2 and 5.2.3 are interesting : the algorithm 5.2.1 does not converge, but with a very simple modification, we obtain a converging algorithm. The algorithm 5.2.3 is also interesting because it provides guidelines to choose a value for μ. We can see that it converges more quickly than the algorithm 5.2.2 with $\mu = 0.1$ or $\mu = 0.2$, which can be expected since the algorithm 5.2.3 is a Newton algorithm, whereas the algorithm 5.2.2, where μ is kept constant, is a gradient algorithm.

On Fig. 5.6, we display the optimal value of μ as given by (5.41). One can see that, in the case $f = 0.006$, the optimal value stays close to 0.84 : this is why algorithms 5.2.1 (that corresponds to $\mu = 1$) and 5.2.2 (with $\mu = 0.8$) give good results : they correspond to a value of μ which is close to the optimal one. Indeed, as can be seen on the right hand side of Fig. 5.4, the function $f_{F,\mu}$ is almost constant for $\mu = 1$ and for $\mu = 0.8$. In the case $f = 0.6$, the optimal value of μ increases from 0.07 to 0.45 in 6 iterations (see Fig. 5.6). So working with $\mu = 1$ is a bad choice : we can see on the right hand side of Fig. 5.5 that the slope on f_F is smaller than -1, and the algorithm 5.2.1 does not converge. The same observation holds for the algorithm 5.2.2 with $\mu = 0.8$ or $\mu = 0.3$. Working with $\mu = 0.2$ or $\mu = 0.1$ is a better choice : the slope of the function $f_{F,\mu}$ is between -1 and 1, and the algorithm converges.

As the algorithm 5.2.3 is a Newton algorithm, it may be too expensive to use. However, we see on this example that, to obtain a converging algorithm, it is not necessary to use the "optimal" value of μ at each iteration. Indeed, if one uses the algorithm 5.2.2 with a constant μ that is chosen close to the optimal value computed by the algorithm 5.2.3 in the first iterations, one obtains a converging algorithm. Indeed, when $f = 0.006$, the optimal value of μ keeps close to 0.84, and the algorithm 5.2.2 with $\mu = 0.8$ and $\mu = 1$ is a converging algorithm. In the case $f = 0.6$, the same observation holds (in this case, a good choice is $\mu = 0.1$).

5.3.1.2 Uzawa algorithm

We now study the algorithm 5.2.5, on the same problem as previously. This algorithm depends on a parameter μ, we will test its influence as well as the optimal

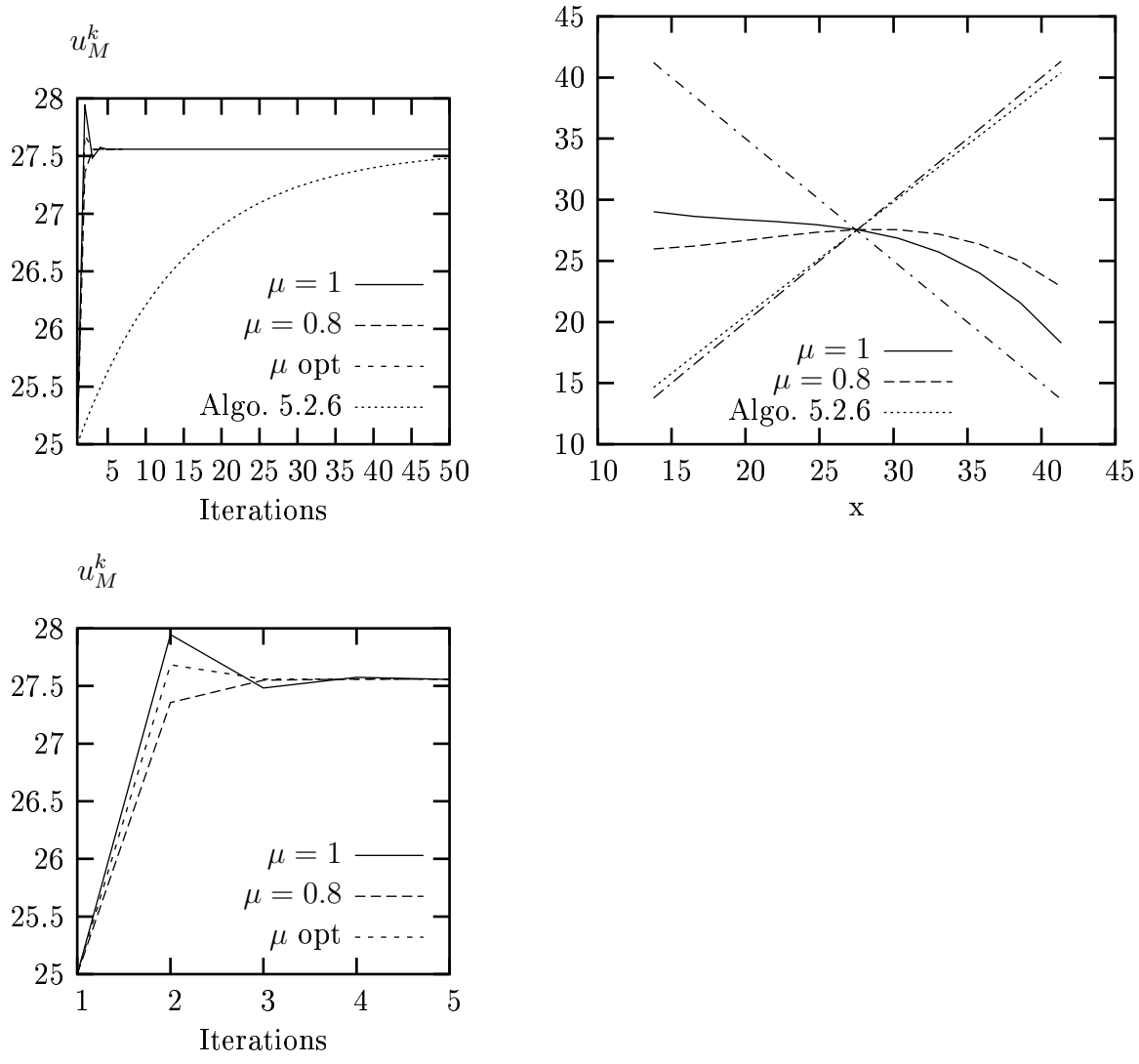

FIG. 5.4 – Convergence of the fixed point algorithms in the case f=0.006 (see (5.80)) for the small system problem. On the left hand side, evolution of the position u_M^k along the iterations k (the bottom figure is a zoom of the top figure; $\mu = 1$ corresponds to the algorithm 5.2.1, $\mu = 0.8$ corresponds to the algorithm 5.2.2 and μ opt corresponds to the algorithm 5.2.3). On the right hand side, $f_F(x)$, $f_{F,\mu}(x)$ (with $\mu = 0.8$) and $f_P(x)$ as a function of x (the curves with slopes 1 and -1 are also plotted).

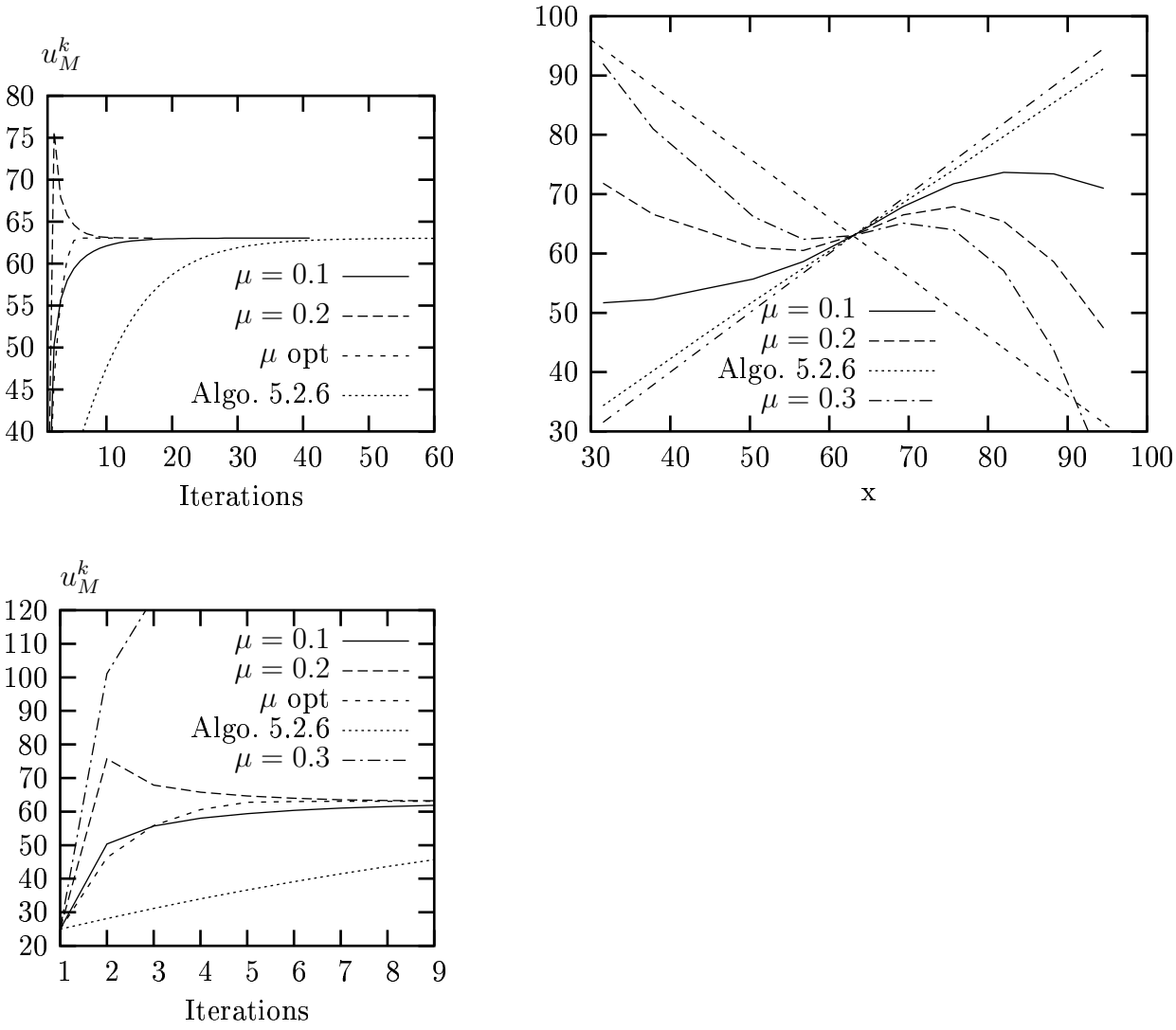

FIG. 5.5 – Convergence of the fixed point algorithms in the case f=0.6 (see (5.80)) for the small system problem. On the left hand side, evolution of the position u_M^k along the iterations k (the bottom figure is a zoom of the top figure; the algorithm 5.2.2 with $\mu = 0.3$ does not converge). On the right hand side, the functions $f_{F,\mu}$ (with $\mu = 0.1, 0.2$ and 0.3) and f_P (the curves with slopes 1 and -1 are also plotted).

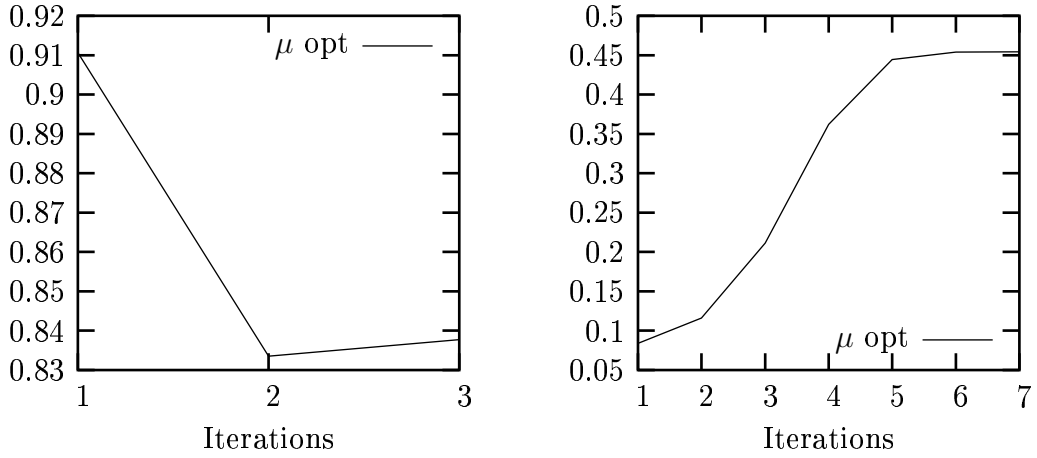

FIG. 5.6 – Optimal value of μ for the algorithm 5.2.3, as given by (5.41), along the iterations (small system problem ; on the left hand side, case f=0.006, and on the right hand side, case f=0.6).

value proposed in (5.66).

The optimal value of μ is displayed on Fig. 5.7. It is almost constant along the iterations. In both cases $f = 0.006$ and $f = 0.6$, we have $\mu^k \approx 0.012$.

As one can see on Fig. 5.8, the convergence rate strongly depends on the value of μ. When μ is smaller than this optimal value, the sequence $\left\{u_M^k\right\}_k$ is monotonous. When μ is larger than the optimal value, the sequence $\left\{u_M^k\right\}_k$ oscillates. If μ is too large, the algorithm 5.2.5 does not converge anymore.

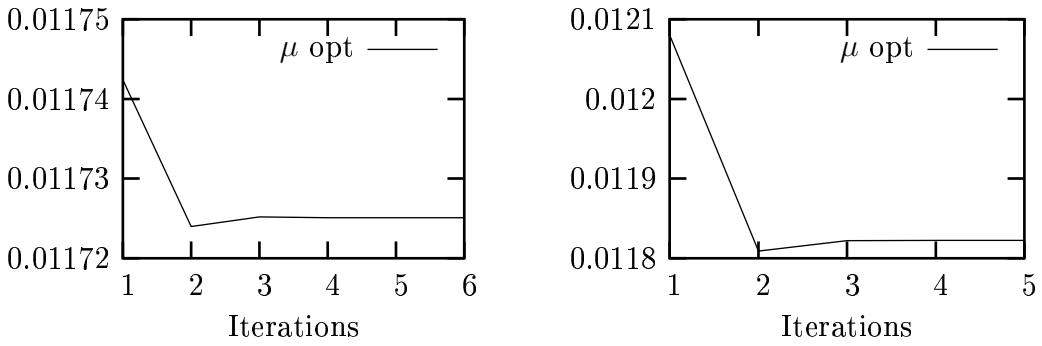

FIG. 5.7 – Optimal value of μ for the algorithm 5.2.5 along the iterations (small system problem ; left hand side : case f=0.006, right hand side : case f=0.6).

5.3.1.3 Conclusion

The most efficient algorithms are the algorithms 5.2.2, 5.2.3 and 5.2.5. However, the algorithm 5.2.5 is not easy to implement (see Section 5.2.3.2).

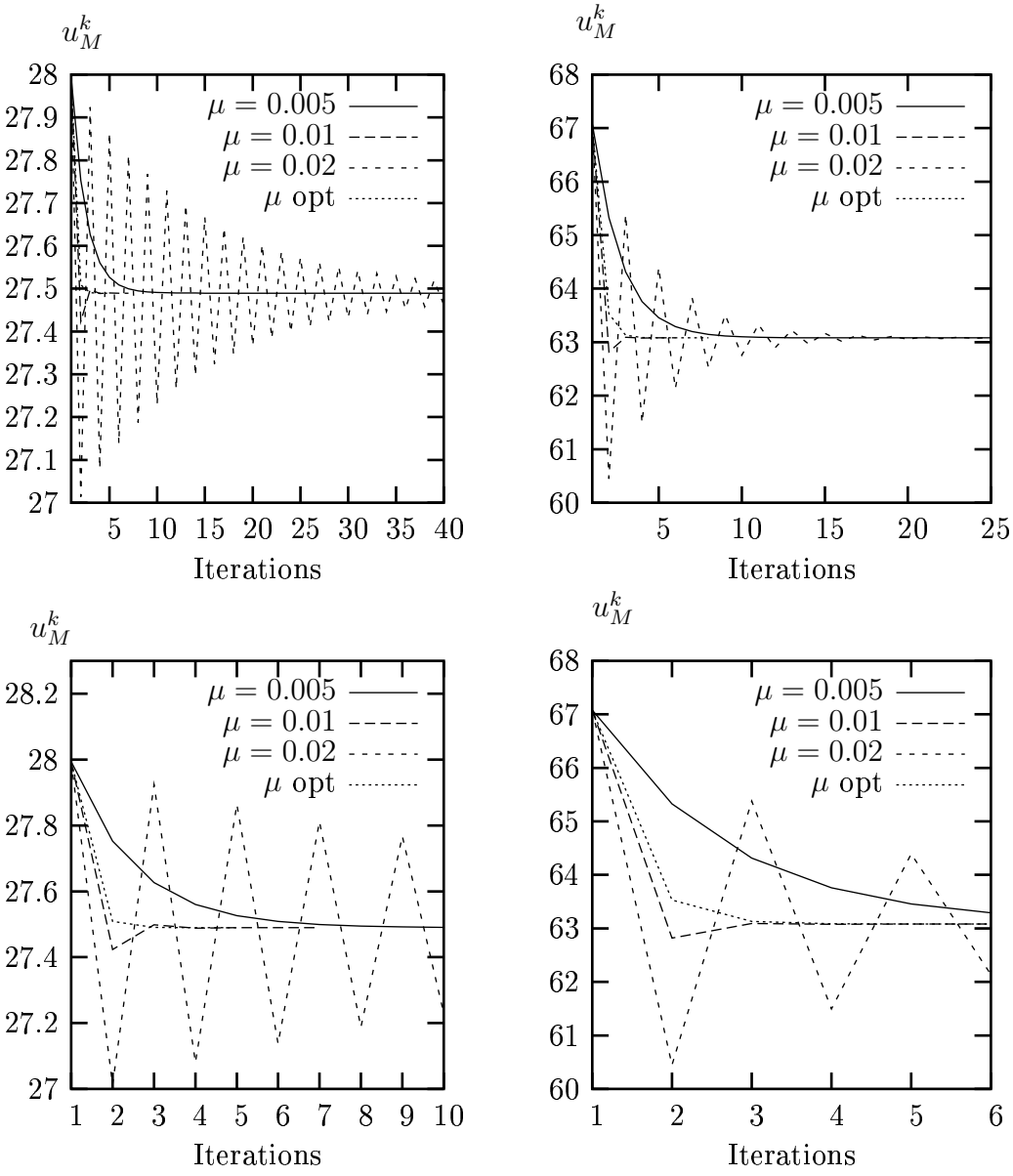

FIG. 5.8 – Algorithm 5.2.5 on the small system problem : evolution of u_M along the iterations for different values of μ (left hand side : case f=0.006 ; right hand side : case f=0.6 ; the bottom figures are a zoom of the corresponding top figures).

5.3.2 Application to larger systems

In this subsection, we test the algorithms 5.2.1, 5.2.2 and 5.2.3 on a larger system. We work again with the interatomic potential W defined by (5.79) (with again $\varepsilon = 0.5$). The whole system is now composed of $N = 10000$ atoms, and the boundary condition on the last atom is $u_N = a = 12500$. We choose to work with the following body forces :

$$\begin{aligned} f_i &= f \quad \text{for } i \leq 1200, \\ f_i &= 0 \quad \text{otherwise,} \end{aligned}$$

with $f = 0.08$.

Three algorithms will be tested : the algorithm 5.2.1, the algorithm 5.2.3 and the algorithm 5.2.2 with a constant value of μ (unless otherwise stated, the value we work with is the optimal value computed by the algorithm 5.2.3 at the second iteration). We show that, if M is large enough, then the three algorithms converge. However, the use of algorithms 5.2.2 and 5.2.3 allows one to solve the same problem with fewer degrees of freedom : there exist values of M such that the algorithms 5.2.2 and 5.2.3 converge whereas the algorithm 5.2.1 does not converge.

Let us start with the choice $M = 4000$. Then the three algorithms converge (see Fig. 5.9). The optimal value of μ, as computed by the algorithm 5.2.3, changes a lot during the first iterations, then stays constant (see Fig. 5.10, left hand side). The value at the second iteration (which is used for the algorithm 5.2.2) is $\mu = 0.55$. On the right hand side of Fig. 5.10, we can see the configuration $(u_i)_{0 \leq i \leq M}$ that has been computed.

If $M = 3000$, then the algorithm 5.2.1 does not converge anymore (the sequence $\left\{u_M^k\right\}_k$ diverges and is not represented here). The algorithms 5.2.2 and 5.2.3 converge (see Fig. 5.11). The evolution of μ^k along the iterations is not displayed but it is similar to that displayed on Fig. 5.10 : it changes a lot during the first iterations then stays almost constant.

The system we study is subjected to body forces on the atoms $i \in [0, 1200]$. So it is not possible to choose $M < 1200$, since, in all the methods that we have analyzed, we have assumed that there was no body force in the exterior domain. We now choose $M = 1300$, a value which is close to the lower bound. As in the case $M = 3000$, the algorithm 5.2.1 does not converge. The algorithm 5.2.3 converges (see Fig. 5.12). The evolution of the optimal value of μ is displayed on Fig. 5.13 : once again, it very quickly reaches a constant value. If we now use algorithm 5.2.2 with $\mu = 0.14$, which is the value of the optimal parameter at the second iteration, we obtain a converging algorithm, though with a slow convergence rate. This can be seen on the left hand side of Fig. 5.12, where we observe large oscillations of the sequence $\left\{u_M^k\right\}_k$. These large oscillations are characteristic of a too large value for

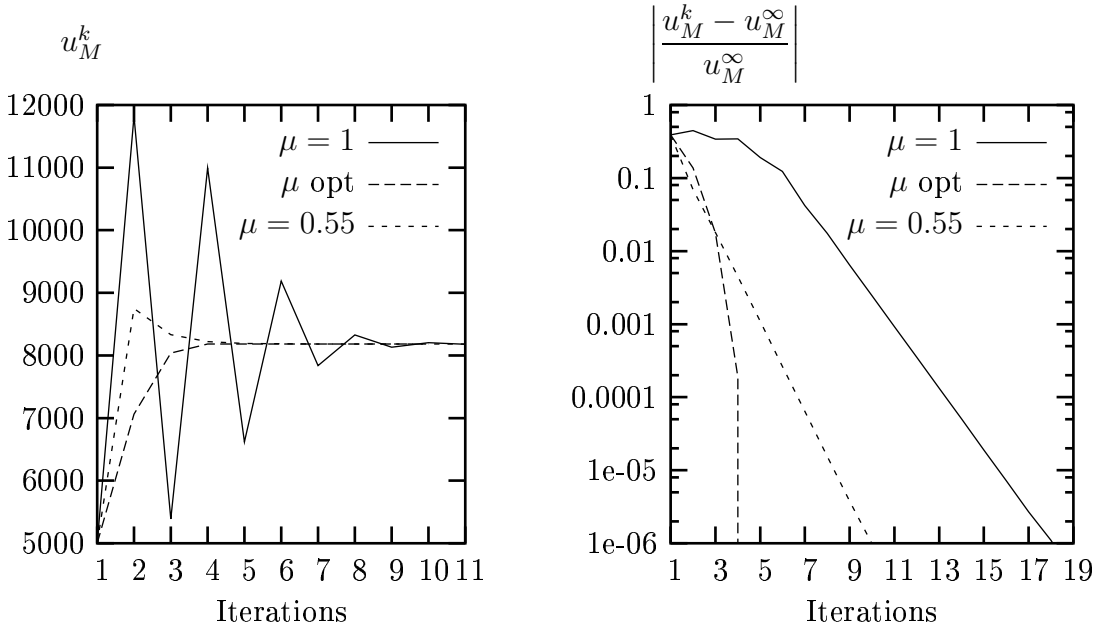

FIG. 5.9 – Evolution of u_M^k along the iterations k (case $M = 4000$) : on the left hand side, evolution of u_M^k, on the right hand side, evolution of the error $\left|\frac{u_M^k - u_M^\infty}{u_M^\infty}\right|$.

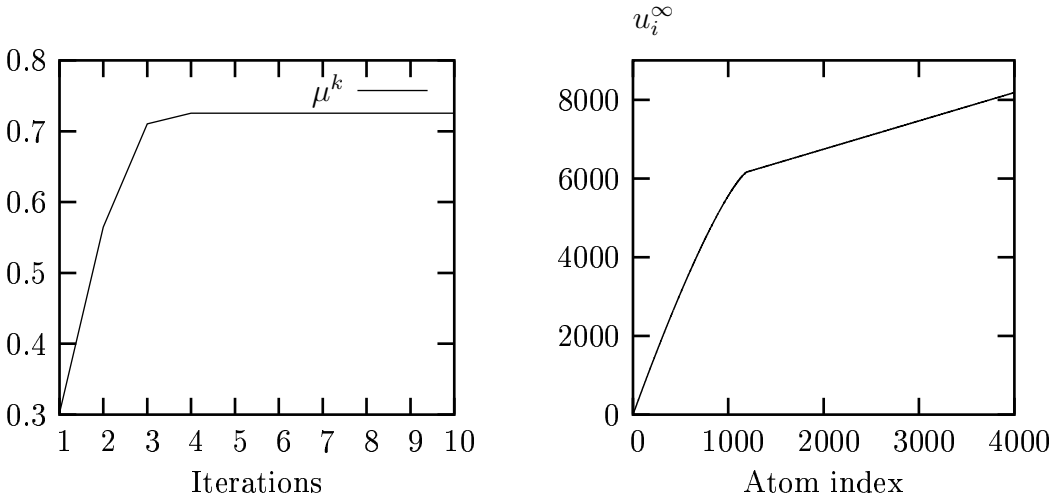

FIG. 5.10 – Left hand side : evolution of the parameter μ^k (algorithm 5.2.3, optimal value computed from (5.41)). Right hand side : equilibrium configuration of the system : the three algorithms give the same result (case $M = 4000$).

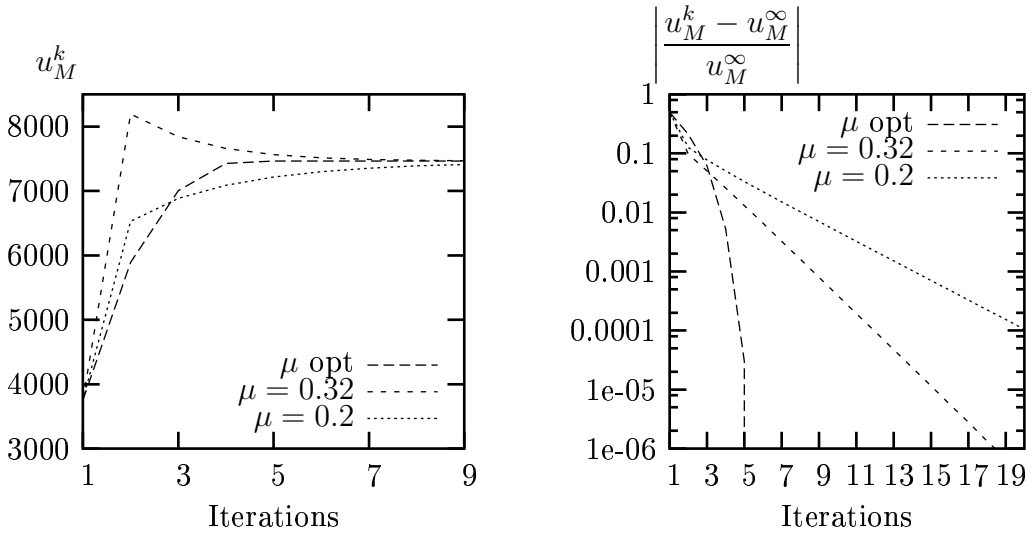

FIG. 5.11 – Case $M = 3000$. Left hand side : evolution of u_M along the iterations k. Right hand side : evolution of the error $\left|\frac{u_M^k - u_M^\infty}{u_M^\infty}\right|$.

μ. If μ is decreased to $\mu = 0.1$, then we obtain an algorithm with better converging properties.

The results obtained on this system confirm the conclusions we have made after the study of the small system : the algorithms 5.2.2 and 5.2.3 have good convergence properties, they can handle situations in which the simulated domain is just a little larger than the domain subjected to body forces. The algorithm 5.2.2 is easier to implement and to use than the algorithm 5.2.3, but it is a first order algorithm, while the algorithm 5.2.3 is a Newton algorithm.

5.4 Conclusion

Most of the methods we have presented need the resolution of the large quadratic minimization problem (5.10). For one dimensional problems with nearest neighbour interactions, an analytical solution has been proposed. In a more general case, several methods can be envisioned :

- one can take advantage of the linearity of the problem to develop specific methods to solve it (for instance by Fourier methods).
- there exists PDEs whose fundamental solution u (in an unbounded domain with the condition $u(r) \to 0$ as $|r| \to +\infty$) is known analytically. Let $G(x, y)$

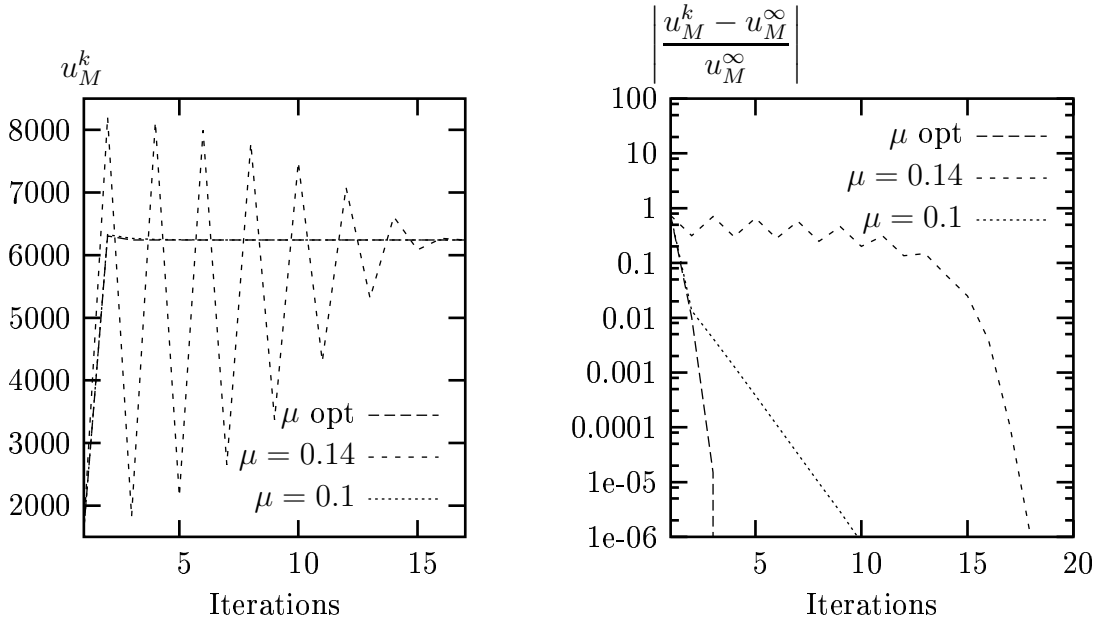

FIG. 5.12 – Evolution of u_M along the iterations (case $M = 1300$). On the left hand side, evolution of u_M^k with the algorithm 5.2.2 with $\mu = 0.14$ and $\mu = 0.1$ and with the algorithm 5.2.3 (μ opt). On the right hand side, evolution of the error $\left|\frac{u_M^k - u_M^\infty}{u_M^\infty}\right|$.

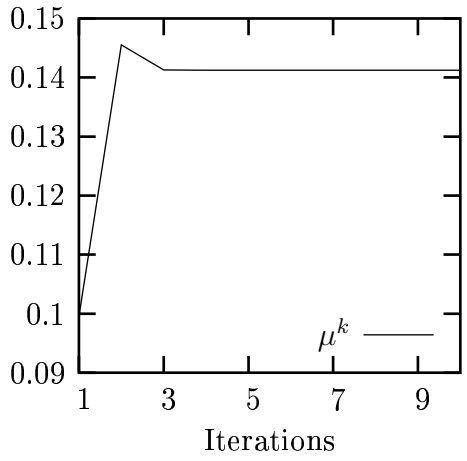

FIG. 5.13 – Evolution of μ (algorithm 5.2.3, optimal value of μ according to (5.41) ; case $M = 1300$).

be such a fundamental solution, that is the response of the system at point x for a sollicitation at point y. Then one can approximate the Green function $G_{k,p}$ of the atomistic system (the response of the atom k to a sollicitation on the atom p) by $G_{k,p} \approx G(k,p)$.
- some expressions are provided by the literature, for some special cases (see for instance [201, 202, 205]).
- another method to compute $G_{k,p}$ is to consider a sollicitation on atom p, and to compute the response of a small system, which includes the atom k, around the atom p. In doing so, we assume that atoms further away than some distance to the atom p do not feel the sollicitation.

In the one dimensional setting we have chosen, several iteratives methods to compute the equilibrium of large atomistic systems have been presented along with their numerical analysis. The most interesting algorithm we have studied is the algorithm 5.2.2, which can be recast as an Uzawa algorithm. It allows for an easy implementation and has good convergence properties.

Chapitre 6

Méthode multiéchelle couplant un modèle atomistique avec un modèle de continuum : analyse d'un cas simple

Ce chapitre a été écrit en collaboration avec Xavier Blanc et Claude Le Bris, et soumis à *Mathematical Modelling and Numerical Analysis* [P4].

Nous nous intéressons ici à un exemple simplifié de méthode multiéchelle couplant un modèle atomistique avec un modèle de continuum, dans le cas unidimensionnel. L'idée est de décrire une partie du domaine de calcul par un modèle atomistique, car des singularités à l'échelle atomistique sont attendues, et d'utiliser dans le reste du domaine un modèle plus grossier. Notre travail met l'accent sur le critère utilisé pour partitionner le domaine et décider de la zone sur laquelle l'un ou l'autre des deux modèles est utilisé. Dans le cas d'une densité d'énergie élastique convexe, nous faisons l'analyse numérique de la méthode la plus naturelle et montrons qu'elle conduit à des résultats satisfaisants. Puis un cas particulier de densité non convexe est étudié, nous soulignons alors les difficultés liées à l'approche la plus naturelle et proposons une modification de cette approche.

Analysis of a prototypical multiscale method coupling atomistic and continuum mechanics

Xavier Blanc[a], Claude Le Bris[b] and Frédéric Legoll[b,c]

[a] *Laboratoire J.L.-Lions, Université Pierre et Marie Curie, Boite courrier 187, 75252 Paris*
[b] *CERMICS, Ecole Nationale des Ponts et Chaussées, 6 et 8 avenue Blaise Pascal, 77455 Marne-la-Vallée Cedex 2*
and
MICMAC, INRIA Rocquencourt, Domaine de Voluceau, 78153 Le Chesnay Cedex
[c] *EDF R & D, Analyse et Modèles Numériques, 1, avenue du Général de Gaulle, 92140 Clamart*

blanc@ann.jussieu.fr, {lebris,legoll}@cermics.enpc.fr

In order to describe a solid which deforms smoothly in some region, but non smoothly in some other region, many multiscale methods have been recently proposed, that aim at coupling an atomistic model (discrete mechanics) with a macroscopic model (continuum mechanics). We provide here a theoretical ground for such a coupling in a one-dimensional setting. We study both the general case of a convex energy and a specific example of a nonconvex energy, the Lennard-Jones case. In the latter situation, we prove that the discretization needs to account in an adequate way for the coexistence of a discrete model and a continuous one. Otherwise, spurious discretization effects may appear. We provide a numerical analysis of the approach.

6.1 Introduction

Traditionally, mechanics makes use of a continuum description of matter [130, 135]. However, when nanoscale localized phenomena arise, the atomistic nature of material cannot be ignored : for instance, to understand how dislocations appear and propagate under a nanoindenter, one has to describe the deformed atomistic lattice. The situation is the same when the material is subjected to singular body forces, or is likely to break because of extensional forces. In all these examples, an appropriate model to describe the localized phenomena is the atomistic model, in which the solid is considered as a set of discrete particles interacting through given interatomic potentials.

Nevertheless, the size of the materials that can be simulated by only resorting to an atomistic description is very small in comparison with the size of the materials

one is interested in. Indeed, for some phenomena we have mentioned above, it is not possible to make accurate computations by just considering a small piece of material, because large scale or bulk effects have to be accounted for. For instance, crack propagation depends on the far stress field (so there is an influence of the coarse scale onto the fine scale[1]), and, at the same time, when crack propagates, it creates stress waves that modify the far stress field (so there is also a feedback influence of the fine scale onto the coarse scale).

Fortunately, in the situations we have considered above, the deformation is smooth in the main part of the solid. Hence, a natural idea is to try to take advantage of the two models, the continuous one and the atomistic one, by coupling them. The atomistic model is used in the zone where the deformation is expected to be non smooth, while the continuum description is used everywhere else. Many methods following this paradigm have been proposed and employed on realistic and complex situations : see [165, 191, 218] for some variational approaches (based on global minimizers), and [171] for time-dependent methods based on hybrid hamiltonians. Notice that alternative ways, consisting in the approximation of the variational problems with a Γ-limit approach [143], or considering local minimizers instead of global ones [183], have also been considered.

We consider here a prototypical example of a variational method that couples a fine scale model in one zone with a coarse-grained model in another zone. This example is a toy-model for more advanced methods such as the Quasi-Continuum Method [159, 161–163, 165–167]. At least from the theoretical standpoint, a first key issue in the method is the *consistency* of the two models. Indeed, if the solution to be determined is smooth, then the solution given by the coarse-grained model should be the same as that given by the fine scale model, within an error controlled by the discretization parameters. A second issue is the *adaptivity* in the determination of the zones : we need to know where to use one model rather than the other one. Several multiscale methods that we have mentioned above include such an adaptation procedure, that seems however to lack from a rigorous ground.

The present work aims at giving such a theoretical ground for the micro-macro variational approach under study. The setting is one-dimensional. It is a clear limitation of the work. We have not been able to extend our analysis to the three-dimensional mechanically relevant case and it is not clear to us, even at the formal level, which of the results contained here may survive in the three-dimensional setting. We however hope that the present study will contribute to a better understanding of the fundamental issues.

6.1.1 The atomistic and continuum problems

Let us consider a one dimensional material, occupying in the reference configuration the domain $\Omega = (0, L)$. Let u be the deformation, i.e. the map defined on Ω such that $u(x)$ is the position, in the current configuration, of a material point that

[1] At this stage, the notion of fine or coarse scales is still vague. It will be made precise below.

is at x in the reference configuration. This material is subjected to body forces f and to Dirichlet boundary conditions $u(0) = 0$ and $u(L) = a > 0$.

The solid will be described at two different space scales :

– the fine scale, at which the atomistic nature of the matter is taken into account.

– the coarse scale, which corresponds to a continuum description.

At the fine scale, the solid is considered as a set of $N + 1$ atoms, whose current positions are $(u^i)_{i=0}^N$. The energy of the system is given by

$$E_\mu(u^0, \ldots, u^N) = h \sum_{i=0}^{N-1} W\left(\frac{u^{i+1} - u^i}{h}\right) - h \sum_{i=0}^{N} u^i \, f(i\,h). \tag{6.1}$$

In this equation, W is the interaction potential between atoms and h is the atomic lattice parameter, which is linked to the number of atoms and the size of the solid by $L = Nh$. For the sake of simplicity, we will assume only nearest neighbour interactions throughout this article. Although we do not have definite conclusions to date, other cases are likely to be analyzed in the same way. The potential W is normalized so that its minimum is attained at 1. The atomistic equilibrium configuration, denoted by $u_\mu = (u_\mu^0, \ldots, u_\mu^N)$, is the solution[2] of the variational problem

$$I_\mu = \inf \left\{ E_\mu(u^0, \ldots, u^N),\ (u^0, \ldots, u^N) \in X_\mu(a) \right\}, \tag{6.2}$$

where the minimizing space is

$$X_\mu(a) = \left\{ (u^0, \ldots, u^N),\ u^0 = 0,\ u^N = a,\ \forall i,\ u^{i+1} > u^i \right\}. \tag{6.3}$$

Recall that a deformation u of the solid is mechanically admissible only if it is an injective function. As we work in a one-dimensional setting and impose $u^N = a > 0 = u^0$, a necessary and sufficient condition for injectivity is that u is increasing, thus the constraint $u^{i+1} > u^i$ in (6.3).

On the other hand, at the coarse scale, the solid deformation is described by a map $u : \Omega \to \mathbb{R}$ chosen in the variational space

$$X_M(a) = \left\{ u \in H^1(\Omega),\ u(0) = 0,\ u(L) = a,\ u \text{ is increasing on } \Omega \right\}. \tag{6.4}$$

The energy of the system reads

$$E_M(u) = \int_\Omega W(u'(x))\, dx - \int_\Omega f(x)\, u(x)\, dx. \tag{6.5}$$

We will give below (see Section 6.2.1) more precise assumptions to ensure that the energy is well-defined as soon as $u \in H^1(\Omega)$. The equilibrium configuration, denoted by $u_M(x)$, is a solution of the variational problem

$$I_M = \inf \left\{ E_M(u),\ u \in X_M(a) \right\}. \tag{6.6}$$

[2]Existence and uniqueness of solutions will be discussed in the next sections.

Remark 6.1.1 *In a two- or three-dimensional setting, some sufficient conditions for the injectivity of a map are given in [130] (see pp. 222-231).*

The question we address in the present work concerns the approximation of problem (6.2). Indeed, for any deformation u of the material, the energy is given by (6.1), but the number of atoms to be considered in the sum, typically of the order 10^{23} in a macroscopic sample of material, makes the computation of (6.1) untractable in practice.

When the deformation u is regular and fixed, it has be shown in [142] that the atomistic energy $E_\mu(u(0), u(h), \ldots, u(Nh))$ converges to $E_M(u)$ when the atomic lattice parameter h goes to 0 and the number of atoms goes to infinity such that Nh remains constant, $Nh = L$. This result ensures the above mentioned *consistency* of the two descriptions, (6.1) on the one hand and (6.5) on the other hand, and also provides with an economical way to compute the sum (6.1), namely by approximating the integral (6.5). It remains that, when deformations that are expected to play a role are not regular enough to allow for the above convergence, the only way to compute the energy seems to be resorting to the atomistic expression (6.1). An economical approach is the coupled approach we consider in the present work.

6.1.2 A coupled problem

Let $\Omega_M \subset \Omega$ be an open subset of the solid in which the deformation u is supposed to be smooth enough so that the atomistic expression of the energy may be replaced by the continuum one. Throughout this article, we suppose that Ω_M satisfies the following property :

Property 6.1.1 *For simplicity, we suppose*

$$\Omega_M = \cup_{j=1}^{J} (a_j h, b_j h) \ \subset \Omega, \tag{6.7}$$

with $a_j, b_j \in \{0, \ldots, N\}$, $a_j < b_j < a_{j+1}$, and where the number J of connected components of Ω_M is bounded by $\mathcal{N}_{cc}$, where $\mathcal{N}_{cc}$ is a given fixed parameter (see Fig. 6.1).

Definition 6.1.1 *Let us denote by*

$$\mathbb{N}_\mu = \left\{ i \in \{0, \ldots, N-1\};\ ih \in \Omega_\mu \text{ and } ih + h \in \Omega_\mu \right\} \tag{6.8}$$

the set of indices i such that both atoms i and $i+1$ are contained in Ω_μ.

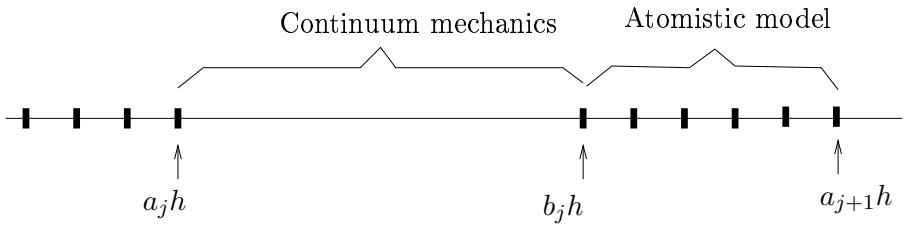

FIG. 6.1 – Partition of Ω into a regular zone Ω_M where the continuum mechanics model is used, and a singular zone Ω_μ where the atomistic model is used. By definition, we have $a_jh \in \Omega_\mu$ and $b_jh \in \Omega_\mu$.

With (6.7), we can see that $\mathbb{N}_\mu = \cup_{j=1}^{J-1}\{b_j, \dots, a_{j+1} - 1\}$.

For any partition $\Omega = \Omega_M \cup \Omega_\mu$, the solid deformation can be described by an element of the hybrid (atomistic/continuum) space

$$X_c(a, \Omega_M) = \left\{ \begin{array}{c} u;\, u_{|\Omega_M} \in X_W(\Omega_M), u_{|\Omega_\mu} \text{is the discrete set of} \\ \text{variables } (u^i)_{ih\in\Omega_\mu},\ u(0) = 0,\ u(L) = a, \\ u\left((a_jh)^+\right) = u^{a_j},\ u\left((b_jh)^-\right) = u^{b_j}, \\ u \text{ is increasing on } \Omega \end{array} \right\}, \tag{6.9}$$

where $X_W(\Omega_M)$ is some functional space that depends on the potential W and that will be made precise below. We have written down the boundary conditions supposing that 0 and L are in $\overline{\Omega_M}$ (otherwise, adequate and simple modifications are in order). The last line in (6.9) is equivalent to the injectivity of u.

A natural idea is to first fix Ω_M, next, for each $u \in X_c(a, \Omega_M)$, to define the energy of the deformed system by

$$E_c(u) = h \sum_{i\in\mathbb{N}_\mu} W\left(\frac{u^{i+1} - u^i}{h}\right) - h \sum_{i, ih\in\Omega_\mu} u^i f(ih) + \int_{\Omega_M} W(u') - uf\ , \tag{6.10}$$

and finally to state a minimization problem at Ω_M fixed,

$$I_c(\Omega_M) = \inf\{E_c(u),\ u \in X_c(a, \Omega_M)\}\,. \tag{6.11}$$

The point is unfortunately that Ω_M is difficult, or even impossible, to determine in advance. In fact, it depends on the minimizer u_μ of (6.2) as it should, vaguely stated, consist of all the zones of regularity of u_μ (in order to allow for both an economical and correct evaluation of the energy), and u_μ, which is the reference (ideal) solution, cannot be computed. Our way to treat the difficulty is the following. In the case of a *convex* interaction potential W, we show that the zones of regularity of u_μ can be approximated by a set, again denoted by Ω_M, that can be computed. This approximation is so that :

1. when the atomistic solution u_μ is smooth in some region of Ω, the set Ω_M is a domain embedding this region ;
2. the minimization problem (6.11) is theoretically well posed ;
3. an algorithm can be proposed to compute a solution of (6.11) ;
4. the error between the computed solution and the reference solution u_μ may be estimated.

The above program is fulfilled in Section 6.2.

In Section 6.3, we show that any minimization problem set with energy (6.10) is ill posed when *nonconvex* interaction potentials W are used (which is the mechanically relevant case). The examination is performed on a special case of a nonconvex elastic energy density W, that is the Lennard-Jones case. Both the continuum model and the atomistic model are unable to sustain traction (see Sections 6.3.1 and 6.3.2), and a fracture appears for any extensional load. With the coupled model, a spurious effect appears in the energy functional (6.10) : the comparison of the energy of a fracture in the zone Ω_μ with that of the same fracture in the zone Ω_M shows that the energetically most favorable situation is the latter (see Section 6.3.3). This rules out the possibility of ever self-consistently adapting the zone to the singularities of the deformation and leads us to a modification of the coupled energy : instead of defining it by (6.10), we define it by

$$E_{\rm mod}(u) = h \sum_{i \in \mathbb{N}_\mu} W\left(\frac{u^{i+1} - u^i}{h}\right) - h \sum_{i, ih \in \Omega_\mu} u^i f(ih) + \int_{\Omega_M} W^h(u') - uf\,, \quad (6.12)$$

with

$$W^h(r) = W(r) + \sqrt{h}\ \tau(r - r_0).$$

In the latter relation, r_0 is some threshold parameter (to be made precise below) and the function τ, which does not depend on h, is a regularization of the function $t \in \mathbb{R} \mapsto t_+ = \max(0, t)$. The minimization problem associated to the energy (6.12) reads

$$I_{\rm mod}(\Omega_M) = \inf\left\{E_{\rm mod}(u),\ u \in X_c(a, \Omega_M)\right\}. \quad (6.13)$$

This modification remedies to the above obstruction, and is also consistent both with the atomistic model energy (6.1) and the continuum mechanics model energy (6.5). Again, as in the convex case, Ω_M should consist of all the zones of regularity of the reference solution u_μ. In Sections 6.3.4 and 6.3.5, we show that this set can be approximated by a set (again denoted by Ω_M) that can be computed. We then show that the solution of the so-obtained problem (6.13) is a converging approximation of the reference solution u_μ, and that, when the solid is subjected to an extensional load, the atomistic domain $\Omega_\mu = \Omega \setminus \Omega_M$ contains the fracture.

6.1.3 Outline of the results

We wish to point out that the main purpose of the present work is to study the coupled (atomistic/continuum) models (6.11) and (6.13), especially in the nonconvex

case, because this is the mechanically relevant case and the most interesting case from numerical analysis standpoint. But before going to this, we need to lay some groundwork. This is the reason why we first study in Section 6.2 the continuum, the atomistic and the coupled problems in the convex case. We also need to study the continuum and the atomistic problems in the nonconvex case (this is done in Sections 6.3.1 and 6.3.2). On the ground of this preliminary work, we address the key models and study the coupled problems in the nonconvex case.

The fully atomistic (resp. continuum) problems of Sections 6.2 and 6.3 have been addressed before in the literature (see [130, 135, 139, 148, 210]). But in the absence of a systematic mathematical study and with a view to self-consistency, we prefer to devote a large part of the present work to them. The reader, either not mathematically oriented, or familiar with such studies, may directly proceed to the study of the coupled models, which is presented in Sections 6.2.2 and 6.2.3 for the convex case and in Sections 6.3.3, 6.3.4 and 6.3.5 for the nonconvex case, where the most original results are discussed.

We collect now the main results of our work. At this stage, they are somewhat vaguely stated, but they will be made precise below.

1. The atomistic problem (6.2) and the continuum problem (6.6) are well posed in the convex case (see Lemma 6.2.1) and in the Lennard-Jones case (see Theorems 6.3.1 and 6.3.3).
2. In the convex case,
   - for any $\Omega_M \subset \Omega$, the coupled problem (6.11) is well posed (see Lemma 6.2.1) ;
   - it is possible to define Ω_M in such a way that the solution of the coupled problem (6.11) is a converging approximation of the solution of the atomistic problem (6.2) (see Definition 6.2.2 and Theorem 6.2.1).
3. In the Lennard-Jones case,
   - for any $\Omega_M \subset \Omega$, the coupled problem (6.13) is well posed (see Theorem 6.3.5) and if a fracture appears, it is located in Ω_μ ;
   - it is possible to define Ω_M in such a way that the solution of the modified coupled problem (6.13) is a converging approximation of the solution of the atomistic problem (6.2) (see Theorem 6.3.6 and Definition 6.3.1).

6.2 The case of a convex elastic energy density W

Let us first make precise the hybrid space $X_c(a, \Omega_M)$ defined in (6.9). In this section, we set $X_W(\Omega_M) = H^1(\Omega_M)$.

6.2.1 Properties of the variational problems

In this subsection, we provide conditions ensuring that the variational problems we consider are well-posed.

Definition 6.2.1 *We suppose that the body forces f satisfy*

$$f \in \mathcal{C}^0\left(\overline{\Omega}\right). \tag{6.14}$$

Let us define F_M and F_μ by

$$\forall x \in \Omega,\ F_M(x) = \int_0^x f(s)\,ds, \tag{6.15}$$

$$F_\mu^0 = 0 \quad \textit{and} \quad \forall i \in \{1,\dots,N\},\ F_\mu^i = h\sum_{j=1}^{i} f(jh). \tag{6.16}$$

For any Ω_M, we also define a function F_c as follows : on Ω_M, we set, for all $x \in (a_j h, b_j h)$, $j \in \{1,\dots,J\}$,

$$F_c(x) = \int_{\Omega_M \cap (0,x)} f(s)\,ds + h\sum_{k=1}^{j-1}\left(f(b_k h) + f(b_k h + h) + \dots + f(a_{k+1}h)\right), \tag{6.17}$$

whereas, on Ω_μ, for all $j \in \{1,\dots,J-1\}$, we set, for all $i \in \{b_j,\dots,a_{j+1}\}$,

$$F_c^i = \int_{\Omega_M \cap (0,ih)} f(s)\,ds + h\sum_{k=1}^{j-1}\left(f(b_k h) + \dots + f(a_{k+1}h)\right) + h\sum_{q=b_j}^{i} f(qh). \tag{6.18}$$

We note that F_c is continuous on Ω_M, continuous at $a_j h$, but not continuous at $b_j h$. In the sequel on this section, we assume that the elastic energy density W satisfies

$$\left\{\begin{array}{l} W \in \mathcal{C}^2(\mathbb{R}), \\ \exists \alpha > 0,\ \forall x \in \mathbb{R},\quad \alpha \leq W''(x), \\ \exists \beta > 0, \forall x \in \mathbb{R},\quad |W'(x)| \leq \beta\,|x-1|\,. \end{array}\right. \tag{6.19}$$

Although W is defined on $\mathbb{R}$, we need in fact to know $W(x)$ only for $x > 0$, due to the injectivity constraint that is included in the variational spaces (6.3), (6.4) and (6.9). Let us set

$$a_M^* = \int_\Omega (W')^{-1}\left(W'(0) + \sup_\Omega F_M - F_M(x)\right)\,dx, \tag{6.20}$$

$$a_\mu^* = h\sum_{i=0}^{N-1} (W')^{-1}\left(W'(0) + \left(\sup_{0\leq i\leq N-1} F_\mu^i\right) - F_\mu^i\right), \tag{6.21}$$

$$\begin{aligned} a_c^* = &\int_{\Omega_M} (W')^{-1}\left(W'(0) + \overline{F_c} - F_c(x)\right)\,dx \\ &+ h\sum_{i\in\mathbb{N}_\mu} (W')^{-1}\left(W'(0) + \overline{F_c} - F_c^i\right), \end{aligned} \tag{6.22}$$

where $\overline{F_c} = \sup\left(\sup_{x\in\Omega_M} F_c(x), \sup_{i\in\mathbb{N}_\mu} F_c^i\right)$.

Lemma 6.2.1 (Existence and uniqueness of solutions) *Let Ω_M be a fixed subdomain of Ω. We assume that the elastic energy density W satisfies (6.19) and that the body forces f satisfy (6.14).*

If $a > a_M^*$, then the continuum problem (6.6) has a unique minimizer u_M, which is in addition in $H^2(\Omega)$. If $a < a_M^*$, the problem (6.6) is not attained.

If $a > a_\mu^*$, then the atomistic problem (6.2) has a unique minimizer u_μ. If $a \leq a_\mu^*$, then the problem (6.2) is not attained.

If $a > a_c^*$, then the coupled problem (6.11) has a unique minimizer u_c. If $a < a_c^*$, the problem (6.11) is not attained.

Proof: The proofs of the three problems follow the same pattern and, for brevity, we only detail here that of the continuum problem (6.6). We first assume $a > a_M^*$. Let us show the existence of a minimizer. For this purpose, let us consider the following minimization problem, *without* the increasing condition :

$$\inf \left\{ E_M(u),\ u \in H^1(\Omega),\ u(0) = 0,\ u(L) = a \right\}. \tag{6.23}$$

With the second line of (6.19), we see that E_M is a α-convex function. The third line of (6.19) implies that E_M is continuous for the H^1 norm. Thus, problem (6.23) has a unique minimizer u_M. We now show that u_M is a minimizer of (6.6). The Euler-Lagrange equation of (6.23) reads

$$u_M'' W''(u_M') + f = 0 \quad \text{a.e. on } \Omega. \tag{6.24}$$

Assumptions (6.19) imply that $u_M \in H^2(\Omega)$. Integrating the previous equation from 0 to x, one obtains $u_M'(x) = (W')^{-1}(\lambda_M - F_M(x))$, where $\lambda_M = W'(u_M'(0)$. Let us set

$$\lambda_M^* = W'(0) + \sup_\Omega F_M,$$

so $a_M^* = \displaystyle\int_\Omega (W')^{-1} (\lambda_M^* - F_M(x))\, dx$. We also have

$$a \ = \ \int_\Omega u_M'(x)\, dx \ = \ \int_\Omega (W')^{-1}(\lambda_M - F_M(x))\, dx.$$

Since $(W')^{-1}$ is an increasing function, we see that $a - a_M^*$ has the same sign as $\lambda_M - \lambda_M^*$.

As $a > a_M^*$, we have $\lambda_M > \lambda_M^*$, hence $u_M' > 0$ on Ω. So u_M is an increasing function, it belongs to the variational space (6.4) and, since it is a minimizer of (6.23), it is a minimizer of (6.6). So, the problem (6.6) has at least one minimizer. Moreover, this minimizer is unique by convexity arguments.

We now turn to the case $a < a_M^*$. One can prove, by a standard technique of regularization [131], that the relaxed problem

$$\inf \left\{ E_M(u),\ u \in H^1(\Omega),\ u(0) = 0,\ u(L) = a,\ u' \geq 0 \text{ a.e. on } \Omega \right\}$$

has a unique minimizer $\tilde{u}_M \in H^2(\Omega)$. The condition $a < a_M^*$ implies that there exists an interval on which $\tilde{u}_M'(x) = 0$. So $\tilde{u}_M$ is not injective, and the problem (6.6) is not attained. □

Before proceeding further, let us write the Euler-Lagrange equation for (6.2). Under the assumption $a > a_\mu^*$, the constraint $u^{i+1} > u^i$ is not active, thus

$$\forall i \in \{1, \dots, N-1\}, \quad W'\left(\frac{u_\mu^i - u_\mu^{i-1}}{h}\right) - W'\left(\frac{u_\mu^{i+1} - u_\mu^i}{h}\right) - hf(ih) = 0. \quad (6.25)$$

An equivalent formulation is obtained easily :

$$\forall i \in \{0, \dots, N-1\}, \quad \frac{u_\mu^{i+1} - u_\mu^i}{h} = (W')^{-1}\left(\lambda_\mu - F_\mu^i\right), \quad (6.26)$$

where $\lambda_\mu = W'\left((u_\mu^1 - u_\mu^0)/h\right)$ and F_μ is defined by (6.16). With similar arguments, it can be shown that the minimizer u_c of (6.11) satisfies

$$\begin{aligned} &\forall x \in \Omega_M, \quad u_c'(x) = (W')^{-1}\left(\lambda_c - F_c(x)\right), \\ &\forall i \in \mathbb{N}_\mu, \quad \frac{u_c^{i+1} - u_c^i}{h} = (W')^{-1}\left(\lambda_c - F_c^i\right), \end{aligned} \quad (6.27)$$

where $\lambda_c = W'(u_c'(0))$ (recall we have assumed $0 \in \overline{\Omega_M}$), the set $\mathbb{N}_\mu$ is defined by (6.8) and F_c is defined by (6.17) and (6.18).

6.2.2 Definition of the partition

We now introduce a criterion in order to define the subdomain Ω_M.

Definition 6.2.2 (Partition in the convex case) *We assume that (6.14) is satisfied. Let $\kappa_f > 0$. We say that the interval $(ih, ih+h)$ is a* regular interval *if $f \in W^{1,1}(ih, ih+h)$ with*

$$\forall x \in (ih, ih+h),\ |f(x)| \leq \kappa_f \quad \textit{and} \quad \int_{ih}^{ih+h} |f'(x)|\, dx \leq h\frac{\kappa_f}{L}.$$

We define

$$\Omega_M = \cup^*_{(ih, ih+h)\ \textit{regular}}(ih, ih+h) \quad \textit{and} \quad \Omega_\mu = \Omega \setminus \Omega_M,$$

where $\cup^*$ means that the point $\{ih\}$ is also included in Ω_M if both $(ih-h, ih)$ and $(ih, ih+h)$ are regular intervals.

Remark 6.2.1 *The subdomain Ω_M as defined above only depends on f. Actually, due to the convexity of W, which allows for elliptic regularity results on the Euler-Lagrange equation (6.25), the singularities of the solution u_μ are solely linked to the singularities of the body forces f. So the subdomain Ω_M as defined above is the zone of regularity of u_μ. The situation will be radically different in the Lennard-Jones case (see Definition 6.3.1).*

By construction, f and f' are bounded on Ω_M :

$$\|f\|_{L^\infty(\Omega_M)} \leq \kappa_f \quad \text{and} \quad \|f'\|_{L^1(\Omega_M)} \leq \kappa_f. \tag{6.28}$$

In Definition 6.2.1, we have introduced the functions F_M, F_μ and F_c (see (6.15), (6.16), (6.17) and (6.18)). In the sequel, we will need an estimate of their difference, which is provided by the following lemma.

Lemma 6.2.2 *We assume that the body forces satisfy (6.14) and (6.28). Then*

$$\lim_{h\to 0} \sup_k \left| F_M(kh) - F_\mu^k \right| = 0, \tag{6.29}$$

and, for all h,

$$\begin{aligned} &\forall k \text{ s.t. } kh \in \Omega_M, \quad |F_c(kh) - F_\mu^k| \leq h\kappa_f(\mathcal{N}_{cc}+1), \\ &\forall k \text{ s.t. } kh \in \Omega_\mu, \quad |F_c^k - F_\mu^k| \leq h\kappa_f(\mathcal{N}_{cc}+1). \end{aligned} \tag{6.30}$$

We skip the proof of this lemma, which relies on a Taylor expansion.

6.2.3 Comparison of the atomistic problem and the coupled problem

In this subsection, we assume that the partition is defined according to Definition 6.2.2. To simplify the notation and since there is no ambiguity, we denote by I_c instead of $I_c(\Omega_M)$ the infimum (6.11).

We now estimate how the coupled problem (6.11) approximates the atomistic problem (6.2). For this purpose, we need the following operators :

Definition 6.2.3 *The operator $\Pi_c : v_\mu \in X_\mu(a) \mapsto \Pi_c v_\mu \in X_c(a, \Omega_M)$ is the interpolation-on-Ω_M operator defined by*

$$\forall x \in \Omega_M,\ (\Pi_c v_\mu)(x) = v_r(x), \quad \textit{and} \quad \forall ih \in \Omega_\mu,\ (\Pi_c v_\mu)^i = v_\mu^i,$$

where v_r is the piecewise linear interpolate of v_μ, at points ih, in Ω_M.

The operator $\Pi_\mu : v_c \in X_c(a, \Omega_M) \mapsto \Pi_\mu v_c \in X_\mu(a)$ is the evaluation operator defined by

$$\forall ih \in \Omega_M,\ (\Pi_\mu v_c)^i = v_c(ih), \quad \textit{and} \quad \forall ih \in \Omega_\mu,\ (\Pi_\mu v_c)^i = v_c^i.$$

To estimate the distance between the respective minimizers u_μ and u_c of problems (6.2) and (6.11), we will argue on the basis of (6.26) and (6.27), that provide expressions of u_μ and u_c as functions of the Lagrange multipliers λ_μ and λ_c.

Remark 6.2.2 *An alternative method to evaluate $u_\mu - u_c$ is based on the observation that minimizing E_μ over $X_\mu(a)$ is equivalent to minimizing E_c over a finite element space $X_c^h(a,\Omega_M) \subset X_c(a,\Omega_M)$ of mesh size h, where the body force integral term is now computed by a numerical integration formula (namely, a Riemann sum). So standard results of FEM theory can be applied in order to obtain an H^1 estimate on $u_\mu - u_c$. Derivating L^∞ estimates from such an argument is more tricky (see [127, 136, 145, 146]), and this is the reason why we prefer here a self-contained argument that we present in the proof of Theorem 6.2.1.*

For any $u \in X_c$, we define

$$\|u\|_{L^\infty(X_c)} = \max\left(\|u\|_{L^\infty(\Omega_M)}, \max_{i,ih\in\Omega_\mu} |u^i|\right), \tag{6.31}$$

$$|u|_{W^{1,\infty}(X_c)} = \max\left(\|u'\|_{L^\infty(\Omega_M)}, \max_{i\in\mathbb{N}_\mu} \left|\frac{u^{i+1}-u^i}{h}\right|\right), \tag{6.32}$$

and for any $u \in X_\mu$, we define

$$\|u\|_{L^\infty(X_\mu)} = \max_{i\in[0,N]} |u^i| \quad \text{and} \quad |u|_{W^{1,\infty}(X_\mu)} = \max_{i\in[0,N-1]} \left|\frac{u^{i+1}-u^i}{h}\right|. \tag{6.33}$$

Lemma 6.2.3 *We assume that (6.14) and (6.19) are satisfied, and that*

$$a > a_M^*, \tag{6.34}$$

where a_M^* is defined by (6.20). We assume that the partition $\Omega = \Omega_M \cup \Omega_\mu$ is defined according to Definition 6.2.2 for some $\kappa_f > 0$.

Then there exists $h_0 \leq 1$ (which depends on κ_f) such that, for all $h \leq h_0$, problems (6.2) and (6.11) have a unique minimizer.

Proof: With (6.29), (6.30) and (6.34), we see that, for h small enough, we have $a > a_\mu^*$ and $a > a_c^*$, where a_μ^* and a_c^* are defined by (6.21) and (6.22). So Lemma 6.2.1 can be applied, it shows that problems (6.2) and (6.11) are well posed. □

Theorem 6.2.1 *We assume that (6.14), (6.19) and (6.34) are satisfied, and that the partition $\Omega = \Omega_M \cup \Omega_\mu$ is defined according to Definition 6.2.2 for some $\kappa_f > 0$. Let Π_c and Π_μ be the operators defined in Definition 6.2.3.*

Then there exist $h_0 \leq 1$ (which depends on κ_f) and constants C_1 and $C_i(\kappa_f)$, $i = 2,\ldots,5$, such that, for all $h \leq h_0$, the minimizers u_μ of (6.2) and u_c of (6.11) satisfy

$$|u_c|_{W^{1,\infty}(X_c)} \leq C_1 \quad \textit{and} \quad |u_\mu|_{W^{1,\infty}(X_\mu)} \leq C_1, \tag{6.35}$$

and are at a distance of order h from one another in the sense that

$$|(\Pi_\mu u_c) - u_\mu|_{W^{1,\infty}(X_\mu)} \leq C_2(\kappa_f) h \kappa_f, \tag{6.36}$$

$$|u_c - (\Pi_c u_\mu)|_{W^{1,\infty}(X_c)} \leq C_2(\kappa_f) h \kappa_f, \tag{6.37}$$

$$\|(\Pi_\mu u_c) - u_\mu\|_{L^\infty(X_\mu)} \leq C_3(\kappa_f) h \kappa_f, \tag{6.38}$$

$$\|u_c - (\Pi_c u_\mu)\|_{L^\infty(X_c)} \leq C_4(\kappa_f) h \kappa_f, \tag{6.39}$$

while the energy infima also differ of an order h :

$$|I_c - I_\mu| \leq C_5(\kappa_f) h \kappa_f. \tag{6.40}$$

The constant C_1 does not depend on κ_f, whereas the functions $\kappa_f \mapsto C_i(\kappa_f)$, $i = 2, \ldots, 5$, are bounded on any compact.

Remark 6.2.3 *The proof yields the following explicit expressions for the constants C_i, $i = 1, \ldots, 5$:*

$$C_1 = \frac{L}{4\alpha}\|f\|_{L^\infty(\Omega)} + 4\frac{a}{L}, \qquad C_2(\kappa_f) = \frac{2+\mathcal{N}_{cc}}{\alpha}\left(1 + \frac{\beta_K(\kappa_f)}{\alpha}\right), \tag{6.41}$$

$$C_3(\kappa_f) = L\, C_2(\kappa_f), \quad C_4(\kappa_f) = C_2(\kappa_f)\, h + C_3(\kappa_f), \tag{6.42}$$

$$C_5(\kappa_f) = 2a + L\left(\beta\, C_2(\kappa_f)(C_1+1) + C_3(\kappa_f)\|f\|_{L^\infty(\Omega)}\right), \tag{6.43}$$

where $\beta_K(\kappa_f) = \max_K W''\left(W'(\cdot)\right)$. Here, K is the closed interval of center 0 and of radius $h_0\kappa_f(1+\mathcal{N}_{cc}) + \beta\,(C_1+1)$.

Remark 6.2.4 *In view of Remark 6.2.2, the atomistic problem can be reinterpreted as a continuum problem posed on a finite element space of mesh size h. Therefore the convergence of order h for the "first derivative" of $u_c - u_\mu$ is sharp (see (6.36) and (6.37)). Using an argument* à la Aubin-Nitsche, *and assuming additionally that $W \in \mathcal{C}^3(\mathbb{R})$ and $f \in H^2(\Omega_M)$, one can improve (6.38) and (6.39) and show that, for h small enough,*

$$\|(\Pi_\mu u_c) - u_\mu\|_{L^\infty(X_\mu)} + \|u_c - \Pi_c u_\mu\|_{L^\infty(X_c)} \leq Ch^2$$

for some constant C that does not depend on h.

Proof of Theorem 6.2.1: Since u_μ is an increasing discrete function, the boundary conditions imply that $0 \leq u_\mu^i \leq a$. With the Euler-Lagrange equations (6.25), one can show that

$$\forall i \in \{1, \ldots, N-1\}, \quad \left|\frac{u_\mu^{i-1} - 2u_\mu^i + u_\mu^{i+1}}{h^2}\right| \leq \frac{1}{\alpha}\, \|f\|_{L^\infty(\Omega)}.$$

Collecting this bound with the bound on u_μ, one obtains

$$\forall i \in \{0, \dots, N-1\}, \quad \left| \frac{u_\mu^{i+1} - u_\mu^i}{h} \right| \leq C_1,$$

where C_1 is defined by (6.41). With the same arguments, we obtain a similar bound on u_c, so estimate (6.35) is proved. With (6.26) (respectively (6.27)), we see that (6.35) implies

$$\forall i \in \{0, \dots, N-1\}, \quad \left|\lambda_\mu - F_\mu^i\right| \leq \max_{[0,C_1]} |W'| \leq \beta\,(C_1 + 1), \tag{6.44}$$

and a similar bound on $\lambda_c - F_c^i$ (for $i \in \mathbb{N}_\mu$) and $\lambda_c - F_c(x)$ (for $x \in \Omega_M$) respectively. With (6.30), we infer from the estimate on $\lambda_c - F_c$ that

$$\forall i \in \{0, \dots, N-1\}, \quad \left|\lambda_c - F_\mu^i\right| \leq \beta\,(C_1 + 1) + h\kappa_f\,(\mathcal{N}_{cc} + 1)\,. \tag{6.45}$$

Let us define the functions

$$p_\mu(\lambda) = h \sum_{i=0}^{N-1} (W')^{-1}\left(\lambda - F_\mu^i\right),$$

$$p_c(\lambda) = \int_{\Omega_M} (W')^{-1}\left(\lambda - F_c(x)\right)\,dx + h \sum_{i \in \mathbb{N}_\mu} (W')^{-1}\left(\lambda - F_c^i\right).$$

In view of (6.26) and (6.27), we have $a = p_\mu(\lambda_\mu) = p_c(\lambda_c)$, so

$$p_c(\lambda_c) - p_\mu(\lambda_c) = p_\mu(\lambda_\mu) - p_\mu(\lambda_c). \tag{6.46}$$

Using this equality, we now estimate $|\lambda_c - \lambda_\mu|$.

With (6.19), we can see that $0 < \left((W')^{-1}\right)'(x) \leq 1/\alpha$ for any $x \in \mathbb{R}$. Furthermore, we see that, for any compact $K \subset \mathbb{R}$, there exists $\beta_K > 0$ such that, for any $x \in K$, $\left((W')^{-1}\right)'(x) > 1/\beta_K$. For any $\lambda \in \mathbb{R}$,

$$\begin{aligned} p_c(\lambda) - p_\mu(\lambda) \;=\; & \int_{\Omega_M} (W')^{-1}(\lambda - F_c(x))\,dx - h \sum_{j=1}^{J} \sum_{i=a_j}^{b_j - 1} (W')^{-1}\left(\lambda - F_c(ih)\right) \\ & + h \sum_{j=1}^{J} \sum_{i=a_j}^{b_j - 1} \left((W')^{-1}\left(\lambda - F_c(ih)\right) - (W')^{-1}(\lambda - F_\mu^i)\right) \\ & + h \sum_{i \in \mathbb{N}_\mu} \left((W')^{-1}(\lambda - F_c^i) - (W')^{-1}(\lambda - F_\mu^i)\right). \end{aligned} \tag{6.47}$$

We bound the first term : with (6.28), one can show that

$$\left| \int_{\Omega_M} (W')^{-1}(\lambda - F_c(x))\,dx - h \sum_{j=1}^{J} \sum_{i=a_j}^{b_j - 1} (W')^{-1}\left(\lambda - F_c(ih)\right) \right| \leq h\kappa_f \frac{L}{\alpha}.$$

For the third line of (6.47), we obtain, with (6.30),

$$\begin{aligned}\left|h\sum_{i\in\mathbb{N}_\mu}\left((W')^{-1}(\lambda-F_c^i)-(W')^{-1}(\lambda-F_\mu^i)\right)\right| &\leq \frac{h}{\alpha}\sum_{i\in\mathbb{N}_\mu}\left|F_c^i-F_\mu^i\right| \\ &\leq |\Omega_\mu|\frac{h\kappa_f}{\alpha}(1+\mathcal{N}_{cc}),\end{aligned}$$

and a similar estimate holds for the second line of (6.47). So we have, for any $\lambda\in\mathbb{R}$,

$$|p_c(\lambda)-p_\mu(\lambda)|\leq h\kappa_f\frac{L}{\alpha}(2+\mathcal{N}_{cc}). \tag{6.48}$$

On the other hand,

$$p_\mu(\lambda_\mu)-p_\mu(\lambda_c)=h\sum_{i=0}^{N-1}\left((W')^{-1}\left(\lambda_\mu-F_\mu^i\right)-(W')^{-1}\left(\lambda_c-F_\mu^i\right)\right).$$

The function $(W')^{-1}$ is increasing, so all the terms $(W')^{-1}\left(\lambda_\mu-F_\mu^i\right)-(W')^{-1}\left(\lambda_c-F_\mu^i\right)$ have the same sign. In view of (6.44) and (6.45), we see that $\lambda_\mu-F_\mu^i$ and $\lambda_c-F_\mu^i$ belong to the closed interval K of center 0 and of radius $\beta\,(C_1+1)+h_0\kappa_f(1+\mathcal{N}_{cc})$. So there exists β_K, which does not depend on h, such that

$$|p_\mu(\lambda_\mu)-p_\mu(\lambda_c)|\geq\frac{L}{\beta_K}\left|\lambda_\mu-\lambda_c\right|. \tag{6.49}$$

Collecting (6.46), (6.48) and (6.49), we finally obtain

$$|\lambda_\mu-\lambda_c|\leq C_0h\kappa_f, \tag{6.50}$$

where $C_0=\beta_K(2+\mathcal{N}_{cc})/\alpha$. We now prove (6.36). With (6.26) and (6.27), we have

$$\begin{aligned}\forall i\in\mathbb{N}_\mu,\quad \left|\frac{u_\mu^{i+1}-u_\mu^i}{h}-\frac{u_c^{i+1}-u_c^i}{h}\right| &\leq \frac{1}{\alpha}\left|\left(\lambda_\mu-F_\mu^i\right)-\left(\lambda_c-F_c^i\right)\right| \\ &\leq \frac{h\kappa_f}{\alpha}\left(C_0+\mathcal{N}_{cc}+1\right), \end{aligned}\tag{6.51}$$

where, in the last line, we have made use of (6.30) and (6.50). With (6.26) and (6.27), we also have, for all $i\notin\mathbb{N}_\mu$,

$$\begin{aligned}\left|\frac{u_\mu^{i+1}-u_\mu^i}{h}-\frac{(\Pi_\mu u_c)^{i+1}-(\Pi_\mu u_c)^i}{h}\right| &= \left|\frac{u_\mu^{i+1}-u_\mu^i}{h}-u_c'(ih+\gamma h)\right| \\ &\leq \left|\frac{u_\mu^{i+1}-u_\mu^i}{h}-u_c'(ih)\right|+h\|u_c''\|_{L^\infty(\Omega_M)} \\ &\leq \frac{1}{\alpha}\left|\left(\lambda_\mu-F_\mu^i\right)-(\lambda_c-F_c(ih))\right|+h\frac{\kappa_f}{\alpha} \\ &\leq \frac{h\kappa_f}{\alpha}\left(C_0+\mathcal{N}_{cc}+2\right), \end{aligned}\tag{6.52}$$

where, in the first line, $\gamma \in (0,1)$ and where we assumed at the second line that $ih \in \Omega_M$ (the case $ih+h \in \Omega_M$ can be treated likewise). Collecting (6.51) and (6.52), we obtain (6.36). The proof of (6.37) follows the same pattern. The estimate (6.38) is obtained by discrete summations of (6.36). We now prove (6.39). We already have, with (6.38),

$$\forall i \text{ s.t. } ih \in \Omega_\mu, \ \left|u_c^i - u_\mu^i\right| \leq C_3 h \kappa_f. \tag{6.53}$$

Let us now consider $x \in \Omega_M$, and let i such that $ih \leq x < ih + h$. We again assume that $ih \in \Omega_M$, and we have

$$|u_c(x) - (\Pi_c u_\mu)(x)| \leq \left|u_c(ih) - u_\mu^i\right| + \left|u_c(x) - u_c(ih) + u_\mu^i - (\Pi_c u_\mu)(x)\right|.$$

The first term is bounded by $C_3 h \kappa_f$ with (6.38). For the second term, we have

$$\left|u_c(x) - u_c(ih) + u_\mu^i - (\Pi_c u_\mu)(x)\right| \leq \int_{ih}^{x} |u_c'(y) - (\Pi_c u_\mu)'(y)|\, dy$$

and the estimate (6.37) allows one to bound the right hand side of the previous inequality. The energy estimate (6.40) is obtained from estimates (6.35), (6.36), (6.37), (6.38) and (6.39) by similar arguments. □

Remark 6.2.5 *It is also possible to prove (6.40) directly, with variational arguments, without resorting to estimates on* $u_c - u_\mu$.

6.3 The Lennard-Jones case

In this section, the interaction potential is the Lennard-Jones potential

$$W_{\rm LJ}(r) = \frac{1}{r^{12}} - \frac{2}{r^6}, \tag{6.54}$$

which attains its minimum at 1 : $W_{\rm LJ}(1) = \inf W_{\rm LJ} = -1$.

Let $W_{\rm LJ}^{**}$ be the convex envelop of $W_{\rm LJ}$ (which is also its quasiconvex envelop), and let us set

$$r_c = \left(\frac{13}{7}\right)^{1/6} \approx 1.11, \tag{6.55}$$

such that $W_{\rm LJ}$ is convex on the interval $(0, r_c)$ and concave on $(r_c, +\infty)$. We also define the functions

$$\psi \ : \ (-\infty, W_{\rm LJ}'(r_c)) \to (0, r_c) \quad \text{such that} \quad \psi \circ W_{\rm LJ}' = \mathrm{Id}, \tag{6.56}$$

$$\varphi \ : \ (0, W_{\rm LJ}'(r_c)) \to (r_c, +\infty) \quad \text{such that} \quad \varphi \circ W_{\rm LJ}' = \mathrm{Id}. \tag{6.57}$$

For $x \leq 0$, we will also make use of the notation $(W_{\rm LJ}')^{-1}(x) = \psi(x)$, as there is in such case no ambiguity. We denote by $H(t)$ the Heaviside function ($H(t) = 0$ if $t < 0$, $H(t) = 1$ otherwise).

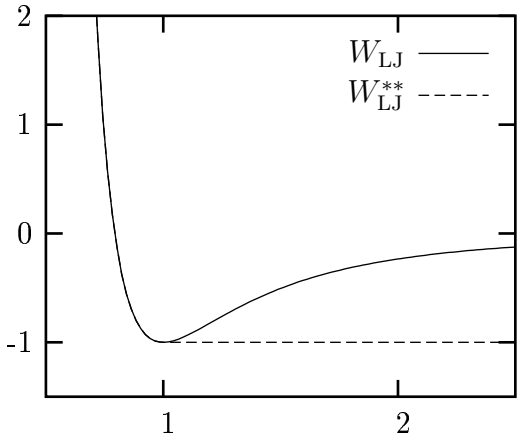

FIG. 6.2 – The Lennard-Jones potential (solid line) and its convex envelop (dashed line).

We use the Lennard-Jones potential as a prototype for a nonconvex interaction potential. One important feature allowing for the appearance of fracture is that $\lim_{r\to\infty} W_{\rm LJ}(r)/r = 0$.

In Section 6.3.1, we study the continuum mechanics problem (6.6), and show that a fracture (a Dirac mass in the derivative of the deformation) appears for any extensional load. The atomistic model (6.2) exhibits the same behavior (see Section 6.3.2). In Section 6.3.3, we study the coupled problem (6.11) where the energy is defined by (6.10). In Section 6.3.4, we study the modified coupled problem (6.13) with the energy (6.12). We then propose a way to build the atomistic-continuum partition (see Definition 6.3.1 below) such that the solution of the modified coupled problem approaches the solution of the atomistic problem. To build the partition, we will make use of a preliminary step : the determination of the solution of the continuum problem, a task which is assumed to be little demanding computationally in comparison with the resolution of a problem with a discrete component. The situation is thus different from the convex case in which we only use the given body forces f, and not the solution of the continuum problem, to define the partition (recall Remark 6.2.1).

Before entering the details, let us point out that the analysis below is highly dependent on the one-dimensional setting we chose here. However, the fact that the most natural way to couple the atomistic and the continuum models creates some difficulties (see Section 6.3.3) is more general, and occurs also in higher dimensions.

Remark 6.3.1 *On the other hand, the results of this section do not depend on the particular choice of the exponents that we have made in (6.54). One would obtain the same results with the potential* $W(r) = \dfrac{q}{r^p} - \dfrac{p}{r^q}$ *with* $p > q > 0$.

6.3.1 The continuum problem

We study in this subsection the continuum problem (6.6) for the Lennard-Jones potential, with $f \in L^1(\Omega)$. The natural variational space is

$$X_M(a) = \left\{ u \in W^{1,1}(\Omega),\ \frac{1}{u'} \in L^{12}(\Omega),\ u' > 0 \text{ a.e.},\ u(0) = 0,\ u(L) = a \right\}, \quad (6.58)$$

which will possibly need to be enlarged in order for the energy (6.5) to have a minimizer, as will be seen below. Let us set

$$\theta_M = \int_\Omega \left(W'_{\mathrm{LJ}}\right)^{-1} \left(\inf F_M - F_M(x)\right)\, dx, \quad (6.59)$$

$$v_1(x) = \left(W'_{\mathrm{LJ}}\right)^{-1} \left(\inf F_M - F_M(x)\right), \quad (6.60)$$

where F_M is defined by (6.15). We also recall (see [213]) the definition of the set

$$SBV(\Omega) = \left\{ u \in \mathcal{D}'(\Omega), u' = D_a u + \sum_{i\in\mathbb{N}} v_i \delta_{x_i}, D_a u \in L^1(\Omega), x_i \in \Omega, \sum_{i\in\mathbb{N}} |v_i| < +\infty \right\}.$$

We now give the main result of the present subsection :

Theorem 6.3.1 (Minimizers of the continuum LJ model) *If $\theta_M \geq a$, where θ_M is defined by (6.59), then the problem*

$$I_M^1 = \inf \left\{ E_M(u),\ u \in X_M(a) \right\}, \quad (6.61)$$

where $X_M(a)$ is defined by (6.58) and E_M is defined by (6.5), has a unique minimizer.

If $\theta_M < a$, then the problem (6.61) is not attained, but the problem

$$I_M^{BV} = \inf \left\{ E_M(u),\ u \in SBV(\Omega),\ \frac{1}{u'} \in L^{12}(\Omega),\ u' > 0,\ u(0) = 0,\ u(L) = a \right\} \quad (6.62)$$

has at least one minimizer. Moreover, $I_M^{BV} = I_M^1$ and the minimizers of the problem (6.62) are the functions

$$u(x) = \int_0^x v_1(t)\, dt + \sum_{i\in\mathbb{I}} \tilde{v}_i H(x - x_i),$$

where $v_1 \in L^1(\Omega)$ is defined by (6.60), $\mathbb{I}$ is any countable set, and $\tilde{v}_i$ and x_i are any real numbers such that

$$\sum_{i\in\mathbb{I}} \tilde{v}_i = a - \theta_M \quad \textit{and} \quad \forall i \in \mathbb{I},\ \tilde{v}_i > 0,\ x_i \in \arg\inf F_M.$$

Remark 6.3.2 *Let $u \in SBV(\Omega)$: its derivative reads $u' = D_a u + \sum_i \tilde{v}_i \delta_{x_i}$ with $D_a u \in L^1(\Omega)$. The notation $u' > 0$ means $D_a u > 0$ a.e. on Ω and $\tilde{v}_i > 0$. When $u' > 0$, we also use the convention $\dfrac{1}{u'} = \dfrac{1}{D_a u}$. The reason is that the inverse of a regularization of u' converges to the inverse of $D_a u$ in the sense of distribution. Since $W_{\rm LJ}(+\infty) = 0$, we will also use the convention $W_{\rm LJ}(u') = W_{\rm LJ}(D_a u)$.*

Proof of Theorem 6.3.1: The energy can also be written

$$E_M(u) = \int_\Omega \left(W_{\rm LJ}(u'(x)) + (F_M(x) - \inf F_M)\, u'(x)\right)\, dx + a\,(\inf F_M - F_M(L))\,. \tag{6.63}$$

We first treat the case $\theta_M \geq a$. Consider the following minimization problem :

$$\overline{I}_M = \inf\left\{\overline{E}_M(v), \quad v \in L^1(\Omega), \quad v > 0 \text{ a.e. }, \quad \frac{1}{v} \in L^{12}(\Omega), \quad \int_\Omega v = a\right\},$$

where

$$\overline{E}_M(v) = \int_\Omega W_{\rm LJ}(v) + (F_M - \inf F_M)\, v.$$

Clearly, $I^1_M = \overline{I}_M + a\,(\inf F_M - F_M(L))$ and u is a minimizer of I^1_M if and only if u' is a minimizer of $\overline{I}_M$. Let us define

$$v_0(x) = (W'_{\rm LJ})^{-1}\,(\inf F_M - F_M(x) - \lambda)\,, \tag{6.64}$$

where $\lambda \geq 0$ is chosen such that $\displaystyle\int_\Omega v_0 = a$. This is possible because

$$\int_\Omega (W'_{\rm LJ})^{-1}\,(\inf F_M - F_M(x))\, dx = \theta_M \geq a$$

and $\displaystyle\lim_{\lambda\to+\infty} \int_\Omega (W'_{\rm LJ})^{-1}\,(\inf F_M - F_M(x) - \lambda)\, dx = L\,(W'_{\rm LJ})^{-1}\,(-\infty) = 0$. We see that $v_0(x)$ is a continuous function that satisfies $v_0(x) \geq (W'_{\rm LJ})^{-1}\,(\inf F_M - \sup F_M - \lambda) > 0$. Thus it is a test function for $\overline{I}_M$. Consider now $v \in L^1(\Omega)$ satisfying $v > 0$, $1/v \in L^{12}(\Omega)$ and $\displaystyle\int_\Omega v = a$. The function $\underline{v} = \inf\{v, 1\}$ satisfies

$$\overline{E}_M(v) \geq \overline{E}_M(\underline{v}). \tag{6.65}$$

Define now for $\alpha \in [0,1]$ the function

$$p(\alpha) = \overline{E}_M\left((1-\alpha)v_0 + \alpha\underline{v}\right).$$

We have that p is convex, because both v_0 and $\underline{v}$ ly in $(0,1]$ where $W_{\rm LJ}$ is convex, and that $p'(0) = -\lambda \displaystyle\int_\Omega (\underline{v} - v_0) \geq 0$, thus p is nondecreasing. Therefore

$$p(0) = \overline{E}_M\,(v_0) \leq p(1) = \overline{E}_M\,(\underline{v}) \leq \overline{E}_M(v).$$

It follows that v_0 is a minimizer of $\overline{I}_M$. On the other hand, if for some v as above, $\overline{E}_M(v_0) = \overline{E}_M(v)$, then $p(\alpha)$ must be constant on $\alpha \in [0,1]$ and $\overline{E}_M(v) = \overline{E}_M(\underline{v})$ also. As the minimum of $W_{\rm LJ}$ is attained at 1, the latter implies that $v = \underline{v}$, while p constant implies that $p'' = 0$, thus $v_0 = \underline{v}$ (because $W_{\rm LJ}$ is strictly convex on $(0,1]$). This proves that v_0 is the unique minimizer of $\overline{I}_M$, and that

$$u_0(x) = \int_0^x v_0(t)\,dt$$

is the unique minimizer of (6.61).

We now turn to the case $\theta_M < a$ and show that (6.62) has at least a minimizer. Consider $u \in SBV(\Omega)$ such that $1/u' \in L^{12}(\Omega)$, $u' > 0$, $u(0) = 0$ and $u(L) = a$. We use the notation $u' = D_a u + \sum_{i\in\mathbb{I}} \tilde{v}_i \delta_{x_i}$ with $D_a u \in L^1(\Omega)$, $D_a u > 0$ and $\tilde{v}_i > 0$. Recall that we use the convention that $W_{\rm LJ}(u') = W_{\rm LJ}(D_a u)$. Thus

$$E_M(u) = \int_\Omega P_x(D_a u(x))\,dx + \sum_{i\in\mathbb{I}} \tilde{v}_i \left(F_M(x_i) - \inf F_M\right) + a\left(\inf F_M - F_M(L)\right), \tag{6.66}$$

where $P_x(t)$ is the function

$$P_x(t) = W_{\rm LJ}(t) + \left(F_M(x) - \inf F_M\right) t. \tag{6.67}$$

On $t \in (0, +\infty)$, this function has a unique minimizer which is the function $v_1(x)$ defined by (6.60). So $P_x(D_a u(x)) \geq P_x(v_1(x))$ on Ω and we infer from (6.66) that

$$E_M(u) \geq \int_\Omega P_x(v_1(x))\,dx + \sum_{i\in\mathbb{I}} \tilde{v}_i \left(F_M(x_i) - \inf F_M\right) + a\left(\inf F_M - F_M(L)\right). \tag{6.68}$$

Consider now a point $x_0 \in \arg\inf F_M$, and define

$$u_0(x) = \int_0^x v_1(t)\,dt + (a - \theta_M) H(x - x_0). \tag{6.69}$$

We have $D_a u_0 = v_1 \in L^1(\Omega)$. Since v_1 is a continuous function that satisfies $v_1(x) \geq \inf v_1 > 0$, we see that $1/v_1 \in L^{12}(\Omega)$, thus u_0 is an admissible function of (6.62) and, making use of (6.66) with $u \equiv u_0$, we obtain

$$E_M(u_0) = \int_\Omega P_x(v_1(x))\,dx + a\left(\inf F_M - F_M(L)\right). \tag{6.70}$$

Collecting (6.68) and (6.70), we obtain

$$E_M(u) \geq E_M(u_0) + \sum_{i\in\mathbb{I}} \tilde{v}_i \left(F_M(x_i) - \inf F_M\right) \tag{6.71}$$

for all test functions u of problem (6.62) such that $u' = D_a u + \sum_{i\in\mathbb{I}} \tilde{v}_i \delta_{x_i}$. Since $\tilde{v}_i > 0$, we have $E_M(u) \geq E_M(u_0)$ and the function u_0 is a minimizer of problem (6.62).

We now look for all the minimizers of problem (6.62). Let u be any test function. From (6.71), we see that $E_M(u) > E_M(u_0)$ if there exists $i \in \mathbb{I}$ such that $x_i \notin \arg\inf F_M$ and $\tilde{v}_i > 0$. In addition, in view of (6.66) and (6.70), we infer from $E_M(u) = E_M(u_0)$ that $\displaystyle\int_\Omega P_x(D_a u(x))\,dx = \int_\Omega P_x(v_1(x))\,dx$. As $v_1(x)$ is the unique minimizer of $t \in (0,+\infty) \mapsto P_x(t)$, we obtain that $v_1 = D_a u$ a.e. on Ω. Hence, any minimizer of (6.62) may be written

$$u(x) = \int_0^x (W'_{\mathrm{LJ}})^{-1} \left(\inf F_M - F_M(t)\right)\, dt + \sum_{i\in\mathbb{I}} \tilde{v}_i H(x - x_i),$$

where $\mathbb{I}$ is any countable set, $x_i \in \arg\inf F_M$ and $\tilde{v}_i > 0$ for all $i \in \mathbb{I}$, and $\displaystyle\sum_{i\in\mathbb{I}} \tilde{v}_i = a - \theta_M$.

Let us now show that $I^{BV}_M = I^1_M$. We already have $I^{BV}_M \leq I^1_M$ and we have shown that $I^{BV}_M = E_M(u_0)$, where u_0 is defined by (6.69). Let u_0^ε be a regularization of u_0 : then u_0^ε is a test function for (6.61) and

$$I^{BV}_M = E_M(u_0) = \lim_{\varepsilon\to 0} E_M(u_0^\varepsilon) \geq I^1_M.$$

Thus $I^{BV}_M = I^1_M$. To prove that problem (6.61) is not attained, we argue by contradiction. Let us assume that problem (6.61) is attained and let $\underline{u}$ be a minimizer. With (6.63), we see that

$$E_M(\underline{u}) = \int_\Omega P_x(\underline{u}'(x))\,dx + a\left(\inf F_M - F_M(L)\right), \tag{6.72}$$

where $P_x(t)$ is defined by (6.67). We have $E_M(\underline{u}) = I^1_M = I^{BV}_M = E_M(u_0)$, thus, in view of (6.70) and (6.72), we obtain $\displaystyle\int_\Omega P_x(v_1(x))\,dx = \int_\Omega P_x(\underline{u}'(x))\,dx$, which implies that $v_1(x) = \underline{u}'(x)$ a.e. on Ω. But this is impossible since $\displaystyle\int_\Omega v_1 = \theta_M < a = \int_\Omega \underline{u}'$. □

6.3.2 The atomistic problem

In this subsection, we study the atomistic problem (6.2), where the energy E_μ is given by (6.1) with $W \equiv W_{\mathrm{LJ}}$. In particular, we show that, for some particular choices of boundary conditions, a fracture appears. This means that the distance between a pair (and actually only one) of consecutive atoms is outstreched. The ideas of the proof are first explained on the simple case $f \equiv 0$. Next we deal with the general case, which is more technical. The main result of this subsection is Theorem 6.3.3, which is a generalization of some results given in [210].

6.3.2.1 The case of no body force

To study problem (6.2), we need the following lemma.

Lemma 6.3.1 *If $a > L$, there exists h_0 such that, for all $h \leq h_0$, there exists a unique pair $(s(h), s_f(h)) \in \mathbb{R}^2$ such that*

$$1 \leq s(h) \leq 1+h,\ W'_{\rm LJ}(s(h)) = W'_{\rm LJ}(s_f(h)) \text{ and } (L-h)s(h) + hs_f(h) = a. \quad (6.73)$$

In addition, we have the estimates

$$s_f(h) \sim_{h\to 0} \frac{a-L}{h} \quad \text{and} \quad s(h) - 1 \sim_{h\to 0} C_0 h^7, \quad (6.74)$$

for some C_0 that does not depend on h.

We skip the proof of Lemma 6.3.1, since it proceeds from an elementary study of the variations of $g(s) = W'_{\rm LJ}(s) - W'_{\rm LJ}\left(\frac{a-(L-h)s}{h}\right)$.

Theorem 6.3.2 *We suppose that there are no body force : $f \equiv 0$.*
If $a \leq L$, then (6.2) has a unique minimizer, defined by $u^i_\mu = ih\ a/L$, $i = 0, \ldots, N$.
If $a > L$, then there exists h_0 such that, for all $h \leq h_0$, the minimizers of (6.2) are exactly the N discrete functions defined for $i_\mu \in \{0, \ldots, N-1\}$ by

$$\frac{u^{i_\mu+1}_\mu - u^{i_\mu}_\mu}{h} = s_f(h) \quad \text{and} \quad \forall i \neq i_\mu,\ \frac{u^{i+1}_\mu - u^i_\mu}{h} = s(h), \quad (6.75)$$

where $s(h)$ and $s_f(h)$ are defined by Lemma 6.3.1.

Remark 6.3.3 *In the case $a > L$, the continuum energy is $I_M = LW_{\rm LJ}(1)$ (see Theorem 6.3.1), whereas it is $I_\mu = LW_{\rm LJ}(1) + h + O(h^7)$ with the atomistic model. In the next subsection, we will see a consequence of the inequality $I_\mu > I_M$.*

Proof of Theorem 6.3.2: As we impose the increasing condition $u^{i+1} > u^i$, any minimizing sequence u^i_n satisfies $0 \leq u^i_n \leq a$ for all i, hence is compact in $\mathbb{R}^{N+1}$. Thus, one can extract a subsequence that converges to a configuration (denoted by u_μ) which is a minimizer of the energy. This configuration satisfies the constraint $u^{i+1}_\mu > u^i_\mu$ for all i. Otherwise, there exists i such that $\lim_{n\to+\infty} u^{i+1}_n - u^i_n = 0$, which implies, as $W_{\rm LJ}(0) = +\infty$, that the infimum (6.2) is $+\infty$, which is a contradiction. Hence, (6.2) has at least one minimizer.

Choosing $u^i = ih\ a/L$ as a test function, we show that $L\,W_{\rm LJ}(a/L) \geq I_\mu$. Bounding $W_{\rm LJ}$ from below by $W^{**}_{\rm LJ}$, and utilizing the convexity of $W^{**}_{\rm LJ}$, we show that $I_\mu \geq W^{**}_{\rm LJ}(a/L)$. Thus

$$L\,W_{\rm LJ}\left(\frac{a}{L}\right) \geq I_\mu \geq L\,W^{**}_{\rm LJ}\left(\frac{a}{L}\right). \quad (6.76)$$

We first treat the case $a \leq L$, where $W_{\rm LJ}(a/L) = W_{\rm LJ}^{**}(a/L)$, thus $I_\mu = L\,W_{\rm LJ}(a/L)$. The configuration $u_\mu^i = ih\ a/L$ is thus a minimizer. Let us now consider a configuration u for which the quantities $\dfrac{u^{i+1}-u^i}{h}, i = 0, \ldots, N-1$, are not all equal to a/L. Then

$$\begin{aligned} E_\mu(u) = h\sum_{i=0}^{N-1} W_{\rm LJ}\left(\frac{u^{i+1}-u^i}{h}\right) &\geq h\sum_{i=0}^{N-1} W_{\rm LJ}^{**}\left(\frac{u^{i+1}-u^i}{h}\right) \\ &> L\,W_{\rm LJ}^{**}\left(\frac{h}{L}\sum_{i=0}^{N-1}\frac{u^{i+1}-u^i}{h}\right), \end{aligned}$$

where we have made use of the strict case of the convexity inequality. Thus we have $E_\mu(u) > L\,W_{\rm LJ}^{**}(a/L) = I_\mu$, and u is not a minimizer. This shows the uniqueness claimed in the theorem.

We now turn to the case $a > L$. Let us consider a minimizer u_μ of problem (6.2). As $W_{\rm LJ}(0) = +\infty$, the constraint $u^{i+1} > u^i$ is not active, so the Euler-Lagrange equation of (6.2) reads as (6.25) :

$$\forall i \in \{1, \ldots, N-1\}, \quad W_{\rm LJ}'\left(\frac{u_\mu^i - u_\mu^{i-1}}{h}\right) = W_{\rm LJ}'\left(\frac{u_\mu^{i+1}-u_\mu^i}{h}\right).$$

If one slope $\dfrac{u_\mu^{i+1}-u_\mu^i}{h}$ is equal to or smaller than 1, then the same holds for the other slopes, and this is in contradiction with the fact that $a > L$. So all the slopes are strictly larger than 1. In addition, due to the variations of $W_{\rm LJ}$, there are two values $s^*(h)$ and $s_f^*(h)$, with $1 < s^*(h) \leq r_c \leq s_f^*(h)$, $W_{\rm LJ}'(s^*(h)) = W_{\rm LJ}'(s_f^*(h))$, such that each of the slopes is either $s^*(h)$ or $s_f^*(h)$. Let

$$k = \text{Card}\left\{ i \in \{0, \ldots, N-1\} \text{ such that } \frac{u_\mu^{i+1}-u_\mu^i}{h} = s_f^*(h) \right\}.$$

We have $(L-kh)s^*(h) + khs_f^*(h) = a$ and $I_\mu = (L-kh)W_{\rm LJ}(s^*(h)) + khW_{\rm LJ}(s_f^*(h))$.

We claim that $k = 1$. For this purpose, we consider the discrete function u_b defined by $u_b^i = ih$ for $i = 0, \ldots, N-1$, and $u_b^N = a$. Its energy is

$$E_\mu(u_b) = (L-h)\,W_{\rm LJ}(1) + hW_{\rm LJ}\left(\frac{a-L}{h}+1\right),$$

and satisfies $\lim\limits_{h\to 0} E_\mu(u_b) = L\,W_{\rm LJ}(1)$. We first show that $k \neq 0$. Otherwise, all slopes $\left(u_\mu^{i+1}-u_\mu^i\right)/h$ are equal to the same value $s^*(h)$, which thus needs to be a/L. As u_μ is a minimizer, $I_\mu = L\,W_{\rm LJ}(a/L)$. However, we also have $I_\mu \leq E_\mu(u_b)$. Letting h go to zero, we obtain $W_{\rm LJ}(a/L) \leq W_{\rm LJ}(1)$, which is a contradiction as $a > L$.

We now show that $k \leq 1$. As $I_\mu \leq E_\mu(u_b)$, one obtains

$$L\left(W_{\rm LJ}(s^*(h)) - W_{\rm LJ}(1)\right) + kh\left(W_{\rm LJ}(s_f^*(h)) - W_{\rm LJ}(s^*(h))\right) \leq -hW_{\rm LJ}(1). \quad (6.77)$$

The left hand side is a sum of two nonnegative terms, for $1 < s^*(h) \leq s_f^*(h)$ and $W_{\rm LJ}$ is an increasing function on $[1, +\infty)$. So we have $\lim_{h\to 0} W_{\rm LJ}(s^*(h)) = W_{\rm LJ}(1)$, hence $\lim_{h\to 0} s^*(h) = 1$, which in turn implies that $\lim_{h\to 0} s_f^*(h) = +\infty$. Inserting this information in (6.77), we obtain $k \leq 3/2$, thus $k = 1$ for h small enough.

We finally identify $(s^*(h), s_f^*(h))$ for h small enough. We have $(L-h)s^*(h) + hs_f^*(h) = a$. So, using $\lim_{h\to 0} s^*(h) = 1$, we see that $s_f^*(h) \sim_{h\to 0} (a-L)/h$. So $W'_{\rm LJ}(s^*(h)) = W'_{\rm LJ}(s_f^*(h)) \sim Ch^7$, and $s^*(h) = 1 + O(h^7)$. As a consequence, $s^*(h) \in [1, 1+h]$. By uniqueness of the pair $(s(h), s_f(h))$ satisfying (6.73) (see Lemma 6.3.1 above), one has $s^*(h) = s(h)$ and $s_f^*(h) = s_f(h)$. So u_μ satisfies (6.75) for some integer i_μ. □

6.3.2.2 The general case

We now assume that the body forces f are in $\mathcal{C}^0\left(\overline{\Omega}\right)$, and are not necessarily zero. This regularity assumption is needed for E_μ to be well-defined. For any configuration $u \in X_\mu(a)$, we define a partition of the set of indices $\{0, \ldots, N-1\}$ in 3 different subsets :

$$\begin{aligned}
G_1(u) &= \left\{ i \in [0, N-1];\ 0 < \frac{u^{i+1} - u^i}{h} \leq 1 \right\}, \\
G_2(u) &= \left\{ i \in [0, N-1];\ 1 < \frac{u^{i+1} - u^i}{h} < r_c \right\}, \\
G_3(u) &= \left\{ i \in [0, N-1];\ r_c \leq \frac{u^{i+1} - u^i}{h} \right\}.
\end{aligned}$$

Let us set

$$\underline{F_\mu} = \inf_{i=0,\ldots,N-1} F_\mu^i, \quad (6.78)$$

$$\theta_\mu = h \sum_{i=0}^{N-1} (W'_{\rm LJ})^{-1}(\underline{F_\mu} - F_\mu^i), \quad (6.79)$$

where F_μ is defined by (6.16). In view of (6.29), we have $\lim_{h\to 0} \theta_\mu = \theta_M$, where θ_M, given by (6.59), is the threshold for appearance of a fracture in the continuum model (see Theorem 6.3.1). We note that, if $f \equiv 0$, $\theta_\mu = \theta_M = L$. To study problem (6.2), we need the following lemma :

Lemma 6.3.2 (Estimate on the Lagrange multiplier for the atomistic problem (6.2)) *Let i_μ be an index such that $F_\mu^{i_\mu} = \underline{F_\mu}$, and let us assume that $a > \theta_M$.*

There exists h_0 such that, for all $h \le h_0$, there exists a unique λ_μ such that

$$0 < \lambda_\mu \le h \quad \text{and} \quad h \sum_{i \in \{0,\dots,N-1\}, i \neq i_\mu} \psi\left(\underline{F_\mu} - F_\mu^i + \lambda_\mu\right) + h\varphi(\lambda_\mu) = a, \tag{6.80}$$

where ψ and φ are defined by (6.56) and (6.57).

In addition, λ_μ does not depend on i_μ and satisfies

$$\lambda_\mu \sim_{h\to 0} C_0 h^7, \tag{6.81}$$

for some C_0 that does not depend on h.

The proof of Lemma 6.3.2 is based on the study of variations, on $[0, h]$, of the function $g_\mu(\lambda) = \lambda - W'_{\rm LJ}\left(\dfrac{a - h\sum_{i\neq i_\mu}\psi\left(\underline{F_\mu} - F_\mu^i + \lambda\right)}{h}\right)$. We now turn to the main result of this section.

Theorem 6.3.3 (Minimizers of the atomistic problem (6.2)) *We assume that $f \in \mathcal{C}^0\left(\overline{\Omega}\right)$. Let θ_μ be defined by (6.79).*

If $a \le \theta_\mu$, then (6.2) has a unique minimizer.

If $a > \theta_M$, there exists h_0 such that, for all $h \le h_0$, the minimizers of (6.2) are exactly the discrete functions defined for $i_\mu \in \{0, \dots, N-1\}$ such that $F_\mu^{i_\mu} = \underline{F_\mu}$ by $u_\mu^0 = 0$ and

$$\frac{u_\mu^{i_\mu+1} - u_\mu^{i_\mu}}{h} = \varphi(\lambda_\mu) \quad \text{and} \quad \forall i \neq i_\mu,\ \frac{u_\mu^{i+1} - u_\mu^i}{h} = \psi(\underline{F_\mu} - F_\mu^i + \lambda_\mu), \tag{6.82}$$

where λ_μ is defined by Lemma 6.3.2 and ψ and φ are defined by (6.56) and (6.57). In addition, $G_3(u_\mu) = \{i_\mu\}$ and

$$\frac{u_\mu^{i_\mu+1} - u_\mu^{i_\mu}}{h} \sim_{h\to 0} \frac{a - \theta_M}{h}. \tag{6.83}$$

Remark 6.3.4 *If $\theta_\mu < \theta_M$, Theorem 6.3.3 does not apply to $a \in (\theta_\mu, \theta_M]$. Note however that $\lim_{h\to 0} \theta_\mu = \theta_M$, thus all boundary conditions a are asymptotically covered.*

Proof of Theorem 6.3.3: If $f \equiv 0$, then Theorem 6.3.3 is identical to Theorem 6.3.2. We now concentrate on the case $f \neq 0$. As in the proof of Theorem 6.3.1, we first reformulate the energy, here in term of the slopes

$$v^i = \frac{u^{i+1} - u^i}{h}, \quad 0 \le i \le N-1. \tag{6.84}$$

The energy (6.1) can be written

$$E_\mu(u) = a\left(\underline{F_\mu} - F_\mu^N\right) + h\sum_{i=0}^{N-1}\left(W_{\rm LJ}\left(\frac{u^{i+1}-u^i}{h}\right) + \left(F_\mu^i - \underline{F_\mu}\right)\frac{u^{i+1}-u^i}{h}\right).$$

Then we consider

$$\overline{I}_\mu = \inf\left\{\overline{E}_\mu(v);\ v \in Y_\mu(a)\right\}, \tag{6.85}$$

where

$$Y_\mu(a) = \left\{v = (v^0,\dots,v^{N-1}) \in \mathbb{R}^N,\ h\sum_{i=0}^{N-1} v^i = a,\ v^i > 0\right\},$$

$$\overline{E}_\mu(v) = h\sum_{i=0}^{N-1}\left(W_{\rm LJ}(v^i) + \left(F_\mu^i - \underline{F_\mu}\right)v^i\right).$$

Clearly, $I_\mu = \overline{I}_\mu + a\left(\underline{F_\mu} - F_\mu^N\right)$. If u is a minimizer of (6.2), then the discrete function v defined from u by (6.84) is a minimizer of (6.85). On the other hand, if v is a minimizer of (6.85), then u defined by $u^i = h\sum_{j=0}^{i-1} v^j$ for $i \geq 1$ and $u^0 = 0$ is a minimizer of (6.2).

Let v_n be a minimizing sequence of $\overline{I}_\mu$. Since $0 < v_n^i \leq a/h$ for all $0 \leq i \leq N-1$ and all n, one can extract a subsequence that converges to the configuration v_μ. Again, since $W_{\rm LJ}(0) = +\infty$ and the infimum (6.85) is not equal to $+\infty$, we have $v_\mu^i > 0$ for all $0 \leq i \leq N-1$ and v_μ is a minimizer of (6.85). Hence (6.85), and therefore (6.2), have at least one minimizer, we denote by v_μ and u_μ two corresponding minimizers of (6.85) and (6.2).

In the case $a \leq \theta_\mu$, the proof of uniqueness follows the same lines as the proof of Theorem 6.3.1, in the case $a \leq \theta_M$.

We now turn to the case $a > \theta_M$. As $\lim_{h\to 0}\theta_\mu = \theta_M$, we choose h small enough such that $a > \theta_\mu$. We introduce the function

$$P_i(t) = W_{\rm LJ}(t) + (F_\mu^i - \underline{F_\mu})t,$$

which attains its minimum with respect to $t \in [0,+\infty)$ at $t = v_1^i$ defined by

$$v_1^i = (W'_{\rm LJ})^{-1}(\underline{F_\mu} - F_\mu^i), \quad i = 0,\dots,N-1. \tag{6.86}$$

We notice that

$$\forall v \in Y_\mu(a), \quad \overline{E}_\mu(v) = h\sum_{i=0}^{N-1} P_i(v^i) \geq \overline{E}_\mu(v_1). \tag{6.87}$$

As $W_{\rm LJ}(0) = +\infty$, the constraint $v^i > 0$ is not active in (6.85), so the Euler-Lagrange equation of (6.85) reads

$$W'_{\rm LJ}(v_\mu^i) + F_\mu^i - \underline{F_\mu} = \lambda_\mu^*, \tag{6.88}$$

where λ_μ^* is the Lagrange multiplier associated to the constraint $h\sum_{i=0}^{N-1} v_\mu^i = a$. Since $a > \theta_\mu$, one can show that $\lambda_\mu^* > 0$ (otherwise, (6.88) leads to $v_\mu^i \leq v_1^i$, and summing these inequalities leads to $a \leq \theta_\mu$). It holds that

$$\forall i \in G_1(u_\mu) \cup G_2(u_\mu), \quad v_\mu^i = \psi(\underline{F_\mu} - F_\mu^i + \lambda_\mu^*). \tag{6.89}$$

Step 1 : $\liminf\ h\,\mathrm{Card}(G_1(u_\mu)) > 0$:

In view of (6.87), we have

$$\overline{E}_\mu(v_1) \leq \overline{I}_\mu. \tag{6.90}$$

Let i_0 be an index such that $F_\mu^{i_0} = \underline{F_\mu}$, and let us consider the function v_b defined by $v_b^i = v_1^i > 0$ for $i \neq i_0$ and $v_b^{i_0}$ such that $h\sum_{i=0}^{N-1} v_b^i = a$. By construction, $v_1^{i_0} = 1$ and $hv_b^{i_0} = a - \theta_\mu + h$, so

$$v_b^{i_0} \sim_{h\to 0} \frac{a-\theta_M}{h}. \tag{6.91}$$

Hence, v_b is a test function for (6.85) and we have $\overline{I}_\mu \leq \overline{E}_\mu(v_b)$. Collecting this inequality with (6.90), we have

$$\overline{E}_\mu(v_1) \ \leq\ \overline{I}_\mu \ \leq\ \overline{E}_\mu(v_b). \tag{6.92}$$

Since $F_\mu^{i_0} = \underline{F_\mu}$, we have $\overline{E}_\mu(v_b) - \overline{E}_\mu(v_1) = h\left(W_{\mathrm{LJ}}(v_b^{i_0}) - W_{\mathrm{LJ}}(v_1^{i_0})\right)$. We know that $v_1^{i_0} = 1$. With (6.91), we obtain

$$0 \leq \overline{E}_\mu(v_b) - \overline{E}_\mu(v_1) \leq -hW_{\mathrm{LJ}}(1). \tag{6.93}$$

With (6.92) and (6.93), we have

$$\lim_{h\to 0} \overline{I}_\mu = \lim_{h\to 0} \overline{E}_\mu(v_1) = \int_\Omega W_{\mathrm{LJ}}(v_1(x)) + (F_M(x) - \inf F_M)v_1(x)\,dx, \tag{6.94}$$

where $v_1(x)$ is defined by (6.60). We now establish a lower bound for $\overline{I}_\mu$. Since $\overline{I}_\mu = h\sum_{i=0}^{N-1}\left(W_{\mathrm{LJ}}(v_\mu^i) + (F_\mu^i - \underline{F_\mu})v_\mu^i\right)$, we have

$$\overline{I}_\mu \geq h\sum_{i=0}^{N-1}\left(W_{\mathrm{LJ}}(1) + (F_\mu^i - \underline{F_\mu})\right) + h\sum_{i\in G_1(u_\mu)} (F_\mu^i - \underline{F_\mu})(v_\mu^i - 1).$$

The previous inequality implies

$$\overline{I}_\mu - h\sum_{i=0}^{N-1}\left(W_{\mathrm{LJ}}(1) + (F_\mu^i - \underline{F_\mu})\right) \geq -h\,\mathrm{Card}(G_1(u_\mu))\left(\sup_{0\leq i\leq N-1} F_\mu^i - \underline{F_\mu}\right).$$

Letting h goes to zero in the above inequality, we obtain, in view of (6.94),

$$\int_\Omega W_{\rm LJ}(v_1) + (F_M - \inf F_M)v_1 - \int_\Omega W_{\rm LJ}(1) + (F_M - \inf F_M) \\ \geq -\liminf_{h\to 0} \left(h\,{\rm Card}(G_1(u_\mu)) \left(\sup_{0\leq i\leq N-1} F_\mu^i - \underline{F_\mu} \right) \right).$$

Since $f \neq 0$, we see that $v_1(x) \neq 1$ somewhere in Ω. As $v_1(x)$ minimizes the function $W_{\rm LJ}(t) + (F_M(x) - \inf F_M)t$ on $[0, +\infty)$, we obtain $\liminf h\,{\rm Card}(G_1(u_\mu)) > 0$, and there exists $L_1 > 0$ such that, for h small enough, $h\,{\rm Card}(G_1(u_\mu)) \geq L_1$.

Step 2 : A first estimate of the Lagrange multiplier λ_μ^* and of v_μ :

In view of (6.92) and (6.93), we have

$$0 \leq \overline{I}_\mu - \overline{E}_\mu(v_1) \leq -hW_{\rm LJ}(1). \tag{6.95}$$

Since $\overline{E}_\mu(v) = h\sum_{i=0}^{N-1} P_i(v^i)$ for any $v \in Y_\mu(a)$ (see (6.87)), we have

$$h \sum_{i\in G_1(u_\mu)} \big(P_i(v_\mu^i) - P_i(v_1^i)\big) + h \sum_{i\notin G_1(u_\mu)} \big(P_i(v_\mu^i) - P_i(v_1^i)\big) \leq -hW_{\rm LJ}(1). \tag{6.96}$$

Since $t \mapsto P_i(t)$ attains its minimum at v_1^i, we see that the left hand side of the above inequality is a sum of two nonnegative terms. We know that $v_1^i \leq 1$. Using the convexity of P_i on $(0, 1]$, one obtains

$$\forall i \in G_1(u_\mu), \quad P_i(v_\mu^i) - P_i(v_1^i) \geq C(v_\mu^i - v_1^i)^2, \tag{6.97}$$

where C stands (here and below) for a generic constant which does not depend on h. With (6.86) and (6.89), we obtain $v_\mu^i - v_1^i \geq C\lambda_\mu^*$ for some constant C. Inserting this information in (6.96) and (6.97), one obtains $(\lambda_\mu^*)^2\, h\ {\rm Card}(G_1(u_\mu)) \leq O(h)$. As $h\,{\rm Card}(G_1(u_\mu)) \geq L_1 > 0$, we have

$$0 < \lambda_\mu^* \leq C\sqrt{h}. \tag{6.98}$$

Collecting (6.86), (6.89) and (6.98), we infer that there exists a constant C independent of h such that, for all i in $G_1(u_\mu) \cup G_2(u_\mu)$, we have $|v_\mu^i - v_1^i| \leq C\lambda_\mu^*$, so

$$\begin{aligned} &\forall i \in G_1(u_\mu) \cup G_2(u_\mu), \quad |v_\mu^i - v_1^i| \leq C\sqrt{h}, \\ &\forall i \in G_3(u_\mu), \qquad\quad v_\mu^i \geq \frac{C}{h^{1/14}}. \end{aligned} \tag{6.99}$$

Step 3 : ${\rm Card}\, G_3(u_\mu) = 1$:

For any $i \in G_1(u_\mu) \cup G_2(u_\mu)$, we see that v_1^i and v_μ^i belong to the interval $(0, 1 + C\sqrt{h}]$, on which $W_{\rm LJ}$ and P_i are convex. So inequality (6.97) is valid for all $i \in G_1(u_\mu) \cup G_2(u_\mu)$, i.e. for all $i \notin G_3(u_\mu)$. Thus, (6.96) implies on the one hand

$$C\,h \sum_{i\notin G_3(u_\mu)} (v_\mu^i - v_1^i)^2 \leq h \sum_{i\notin G_3(u_\mu)} \big(P_i(v_\mu^i) - P_i(v_1^i)\big) \leq -hW_{\rm LJ}(1). \tag{6.100}$$

On the other hand, for any $i \in G_3(u_\mu)$, $P_i(v_1^i) \leq P_i(1) = W_{\rm LJ}(1) + F_\mu^i - \underline{F_\mu}$ and, with (6.99),

$$P_i(v_\mu^i) \geq W_{\rm LJ}(v_\mu^i) + F_\mu^i - \underline{F_\mu} \geq W_{\rm LJ}\left(\frac{C}{h^{1/14}}\right) + F_\mu^i - \underline{F_\mu}.$$

Inserting this information in (6.96), we obtain

$$\mathrm{Card}\, G_3(u_\mu)\left(W_{\rm LJ}\left(\frac{C}{h^{1/14}}\right) - W_{\rm LJ}(1)\right) \leq \sum_{i\in G_3(u_\mu)} \left(P_i(v_\mu^i) - P_i(v_1^i)\right) \leq -W_{\rm LJ}(1).$$

So, for h small enough, $\mathrm{Card}\, G_3(u_\mu) \leq 1$. If $G_3(u_\mu) = \emptyset$, then, with (6.100), we obtain

$$0 \;<\; a - \theta_\mu \;=\; h\sum_{i=0}^{N-1} v_\mu^i - v_1^i \;\leq\; \sqrt{L}\,\sqrt{h\sum_{i=0}^{N-1}(v_\mu^i - v_1^i)^2} \;\leq\; O(\sqrt{h}),$$

and we come to a contradiction with $a > \theta_M = \lim_{h\to 0}\theta_\mu$. So $\mathrm{Card}\, G_3(u_\mu) = 1$, and we denote by i_μ its unique index : $G_3(u_\mu) = \{i_\mu\}$.

If $F_\mu^{i_\mu} > \underline{F_\mu}$, let i_0 be an index such that $F_\mu^{i_0} = \underline{F_\mu}$. By exchanging $v_\mu^{i_\mu}$ and $v_\mu^{i_0}$, one can lower the energy of v_μ. This is in contradiction with the fact that v_μ is minimizer. So $F_\mu^{i_\mu} = \underline{F_\mu}$.

Step 4 : Identification of the Lagrange multiplier :

We have

$$a - \theta_M = a - \theta_\mu + o(1) = o(1) + h\sum_{i\notin G_3(u_\mu)} \left(v_\mu^i - v_1^i\right) + h\left(v_\mu^{i_\mu} - v_1^{i_\mu}\right).$$

With (6.99), we see that $h\sum_{i\notin G_3(u_\mu)} \left(v_\mu^i - v_1^i\right) = o(1)$. With (6.86), we have $v_1^{i_\mu} = 1$, so $a - \theta_M = o(1) + hv_\mu^{i_\mu}$. Hence,

$$v_\mu^{i_\mu} \sim_{h\to 0} \frac{a - \theta_M}{h},$$

which implies (6.83). As a consequence, $W'_{\rm LJ}\left(v_\mu^{i_\mu}\right) \sim Ch^7$. From (6.88), we obtain $\lambda_\mu^* = W'_{\rm LJ}\left(v_\mu^{i_\mu}\right)$. So, for h small enough, we have $\lambda_\mu^* \in (0, h]$ and, since $h\sum v_\mu^i = a$, the Lagrange multiplier λ_μ^* satisfies (6.80). Since λ_μ, defined by Lemma 6.3.2, is the unique solution of (6.80), we have $\lambda_\mu^* = \lambda_\mu$. Collecting this equality with (6.88), we obtain (6.82). □

6.3.3 The natural coupled approach

We study in this subsection the problem (6.11). We will see that this coupled problem, though natural, has a flaw. In order to illustrate this fact, we restrict ourselves to the case $f \equiv 0$ (see Remark 6.3.7 below for the case $f \neq 0$). We also assume that

$$a > L,$$

because otherwise the minimizers of the continuum and atomistic problems are equal and no singularity appears, and it is therefore not interesting to use a coupled method. We assume that $\Omega = (0, L)$ is partitioned into two subsets Ω_M and $\Omega_\mu = \Omega \setminus \Omega_M$, and that, for simplicity,

$$\Omega_\mu = (0, Kh], \quad \Omega_M = (Kh, L).$$

We will see in Remark 6.3.5 below that the treatment of the general case follows exactly the same lines. Our aim is to study the minimization problem (6.11), where the variational space $X_c(a, \Omega_M)$ is defined by (6.9), with $X_W(\Omega_M) = SBV(\Omega_M)$, and where the corresponding energy E_c is given by (6.10), with $W \equiv W_{\rm LJ}$. The key ingredient of the mathematical analysis is the following observation. Let us choose $x_0 \in \Omega_M$ and let us consider the configurations $u_1 \in X_c(a, \Omega_M)$ and $u_2 \in X_c(a, \Omega_M)$ defined by

$$\begin{array}{ll} \forall i \in \{0, 1, \ldots, K\},\ u_1^i = ih, & \forall x \in \Omega_M,\ u_1(x) = x + (a - L)H(x - x_0), \\ \forall i \in \{1, \ldots, K\},\ u_2^i = ih + (a - L), & \forall x \in \Omega_M,\ u_2(x) = x + (a - L) \end{array} \tag{6.101}$$

and $u_2^0 = 0$. Within the configuration u_1 (resp. u_2), a fracture appears in Ω_M (resp. in Ω_μ). We have $E_c(u_1) = E_c(F \in \Omega_M) = L\, W_{\rm LJ}(1)$ and $E_c(u_2) = E_c(F \in \Omega_\mu) = L\, W_{\rm LJ}(1) + h + O(h^7)$, so

$$E_c(F \in \Omega_\mu) > E_c(F \in \Omega_M). \tag{6.102}$$

So, if the energy is defined by (6.10), a fracture costs less energy when it lies in Ω_M ($F \in \Omega_M$) than when it lies in Ω_μ ($F \in \Omega_\mu$). Hence, the fracture of the minimizers of (6.11) appears in Ω_M, as stated by the following lemma.

Lemma 6.3.3 *For h small enough, the minimizers of problem (6.11) are of the form*

$$u^i = ih \quad \forall i \in \{0, 1, \ldots, K\}, \tag{6.103}$$

$$u(x) = x + \sum_{i \in \mathbb{I}} \tilde{v}_i H(x - x_i), \quad \forall i,\ x_i \in (Kh, L),\ \tilde{v}_i > 0, \tag{6.104}$$

with $\mathbb{I} \subset \mathbb{N}$ and $\displaystyle\sum_{i \in \mathbb{I}} \tilde{v}_i = a - L$.

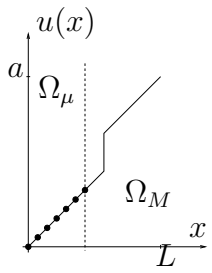

FIG. 6.3 – A minimizer of problem (6.11).

Proof: Since the minimum of $W_{\rm LJ}$ is attained at 1, we have

$$E_c(u) \geq KhW_{\rm LJ}(1) + (L - Kh)W_{\rm LJ}(1),$$

with a strict inequality if for some $i \in \{0, 1, \dots K-1\}$, $\dfrac{u^{i+1} - u^i}{h} \neq 1$. Moreover, this value is attained when u is of the form (6.104). Conversely, if u is a minimizer, we necessarily have (6.103), and the restriction of u to Ω_M is thus a minimizer of

$$I_M^{Kh,L}(a - Kh) = \inf \left\{ \int_{Kh}^{L} W_{\rm LJ}(u'), \quad u \in SBV(Kh, L),\ u' > 0, \right.$$
$$\left. \frac{1}{u'} \in L^{12}(Kh, L), \quad u(Kh) = Kh, \quad u(L) = a \right\}.$$

Now, Theorem 6.3.1 implies that a minimizer of this problem satisfies (6.104). □

Remark 6.3.5 *The case of a general set Ω_μ may be treated likewise. Then, any minimizer of problem (6.11) satisfies :*

$$\begin{aligned} u^{i+1} - u^i &= h \quad \forall i \text{ s.t. } ih \in \Omega_\mu \text{ and } ih + h \in \Omega_\mu, \\ u' &= 1 + \sum_{i \in \mathbb{I}} \tilde{v}_i \delta_{x_i} \text{ in } \Omega_M, \quad \forall i,\ x_i \in \Omega_M, \quad \tilde{v}_i > 0, \quad \text{and } \sum_{i \in \mathbb{I}} \tilde{v}_i = a - L. \end{aligned}$$

The above argument shows that, in the simplest case where we expect a fracture of the material, for any partition of the domain Ω into a regular domain Ω_M and a singular one Ω_μ, the fracture naturally appears in the regular one.

Hence, if the following algorithm is used to compute an approximation of u_μ : initialize Ω_μ to $\emptyset$,

1. Solve (6.11) for $\Omega_M = \Omega \setminus \Omega_\mu$,
2. Find the set where the solution u_c of (6.11) has a large derivative, enlarge Ω_μ correspondingly if necessary and go back to 1,

then Ω_μ converges to Ω, because, at each step, the above algorithm computes a solution having a singularity in the set Ω_M. The latest iterations are therefore as costly as the determination of the atomistic solution.

Remark 6.3.6 *A way around the above difficulty might be to allow the set Ω_μ to shrink back if the computed minimizer happens to be regular enough (in some sense) in Ω_μ. However, we have not been able to solve the difficulty with this alternative strategy.*

Remark 6.3.7 *The preceding analysis can be carried out in the case of a force $f \neq 0$ satisfying $f \in \mathcal{C}^0(\overline{\Omega})$. Let us assume that $a > \theta_M$, where θ_M is defined by (6.59), and that $\inf_{x\in\Omega_M} F_c(x) = \inf_{i\in\mathbb{N}_\mu} F_c^i$, where F_c is defined by (6.17) and (6.18). Let $E_c(F \in \Omega_M)$ (resp. $E_c(F \in \Omega_\mu)$) be the energy of a configuration with a fracture in Ω_M (resp. Ω_μ). Then the inequality (6.102) holds.*

6.3.4 A modified coupled approach

In this subsection, we propose a way to build a coupled problem that remedies to the difficulties observed in the previous subsection.

We again assume, for the time being, that the partition $\Omega = \Omega_M \cup \Omega_\mu$ is given. We show in Theorem 6.3.5 that the modified coupled variational problem (6.13) is well-posed. In Theorem 6.3.6, we show that its solution is a converging approximation of the solution of the atomistic problem (6.2), and in Subsection 6.3.5, we will propose a definition of the partition (see Definition 6.3.1 below).

The variational space we work with is $X_c(a, \Omega_M)$ defined by (6.9) with

$$X_W(\Omega_M) = \left\{ u;\ u \in W^{1,1}(\Omega_M),\ \frac{1}{u'} \in L^{12}(\Omega_M) \right\}.$$

As announced in the Introduction, the modified coupled energy is given by

$$E_{\rm mod}(u) = h \sum_{i\in\mathbb{N}_\mu} W_{\rm LJ}\left(\frac{u^{i+1}-u^i}{h}\right) - h \sum_{i,ih\in\Omega_\mu} u^i f(ih) + \int_{\Omega_M} W_{\rm LJ}^h(u') - uf\,,$$

with

$$W_{\rm LJ}^h(r) = W_{\rm LJ}(r) + \sqrt{h}\,\tau(r - r_0).$$

Here, r_0 is any real number in $(1, r_c)$ and the function τ is a regularization of the function $t \in \mathbb{R} \mapsto t_+ = \max(0, t)$ (in particular, it does not depend on h).

The energy $E_{\rm mod}$ differs from the natural coupled energy E_c given by (6.10) by the use of the energy density $W_{\rm LJ}^h$ instead of $W_{\rm LJ}$ on the continuum domain Ω_M. Let us explain this choice, assuming that there are no body forces. According to the definition (6.12), the energy of the fractured configurations (6.101) reads

$E_{\rm mod}(F \in \Omega_\mu) = LW_{\rm LJ}(1) + h + O(h^7)$ and $E_{\rm mod}(F \in \Omega_M) = LW_{\rm LJ}(1) + (a-L)\sqrt{h}$, so

$$E_{\rm mod}(F \in \Omega_\mu) < E_{\rm mod}(F \in \Omega_M). \tag{6.105}$$

If we compare (6.102) and (6.105), we see that, with the modified definition of the coupled energy, a fracture costs now less energy when it lies in Ω_μ than when it lies in Ω_M.

Let us now assume that the solution u_μ of the atomistic problem (6.2) shows a fracture (on the atom i_μ), and that we want to use a coupled model to compute an approximation of u_μ. At this point, the domain Ω_M is unknown. In order to determine both Ω_M and an approximation of u_μ, a possible algorithm is the one already given in the previous subsection, which consists in iterating over two steps, first solve a coupled problem with Ω_M fixed, second modify the partition according to the computed solution. Assume now that, at some moment, a "correct" partition has been found, in the sense that the atom i_μ (where we expect the fracture to take place) belongs to Ω_μ. At that moment, we want the algorithm to stop, because the zone Ω_μ is satisfactory. Let us consider the minimization problem of the first step of the algorithm. The position of the fracture is such that its energy cost is minimal. If one works with the coupled energy (6.10), then, as explained in the previous subsection, the fracture is located in Ω_M, which is not satisfactory. If one works with the modified coupled energy (6.12), then, in view of (6.105), the fracture is located in Ω_μ and the computed solution is smooth in Ω_M, so the partition is not updated and the algorithm stops.

The difference between E_c defined by (6.10) and $E_{\rm mod}$ defined by (6.12) is of order $\sqrt{h}$, that is a small quantity (recall that h is the atomic lattice parameter). However, even if it is small, this correction has an influence on the choices of the zones, since it allows for the zone Ω_μ to contain the fracture.

Before studying the modified coupled problem (6.13), we study the continuum problem with energy density $W^h_{\rm LJ}$. The following Theorem (which is to be compared with Theorem 6.3.1) will be needed in the sequel.

Theorem 6.3.4 *Consider the energy E^h_M defined by*

$$E^h_M(u) = \int_\Omega W^h_{\rm LJ}(u'(x)) - f(x)u(x)\,dx.$$

Let us set $\beta^h(x) \in (0, r_c)$ such that $W'_{\rm LJ}(\beta^h(x)) = \sqrt{h} + \inf F_M - F_M(x)$ and

$$\theta^h_M = \int_\Omega \beta^h(x)\,dx.$$

If $\theta^h_M \geq a$, then the problem

$$\inf\left\{E^h_M(u),\, u \in W^{1,1}(\Omega),\, \frac{1}{u'} \in L^{12}(\Omega),\, u' > 0 \text{ a.e.},\, u(0) = 0,\, u(L) = a\right\} \tag{6.106}$$

has a unique minimizer which reads

$$u(x) = \int_0^x \psi\left(-\lambda + \inf F_M - F_M(s)\right) ds \tag{6.107}$$

for some $\lambda \geq -\sqrt{h}$ *and where* ψ *is defined by (6.56).*

If $\theta_M^h < a$*, then (6.106) is not attained, but the problem*

$$\inf\left\{ E_M^h(u),\ u \in SBV(\Omega),\ \frac{1}{u'} \in L^{12}(\Omega),\ u' > 0,\ u(0) = 0,\ u(L) = a \right\} \tag{6.108}$$

has at least one minimizer. Moreover, the minimizers of (6.108) are exactly the functions

$$u(x) = \int_0^x \beta^h(t)\, dt + \sum_{i \in \mathbb{I}} \tilde{v}_i H(x - x_i),$$

where $\mathbb{I}$ *is any countable set, and* $\tilde{v}_i$ *and* x_i *are any real numbers such that*

$$\sum_{i \in \mathbb{I}} \tilde{v}_i = a - \theta_M^h \quad \textit{and} \quad \forall i \in \mathbb{I},\ \tilde{v}_i > 0,\ x_i \in \arg\inf F_M.$$

We skip the proof of Theorem 6.3.4, which is an easy adaptation of the proof of Theorem 6.3.1.

We now study the existence and the uniqueness of solutions of the modified coupled problem (6.13). Let F_c be defined by (6.17) and (6.18), and let us set

$$\underline{F_c} = \inf\left(\inf_{\Omega_M} F_c,\ \inf_{i \in \mathbb{N}_\mu} F_c^i \right).$$

The threshold for the appearance of a fracture will be showed to be

$$\theta_{\text{mod}} = \int_{\Omega_M} \left(W'_{\text{LJ}}\right)^{-1} \left(\underline{F_c} - F_c(x)\right)\, dx + h \sum_{i \in \mathbb{N}_\mu} \left(W'_{\text{LJ}}\right)^{-1} \left(\underline{F_c} - F_c^i\right). \tag{6.109}$$

Remark 6.3.8 *The modified coupled problem (6.13) could have been introduced in the convex case, although it was not needed (the coupled problem (6.11) leads to satisfactory results). In this case, one would have proved results similar to those given in Lemmae 6.2.1 and 6.2.3 and Theorem 6.2.1.*

To study the problem (6.13), we need the following lemma :

Lemma 6.3.4 *Let us assume that* $a > \theta_M$*, where* θ_M *is defined by (6.59), and that the partition is such that, for any* h *small enough, there exists* $i_{mod} \in \mathbb{N}_\mu$ *such that*

$$\underline{F_c} \leq F_c^{i_{mod}} \leq \underline{F_c} + Ch \tag{6.110}$$

for some constant C that does not depend on h.

Then there exists h_0 such that, for all $h \leq h_0$, there exists a unique λ_{mod} such that $0 \leq \lambda_{mod} \leq Ch$ and

$$h \sum_{i \in \mathbb{N}_\mu, i \neq i_{mod}} \psi\left(\underline{F_c} - F_c^i + \lambda_{mod}\right) + h\varphi(\underline{F_c} - F_c^{i_{mod}} + \lambda_{mod}) + \int_{\Omega_M} \psi\left(\underline{F_c} - F_c(x) + \lambda_{mod}\right) dx = a \quad (6.111)$$

where ψ and φ are defined by (6.56) and (6.57).

Proof: Let us define $g_{\rm mod}(\lambda) = \lambda + \underline{F_c} - F_c^{i_{\rm mod}} - W'_{\rm LJ}\left(\dfrac{p_{\rm mod}(\lambda)}{h}\right)$ on $[0, 2Ch]$, where

$$p_{\rm mod}(\lambda) = a - h \sum_{i \in \mathbb{N}_\mu, i \neq i_{\rm mod}} \psi\left(\underline{F_c} - F_c^i + \lambda\right) - \int_{\Omega_M} \psi\left(\underline{F_c} - F_c(x) + \lambda\right) dx.$$

For h small enough, $g_{\rm mod}$ is an increasing function and

$$g_{\rm mod}(2Ch) \geq Ch - W'_{\rm LJ}\left(\frac{a - \theta_M}{2h}\right) > 0.$$

For h small enough, we have $g_{\rm mod}(0) < 0$. Hence there exists a unique $\lambda_{\rm mod} \in [0, 2Ch]$ such that $g_{\rm mod}(\lambda_{\rm mod}) = 0$. □

Theorem 6.3.5 (Minimizers of the modified coupled problem) *We assume that $f \in \mathcal{C}^0(\overline{\Omega})$.*

If $a \leq \theta_{mod}$, then problem (6.13) has a unique minimizer u_{mod}, which is smooth : there exists C_0 independant of h such that

$$|u_{mod}|_{W^{1,\infty}(X_c)} \leq C_0, \quad (6.112)$$

where $|\cdot|_{W^{1,\infty}(X_c)}$ is defined by (6.32).

If $a > \theta_M$, and if

$$\exists i_0 \in \mathbb{N}_\mu \text{ such that } \underline{F_\mu} \leq F_\mu^{i_0} \leq \underline{F_\mu} + C_1 h, \quad (6.113)$$

for some constant $C_1 \geq 0$ independent of h (recall $\underline{F_\mu}$ is defined by (6.78)), then there exists $h_0 > 0$ such that, for all $h \leq h_0$, the minimizers of (6.13) are exactly the functions defined, for $i_{mod} \in \mathbb{N}_\mu$ such that $F_c^{i_{mod}} = \inf\limits_{i \in \mathbb{N}_\mu} F_c^i$, by $u_{mod}(0) = 0$ and

$$\begin{cases} u'_{mod}(x) = \psi\left(\lambda_{mod} + \underline{F_c} - F_c(x)\right) \text{ on } \Omega_M, \\ \dfrac{u_{mod}^{i+1} - u_{mod}^i}{h} = \psi\left(\lambda_{mod} + \underline{F_c} - F_c^i\right) \text{ for all } i \in \mathbb{N}_\mu, i \neq i_{mod}, \\ \dfrac{u_{mod}^{i_{mod}+1} - u_{mod}^{i_{mod}}}{h} = \varphi\left(\lambda_{mod} + \underline{F_c} - F_c^{i_{mod}}\right), \end{cases} \quad (6.114)$$

where λ_{mod} is defined by Lemma 6.3.4.

Remark 6.3.9 *If $\theta_{mod} < \theta_M$, Theorem 6.3.5 does not apply to $a \in (\theta_{mod}, \theta_M]$. Note however that $\lim_{h\to 0} \theta_{mod} = \theta_M$, thus all boundary conditions a are asymptotically covered.*

Proof of Theorem 6.3.5: The proof in the case $\theta_{\rm mod} \geq a$ follows the same pattern as the proof of Theorem 6.3.1 (in the case $\theta_M \geq a$). We concentrate on the case $a > \theta_M$. Let us set

$$P_x^c(t) = W_{\rm LJ}^h(t) + \left(F_c(x) - \underline{F_c}\right) t, \tag{6.115}$$
$$P_i^c(t) = W_{\rm LJ}(t) + (F_c^i - \underline{F_c})t, \tag{6.116}$$

and, for $i \in \mathbb{N}_\mu$,

$$v_{1,c}^i = (W'_{\rm LJ})^{-1}\left(\underline{F_c} - F_c^i\right). \tag{6.117}$$

For $x \in \Omega_M$, we also define $\beta_x^c \in (0, r_c)$ such that

$$W'_{\rm LJ}(\beta_x^c) = \sqrt{h} + \underline{F_c} - F_c(x). \tag{6.118}$$

As in the proof of Theorem 6.3.1, we first reformulate the energy $E_{\rm mod}$, which reads

$$E_{\rm mod}(u) = \int_{\Omega_M} P_x^c(u'(x))\,dx + h\sum_{i\in\mathbb{N}_\mu} P_i^c\left(\frac{u^{i+1}-u^i}{h}\right) + a\left(\underline{F_c} - F_c(L)\right).$$

Let us consider the problem

$$\overline{I}_{\rm mod} = \inf\left\{\overline{E}_{\rm mod}(v),\ v \in Y_c(a,\Omega_M)\right\}, \tag{6.119}$$

where $\overline{E}_{\rm mod}$ is given by

$$\overline{E}_{\rm mod}(v) = \int_{\Omega_M} P_x^c(v(x))\,dx + h\sum_{i\in\mathbb{N}_\mu} P_i^c(v^i), \tag{6.120}$$

and where $Y_c(a,\Omega_M)$ is given by

$$Y_c(a,\Omega_M) = \left\{ \begin{array}{c} v;\ v \in L^1(\Omega_M),\ \dfrac{1}{v} \in L^{12}(\Omega_M),\ v_{|\mathbb{N}_\mu} \text{ is the discrete set} \\ \text{of variables } (v^i)_{i\in\mathbb{N}_\mu},\ v > 0,\ \displaystyle\int_{\Omega_M} v + h\sum_{i\in\mathbb{N}_\mu} v^i = a \end{array} \right\}. \tag{6.121}$$

Clearly, $I_{\rm mod} = \overline{I}_{\rm mod} + a\left(\underline{F_c} - F_c(L)\right)$ and u is a minimizer of (6.13) if and only if v, defined by

$$\forall x \in \Omega_M,\ v(x) = u'(x) \quad \text{and} \quad \forall i \in \mathbb{N}_\mu,\ v^i = \frac{u^{i+1}-u^i}{h},$$

is a minimizer of (6.119).

Step 1 : A lower bound on the atomistic deformation :

Let u_n be a minimizing sequence of problem (6.13) and let v_n be the associated minimizing sequence of problem (6.119). Since u_n is an increasing function, we have $0 \leq v_n^i \leq a/h$ for all i in $\mathbb{N}_\mu$. So, up to a subsequence extraction, we can assume that the sequence $\left(v_n^i\right)_n$ converges. Let us set

$$a_\mu^\infty = \lim_{n\to+\infty} h \sum_{i\in\mathbb{N}_\mu} v_n^i. \tag{6.122}$$

As $\displaystyle\int_{\Omega_M} v_n + h\sum_{i\in\mathbb{N}_\mu} v_n^i = a$, we can also define

$$a_M^\infty = \lim_{n\to+\infty} \int_{\Omega_M} v_n(x)\,dx = a - a_\mu^\infty. \tag{6.123}$$

Let us establish a lower bound on a_μ^∞. For all $t > 0$, we have

$$\begin{aligned} P_x^c(t) &\geq P_x^c(\beta_x^c) + \sqrt{h}\,(t - \beta_x^c), &(6.124)\\ P_i^c(t) &\geq P_i^c(v_{1,c}^i), &(6.125)\end{aligned}$$

where P_x^c, P_i^c, β_x^c and $v_{1,c}$ are defined by (6.115), (6.116), (6.118) and (6.117). We now consider (6.120) with $v \equiv v_n$. With (6.124) and (6.125), we obtain

$$\overline{I}_{\rm mod} \geq \int_{\Omega_M} P_x^c(\beta_x^c)\,dx + \sqrt{h}\left(a_M^\infty - \int_{\Omega_M} \beta_x^c\,dx\right) + h\sum_{i\in\mathbb{N}_\mu} P_i^c(v_{1,c}^i). \tag{6.126}$$

Let us now choose $i_{\rm mod} \in \mathbb{N}_\mu$ such that $F_c^{i_{\rm mod}} = \inf\limits_{i\in\mathbb{N}_\mu} F_c^i$. With (6.113) and (6.30), we see that

$$\underline{F_c} \leq F_c^{i_{\rm mod}} \leq \underline{F_c} + Ch \tag{6.127}$$

for some constant C that does not depend on h. Let us consider the test function v_b defined by $v_b(x) = \beta_x^c$ in Ω_M, $v_b^i = v_{1,c}^i$ for all $i \in \mathbb{N}_\mu$, $i \neq i_{\rm mod}$, and $v_b^{i_{\rm mod}}$ such that $\displaystyle\int_{\Omega_M} v_b + h\sum_{i\in\mathbb{N}_\mu} v_b^i = a$. By construction, we have

$$v_b^{i_{\rm mod}} \sim_{h\to 0} (a - \theta_M)/h. \tag{6.128}$$

The function β_x^c is continuous and positive, and $v_b^i > 0$ for all $i \in \mathbb{N}_\mu$. So v_b is a test function for (6.119), and we have

$$\begin{aligned} \overline{I}_{\rm mod} \leq \overline{E}_{\rm mod}(v_b) \;=\; & \int_{\Omega_M} P_x^c(\beta_x^c)\,dx + h\sum_{i\in\mathbb{N}_\mu} P_i^c(v_{1,c}^i) \\ & + hP_{i_{\rm mod}}^c(v_b^{i_{\rm mod}}) - hP_{i_{\rm mod}}^c(v_{1,c}^{i_{\rm mod}}). \end{aligned} \tag{6.129}$$

We now bound the last line of the previous equation, which can be written

$$\begin{aligned} hP^c_{i_{\rm mod}}(v_b^{i_{\rm mod}}) - hP^c_{i_{\rm mod}}(v_{1,c}^{i_{\rm mod}}) &= h\left(W_{\rm LJ}(v_b^{i_{\rm mod}}) - W_{\rm LJ}(v_{1,c}^{i_{\rm mod}})\right) \\ &+ h(F_c^{i_{\rm mod}} - \underline{F_c})(v_b^{i_{\rm mod}} - v_{1,c}^{i_{\rm mod}}). \end{aligned} \tag{6.130}$$

From (6.117), we have $0 < v_{1,c}^{i_{\rm mod}} \leq 1$, thus $W_{\rm LJ}(v_{1,c}^{i_{\rm mod}}) \geq W_{\rm LJ}(1)$. With this inequality and (6.128), we obtain

$$h\left(W_{\rm LJ}(v_b^{i_{\rm mod}}) - W_{\rm LJ}(v_{1,c}^{i_{\rm mod}})\right) \leq Ch \tag{6.131}$$

for some C that does not depend on h. For the second line of (6.130), we see that $\lim_{h\to 0} h(v_b^{i_{\rm mod}} - v_{1,c}^{i_{\rm mod}}) = a - \theta_M$, and, with (6.127), we obtain

$$h(F_c^{i_{\rm mod}} - \underline{F_c})(v_b^{i_{\rm mod}} - v_{1,c}^{i_{\rm mod}}) \leq Ch \tag{6.132}$$

for some C that does not depend on h. Collecting (6.131) and (6.132), we obtain $hP^c_{i_{\rm mod}}(v_b^{i_{\rm mod}}) - hP^c_{i_{\rm mod}}(v_{1,c}^{i_{\rm mod}}) \leq Ch$. Inserting this inequality in (6.129) and making use of (6.126), we see that $a_M^\infty - \displaystyle\int_{\Omega_M} \beta_x^c\, dx \leq C\sqrt{h}$. Since

$$\theta_M = \lim_{h\to 0}\left(\int_{\Omega_M} \beta_x^c\, dx + \int_{\Omega_\mu} (W'_{\rm LJ})^{-1}\left(\inf F_M - F_M(x)\right)\, dx\right),$$

we obtain, for h small enough, the following lower bound on a_μ^∞ :

$$\frac{a-\theta_M}{2} + \int_{\Omega_\mu} (W'_{\rm LJ})^{-1}\left(\inf F_M - F_M(x)\right)\, dx \leq a_\mu^\infty. \tag{6.133}$$

Step 2 : An auxiliary problem :

For any positive real numbers a_M and a_μ, let us define

$$I_M(a_M) = \inf\left\{\int_{\Omega_M} P_x^c(v(x))\, dx,\ v \in L^1(\Omega_M),\ \frac{1}{v} \in L^{12}(\Omega_M),\right.$$
$$\left. v > 0 \text{ a.e. on } \Omega_M,\ \int_{\Omega_M} v = a_M\right\} \tag{6.134}$$

and

$$I_\mu(a_\mu) = \inf\left\{h\sum_{i\in\mathbb{N}_\mu} P_i^c(v^i),\ v^i > 0,\ h\sum_{i\in\mathbb{N}_\mu} v^i = a_\mu\right\}. \tag{6.135}$$

Problem (6.134) is similar to the problem studied in Theorem 6.3.4 : let

$$a_M^{th} = \int_{\Omega_M} \psi\left(\sqrt{h} + \inf_{\Omega_M} F_c - F_c(x)\right) dx. \tag{6.136}$$

If $a_M \leq a_M^{th}$, then problem (6.134) has a unique minimizer which is continuous, and if $a_M > a_M^{th}$, then problem (6.134) has no minimizer (any minimizing sequence converges to some measure which includes Dirac masses).

Problem (6.135) is similar to the problem (6.85) studied in Theorem 6.3.3. Let

$$a_\mu^{th} = \int_{\Omega_\mu} (W'_{\rm LJ})^{-1} \left(\inf_\Omega F_M - F_M(x) \right) dx. \tag{6.137}$$

By arguments similar to the ones used in Theorem 6.3.3, one can show that, if $a_\mu > a_\mu^{th}$, then a fracture appears in the minimizer(s) of (6.135).

With (6.119), (6.120) and (6.121), we see that

$$\overline{I}_{\rm mod} = \inf \left\{ I_M(a - \overline{a}) + I_\mu(\overline{a}),\ \overline{a} \in [0, a] \right\}.$$

Let us define the function g by

$$g(\overline{a}) = I_M(a - \overline{a}) + I_\mu(\overline{a}), \tag{6.138}$$

and let $\overline{a}^*$ be any real number in $[0, a]$ such that $\overline{I}_{\rm mod} = \inf g = g(\overline{a}^*)$. One can consider a minimizing sequence of problem (6.135) with $a_\mu = \overline{a}^*$, and one can also consider a minimizing sequence of problem (6.134) with $a_M = a - \overline{a}^*$. So we can build a minimizing sequence v_n of problem (6.119) and apply results of Step 1 to this sequence. By construction, $h \sum_{i \in \mathbb{N}_\mu} v_n^i = \overline{a}^*$, so the real number a_μ^∞ defined by (6.122) reads $a_\mu^\infty = \overline{a}^*$. Hence, with (6.133) and (6.137), we see that $\overline{a}^* \in I_g$, where we set

$$I_g = \left[\frac{a - \theta_M}{2} + a_\mu^{th}\ ,\ a \right].$$

So $\overline{I}_{\rm mod} = \inf \{ g(\overline{a}),\ \overline{a} \in I_g \}$. In the sequel on this Step, we study the variations of the function g to show that, on the interval I_g, g has a unique minimizer.

Let us choose $\overline{a} \in I_g$, let us set $a_\mu = \overline{a}$ and $a_M = a - \overline{a}$ and let us consider problems (6.134) and (6.135). Since $\overline{a} > a_\mu^{th}$, problem (6.135) is an atomistic problem with boundary conditions so that a fracture appears. With results of Section 6.3.2, one can show that there exists $\lambda_\mu \in [0, h]$ such that the minimizers v of (6.135) read

$$v^{i_{\rm mod}} = \varphi(\lambda_\mu) \quad \text{and} \quad \forall i \neq i_{\rm mod},\ v^i = \psi(\inf_{i \in \mathbb{N}_\mu} F_c^i - F_c^i + \lambda_\mu),$$

where $i_{\rm mod}$ is any index such that $F_c^{i_{\rm mod}} = \inf_{i \in \mathbb{N}_\mu} F_c^i$. We have also shown that λ_μ depends on a_μ but not on $i_{\rm mod}$. As a consequence, we can compute $I_\mu(a_\mu)$ and show that

$$\frac{dI_\mu(a_\mu)}{da_\mu} = \lambda_\mu + \inf_{i \in \mathbb{N}_\mu} F_c^i - \underline{F_c}. \tag{6.139}$$

In addition,

$$\frac{d\lambda_\mu}{da_\mu} \sim_{h \to 0} -Ch^7 \tag{6.140}$$

for some constant $C > 0$. We now study problem (6.134). By similar arguments to the ones used to prove Theorem 6.3.4, we can show that, if $a_M \leq a_M^{th}$ defined by (6.136), then problem (6.134) has a unique minimizer v which reads

$$v(x) = \psi\left(-\lambda_M + \inf_{\Omega_M} F_c - F_c(x)\right)$$

for some $\lambda_M \geq -\sqrt{h}$. If $a_M > a_M^{th}$, then problem (6.134) has one or many minimizers v which read

$$v(x) = \psi\left(\sqrt{h} + \inf_{\Omega_M} F_c - F_c(x)\right) + \text{ Dirac masses .}$$

So we can compute $I_M(a_M)$ and show that

$$\frac{dI_M(a_M)}{da_M} = -\lambda_M + \inf_{\Omega_M} F_c - \underline{F_c}, \tag{6.141}$$

where $\lambda_M \in [-\sqrt{h}, +\infty)$. If $a_M \geq a_M^{th}$, then $\lambda_M = -\sqrt{h}$, and if $a_M \leq a_M^{th}$, then

$$\frac{d\lambda_M}{da_M} \leq -\frac{W''_{\rm LJ}(1)}{2\,|\Omega_M|}. \tag{6.142}$$

We now study the variations of the function g defined by (6.138). In view of (6.139) and (6.141), it satisfies

$$g'(\overline{a}) = \lambda_M - \inf_{\Omega_M} F_c + \lambda_\mu + \inf_{i\in\mathbb{N}_\mu} F_c^i. \tag{6.143}$$

If $\inf\limits_{i\in\mathbb{N}_\mu} F_c^i \geq \inf\limits_{\Omega_M} F_c$, then $\underline{F_c} = \inf\limits_{\Omega_M} F_c$, and with (6.127), we obtain

$$\inf_{i\in\mathbb{N}_\mu} F_c^i - \inf_{\Omega_M} F_c \leq Ch. \tag{6.144}$$

The above inequality also holds if $\inf\limits_{i\in\mathbb{N}_\mu} F_c^i < \inf\limits_{\Omega_M} F_c$. Inserting (6.144) in (6.143), and since $\lambda_\mu \leq h$, we obtain

$$g'(\overline{a}) \leq \lambda_M + Ch$$

for some constant C that does not depend on h.

If $a - \overline{a} \geq a_M^{th}$ then $\lambda_M = -\sqrt{h}$, so $g'(\overline{a}) < 0$. If $a - \overline{a} \leq a_M^{th}$, we differentiate (6.143) with respect to $\overline{a}$, and with (6.140) and (6.142), one can show that $g''(\overline{a}) > 0$. Since $g'(a) = +\infty$, there exists a unique $\overline{a}^*$ which minimizes g on I_g, and $\overline{a}^* \in (a - a_M^{th}, a)$.

Step 3 : Conclusion :

We know that $\overline{I}_{\rm mod} = \inf g = g(\overline{a}^*)$, and that there exists minimizers for problem (6.134) with $a_M = a - \overline{a}^* < a_M^{th}$ and for problem (6.135) with $a_\mu = \overline{a}^*$. So problems (6.119) and (6.13) have a minimizer. Let now $u_{\rm mod}$ be a minimizer of problem (6.13)

and let $v_{\rm mod}$ be its first derivative, which is a minimizer of (6.119). As g has a unique minimizer $\overline{a}^*$, we see that

$$\int_{\Omega_M} v_{\rm mod}(x)\,dx = a - \overline{a}^* < a_M^{th},$$

so that $u_{\rm mod}$ has no fracture on Ω_M. We also see that

$$h\sum_{i\in\mathbb{N}_\mu} v_{\rm mod}^i = \overline{a}^* \geq \frac{a-\theta_M}{2} + a_\mu^{th},$$

so $u_{\rm mod}$ has a unique fracture on Ω_μ. We apply Theorem 6.3.3 on problem (6.135) and Theorem 6.3.4 on problem (6.134), and we obtain (6.114) for some Lagrange multiplier $\lambda_{\rm mod}^* \in (0, Ch)$.

We now identify $\lambda_{\rm mod}^*$. We know that $\lambda_{\rm mod}^* \in (0, Ch)$, and we see from (6.114) that $\lambda_{\rm mod}^*$ satisfies (6.111). In addition, we see that (6.113) implies (6.110). Since $a > \theta_M$, we can apply Lemma 6.3.4, that defines $\lambda_{\rm mod}$, and we have $\lambda_{\rm mod}^* = \lambda_{\rm mod}$. □

We now estimate the difference between a minimizer $u_{\rm mod}$ of the modified coupled problem (6.13) and a minimizer u_μ of the atomistic problem (6.2).

Theorem 6.3.6 (Estimate on the minimizers) *We assume that $f \in \mathcal{C}^0(\overline{\Omega})$ and that there exists $\kappa_f > 0$ such that (6.28) is satisfied. Let u_μ be a minimizer of problem (6.2) and u_{mod} be a minimizer of problem (6.13).*

If $a < \theta_M$, then there exist $h_0 \leq 1$ and $C(\kappa_f)$, that both depend on κ_f, such that, for all $h \leq h_0$, the minimizers u_μ and u_{mod}, as well as their respective energies, are at distance of order h in the sense that

$$|(\Pi_\mu u_{mod}) - u_\mu|_{W^{1,\infty}(X_\mu)} + |u_{mod} - (\Pi_c u_\mu)|_{W^{1,\infty}(X_c)} \leq C(\kappa_f)h\kappa_f, \tag{6.145}$$

$$\|(\Pi_\mu u_{mod}) - u_\mu\|_{L^\infty(X_\mu)} + \|u_{mod} - (\Pi_c u_\mu)\|_{L^\infty(X_c)} \leq C(\kappa_f)h\kappa_f, \tag{6.146}$$

$$|I_{mod} - I_\mu| \leq C(\kappa_f)h\kappa_f, \tag{6.147}$$

where Π_c and Π_μ are the operators defined in Definition 6.2.3, and $\|\cdot\|_{L^\infty(X_c)}$, $|\cdot|_{W^{1,\infty}(X_c)}$, $\|\cdot\|_{L^\infty(X_\mu)}$ and $|\cdot|_{W^{1,\infty}(X_\mu)}$ are the norms defined by (6.31), (6.32) and (6.33). In addition, the function $\kappa_f \mapsto C(\kappa_f)$ is bounded on any compact.

If $a > \theta_M$, and if the partition $\Omega = \Omega_M \cup \Omega_\mu$ satisfies (6.113), then there exist $h_0 \leq 1$ and a constant C (that both depend on κ_f) such that, for all $h \leq h_0$, there

exist $i_\mu \in \{0, \dots, N-1\}$ *and* $i_{mod} \in \mathbb{N}_\mu$ *so that*

$$\|u'_{mod} - (\Pi_c u_\mu)'\|_{L^\infty(\Omega_M \setminus [i_\mu h, i_\mu h + h])} \leq Ch, \tag{6.148}$$

$$\sup_{i \in \mathbb{N}_\mu, i \neq i_\mu, i \neq i_{mod}} \left| \frac{u^{i+1}_{mod} - u^i_{mod}}{h} - \frac{u^{i+1}_\mu - u^i_\mu}{h} \right| \leq Ch, \tag{6.149}$$

$$\left| (u^{i_{mod}+1}_{mod} - u^{i_{mod}}_{mod}) - (u^{i_\mu+1}_\mu - u^{i_\mu}_\mu) \right| \leq Ch, \tag{6.150}$$

$$u^{i_{mod}+1}_{mod} - u^{i_{mod}}_{mod} \sim_{h \to 0} a - \theta_M, \qquad u^{i_\mu+1}_\mu - u^{i_\mu}_\mu \sim_{h \to 0} a - \theta_M, \tag{6.151}$$

$$|I_{mod} - I_\mu| \leq Ch. \tag{6.152}$$

In the case $a > \theta_M$, we see that both u_μ and $u_{\rm mod}$ have a singularity, that is localized on a unique atom pair (see (6.151)). With (6.150), we can see that the difference between the discontinuity of $u_{\rm mod}$ and of u_μ converges to 0 when h goes to zero.

Proof: We first treat the case $a < \theta_M$. Then, for h small enough, the atomistic problem (6.2) and the modified coupled problem (6.13) are well-posed (see Theorems 6.3.3 and 6.3.5), and the analysis can be conducted in exactly the same way as in the convex case (see Theorem 6.2.1). We thus obtain estimates (6.145), (6.146) and (6.147), which are similar to (6.36), (6.37), (6.38), (6.39) and (6.40).

We now treat the case $a > \theta_M$. Then, for h small enough, the configuration u_μ, which is a minimizer of the atomistic problem (6.2), is given by (6.82) for some $i_\mu \in \{0, \dots, N-1\}$ and some $\lambda_\mu \in (0, h]$ (see Theorem 6.3.3). The configuration $u_{\rm mod}$, which is a minimizer of problem (6.13), is given by (6.114) for some $i_{\rm mod} \in \mathbb{N}_\mu$ and some $\lambda_{\rm mod} \in (0, Ch]$ (see Theorem 6.3.5). Hence, we obtain

$$|\lambda_\mu - \lambda_{\rm mod}| \leq Ch, \tag{6.153}$$

for some constant C that does not depend on h. Collecting (6.82), (6.114) and (6.153), we obtain estimates (6.148) and (6.149). Using boundary conditions, one can prove (6.150). The estimate (6.83) implies the estimate (6.151) on u_μ, and we infer from the latter and (6.150) the estimate (6.151) on $u_{\rm mod}$. Collecting (6.148), (6.149) and (6.150), we obtain energy estimate (6.152). □

6.3.5 Definition of the partition

In the statement of Theorem 6.3.6, we have supposed that we were given some body forces f and a partition $\Omega = \Omega_M \cup \Omega_\mu$ satisfying (6.28) and (6.113). In the sequel, we describe a strategy to find such a partition without resorting to the computation of F_μ, an object which is expensive to compute.

Definition 6.3.1 (Partition in the Lennard-Jones case) *We assume that* $f \in W^{1,1}(\Omega)$. *Let us fix* $\kappa_f > 0$, *and let* $u_M \in SBV(\Omega)$ *be a minimizer of the continuum problem (6.62).*

The interval $(ih, ih+h)$ *is said to be a* regular interval *if* f *satisfies*

$$\forall x \in (ih, ih+h),\ |f(x)| \leq \kappa_f \quad \textit{and} \quad \int_{ih}^{ih+h} |f'(x)|\,dx \leq h\frac{\kappa_f}{L},$$

and if u_M *is continuous on* $[ih, ih+h]$.

We define Ω_M *by*

$$\Omega_M = \cup^*_{(ih,ih+h)\ \textit{regular}}(ih, ih+h) \textit{ and } \Omega_\mu = \Omega \setminus \Omega_M,$$

where $\cup^*$ *means that the point* $\{ih\}$ *is also included in* Ω_M *if both* $(ih-h, ih)$ *and* $(ih, ih+h)$ *are regular intervals.*

Note that we make a stronger assumption on f than before (until now, we have only assumed $f \in \mathcal{C}^0(\overline{\Omega})$).

Remark 6.3.10 *In the case* $a \leq \theta_M$*, one can show that problem (6.62) has a unique minimizer* u_M*, which satisfies* $u'_M = v_0$*, where* v_0 *is defined by (6.64) (the proof of this fact is similar to the one presented in Theorem 6.3.1, in the case* $a \leq \theta_M$*). So* u_M *is continuous on* Ω *and the last condition for an interval to be regular is always satisfied.*

Theorem 6.3.7 *Let us consider a partition* $\Omega = \Omega_M \cup \Omega_\mu$ *as defined by Definition 6.3.1. Then the body forces* f *satisfy (6.28). If* $a > \theta_M$*, then the function* F_μ *defined by (6.16) satisfies (6.113).*

Proof: Estimates (6.28) are a direct consequence of the definition of Ω_M. We now assume $a > \theta_M$ and prove (6.113). Let us first note that the assumption $f \in W^{1,1}(\Omega)$ implies that

$$\forall k \in \{0, \ldots, N\}, \quad \left|F_M(kh) - F_\mu^k\right| \leq h\|f'\|_{L^1(\Omega)}, \tag{6.154}$$

which is a better estimate than (6.29).

Theorem 6.3.1 shows that the minimizers u_M of (6.62) read

$$u_M(x) = \int_0^x v_1(t)\,dt + \sum_{i\in\mathbb{I}} \tilde{v}_i H(x - x_i),$$

where $v_1 \in L^1(\Omega)$ is defined by (6.60), $\mathbb{I}$ is any countable set, and $\tilde{v}_i$ and x_i are such that, for all $i \in \mathbb{I}$, $\tilde{v}_i > 0$ and $x_i \in \arg\inf F_M$. So u_M is not continuous at x_1. Let σ_1 be such that $x_1 \in [\sigma_1 h, \sigma_1 h + h)$. So the interval $(\sigma_1 h, \sigma_1 h + h)$ is not regular, thus

$$[\sigma_1 h, \sigma_1 h + h] \subset \Omega_\mu, \text{ i.e. } \sigma_1 \in \mathbb{N}_\mu.$$

With (6.154), we see that

$$|F_M(\sigma_1 h) - F_\mu^{\sigma_1}| \leq h\|f'\|_{L^1(\Omega)}. \tag{6.155}$$

We also have

$$|F_M(x_1) - F_M(\sigma_1 h)| \leq \int_{\sigma_1 h}^{x_1} |f(s)| ds \leq h \|f\|_{L^\infty(\Omega)}. \tag{6.156}$$

We also infer from (6.154) that

$$|\inf F_M - \underline{F_\mu}| \leq Ch \tag{6.157}$$

for some C that does not depend on h, and where $\underline{F_\mu} = \inf_{0 \leq i \leq N-1} F_\mu^i$. As $F_M(x_1) = \inf F_M$, we infer from (6.155), (6.156) and (6.157) that

$$|F_\mu^{\sigma_1} - \underline{F_\mu}| \leq Ch.$$

As $\sigma_1 \in \mathbb{N}_\mu$, we obtain (6.113). □

Chapitre 7

Homogénéisation numérique de polycristaux

Les résultats présentés dans ce chapitre font l'objet d'un article à paraître dans *Computational and Applied Mathematics* [P5].

Nous nous intéressons ici à des polycristaux élasto-viscoplastiques. Un polycristal est un agrégat de monocristaux (les grains), et constitue donc un matériau hétérogène, où la taille caractéristique de variation des propriétés mécaniques est celle du grain. Nous étudions d'un point de vue numérique l'homogénéisation d'un tel matériau, en cherchant à déterminer une loi de comportement effective pour un polycristal composé d'un nombre suffisamment grand de monocristaux.

Numerical homogenization of nonlinear viscoplastic two-dimensional polycrystals

Frédéric Legoll[a,b]

[a] *CERMICS, Ecole Nationale des Ponts et Chaussées, 6 et 8 avenue Blaise Pascal, 77455 Marne-la-Vallée Cedex 2*
and
MICMAC, INRIA Rocquencourt, Domaine de Voluceau, 78153 Le Chesnay Cedex
[b] *EDF R & D, Analyse et Modèles Numériques, 1, avenue du Général de Gaulle, 92140 Clamart*

legoll@cermics.enpc.fr

In this article, we numerically determine the effective stress-strain relation of some two-dimensional polycrystals. These are aggregates of a few tens of perfectly bonded single-crystal (hexagonal atomic lattice) grains, with varying orientations. Each grain obeys a given nonlinear viscoplastic stress-strain relation, which depends on the orientation of the grain. Precise calculations performed with this microscopic model are compared with calculations done with a macroscopic approximate model (in which matter has no microstructure) in order to determine the macroscopic constitutive law. We find an effective behaviour for the stationary response which appears to be also consistent for the transient response. The influence of the number of grains as well as that of the distribution of grain orientations are investigated.

7.1 Introduction

The theoretical prediction of the effective response of a heterogeneous material is still an essentially open question. In some few simple cases, an analytic closed form expression is known. For instance, this is the case for a linear elastic matrix with linear elastic inclusions, in the dilute limit (that is, inclusions are considered too far away from one another to have an interaction) [157].

A more general case is that of a material presenting *material nonlinearity* [130, 148]. The constitutive law (also named the *stress-strain relation*) of such a material reads

$$\varepsilon(x,t) = \frac{\partial U}{\partial \sigma}(x, \sigma(x,t)), \tag{7.1}$$

where the real-valued function U is the heterogeneous *elastic stress potential*, $\sigma(x,t)$ is the *stress tensor*, and $\varepsilon(x,t)$ is the *strain tensor*. Throughout this article, we work

under the assumption of small perturbations (small strain, small displacement). All fields are defined on the reference configuration, and $\varepsilon(x,t)$ is linked to the displacement field $u(x)$ by the linearized *compatibility equation*

$$\varepsilon = \frac{1}{2}\left(\nabla u + {}^t\nabla u\right). \tag{7.2}$$

In this setting, one can derive various bounds and estimates on the effective behaviour [192, 194, 198–200]. Let us note that, in general, no closed form expression for the effective elastic stress potential is available.

In this article, we will consider the *elasto-viscoplastic* materials whose constitutive law reads

$$\dot{\varepsilon}(x,t) = \frac{\partial U^{vp}}{\partial \sigma}(x,\sigma(x,t)) + \frac{\partial U^{e}}{\partial \dot{\sigma}}(x,\dot{\sigma}(x,t)), \tag{7.3}$$

where U^{vp} is the *viscoplastic stress potential* (also referred to as the dissipation potential) and U^e is the *elastic stress potential*. So the *strain rate tensor* $\dot{\varepsilon}(x,t)$, which is the time derivative of the strain tensor, depends both on the stress tensor σ and the stress rate tensor $\dot{\sigma}$. In such a case, when the stress-strain relation cannot be written with a unique potential, there are no theoretical bounds known.

In the following, we numerically investigate the effective behaviour of a heterogeneous polycrystal obeying such an elasto-viscoplastic law [154]. With a view to studying a more realistic and complex model in the future, we want to check here whether an effective constitutive law of type (7.3) can be inferred from the examination of the material at lower scale.

The article is organized as follows. The polycrystal model is presented in Section 7.2. Let us just mention in this Introduction that a *polycrystal* is an aggregate of perfectly bonded single-crystal grains, and that each grain is homogeneous and obeys a given nonlinear stress-strain relation. This relation depends on parameters which are not the same from one grain to another one, thus making the polycrystal heterogeneous. Section 7.3 is dedicated to the theoretical study of such a heterogeneous law. We first recall some definitions and classical results on the derivation of an effective law for heterogeneous materials, by a homogenization procedure. As above stated, the classical procedure does not apply for our model, since the microscopic law cannot be written by using a single potential. We however decide to make use of the classical procedure separately on the elastic potential and on the viscoplastic potential, thus obtaining an effective elastic potential and an effective viscoplastic potential, up to some unknown parameters. Collecting these two effective potentials, we are able to postulate some expression for the effective constitutive law (see (7.17) below).

Our aim is to use, in the future, the effective law in the following way. Computing the response of a structure (composed of a large number of grains) by using the microscopic law is very expensive. Recall that, if one uses a finite element method, the mesh size has to be smaller than the grain size. Using an effective homogeneous law is much cheaper, for it allows for larger mesh sizes. In this article, as a first step,

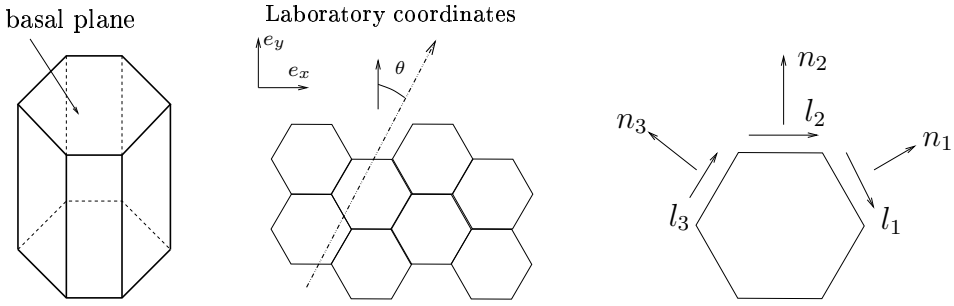

FIG. 7.1 – Atomic lattice inside a grain : 3D unit cell (left), 2D section along the basal plane inside a grain (center), the 3 slip systems we take into account (right). The orientation of the grain is given by the angle θ.

we look for an effective constitutive law which is consistent with the microscopic law. This consistency is checked by comparing the numerical results that are obtained on the basis of the effective law with the numerical results that are obtained (through a costly calculation) with the microscopic law. For this purpose, we choose some test problems, and make two computations, one with the macroscopic model, one with the microscopic model (by using a very fine finite element mesh). Numerical results are given in Section 7.4.

7.2 The microscopic model

The materials we deal with are metals that have a hexagonal atomic lattice (see Fig. 7.1). The orientation of the lattice is not uniform in the material : by definition, a *grain* is a domain of the material in which the orientation stays constant, and a polycrystal is a set of a large number of perfectly-bonded grains [154]. We will only consider polycrystals made of grains of isotropic shape (there is no special direction in the grain shape). For the materials we deal with, the characteristic size of a grain is 5.10^{-6} m, which is much larger than the atomic scale (10^{-10} m) : so it is possible to use a continuum model to describe the constitutive relation inside a grain. At this scale, the stress tensor is $\sigma_\mu(x,t)$, the displacement is $u_\mu(x,t)$ and the strain tensor $\varepsilon_\mu(x,t)$ is linked to the displacement by the linearized compatibility equation (7.2). We do not include in our model any grain interface properties, and we only suppose that the displacement and the stress vector are continuous at the grain interfaces. The stress-strain relation inside a grain depends on its orientation, and the heterogeneity in the polycrystal comes from the fact that this orientation is not the same from one grain to another one.

We suppose in the following that for all grains, the *basal plane* of the atomic lattice (see Fig. 7.1) is the same, namely the (e_x, e_y) plane. So, the grain orientation is defined by an angle between 0 and $\pi/6$. We also assume that the grain orientations occur with equal probability (there are actually very few experimental data for the

metals we deal with, so this assumption is the most sensible one).

Let us now write the stress-strain relation inside a grain. In the metals we study, there are 12 preferred *slip systems*, defined by the plane in which the slip takes place (the normal direction to this plane is denoted by n_s), and by the slip direction l_s. Here, the vectors $n_s(x)$ and $l_s(x)$ depend on the space variable x, as they change from one grain to another one. In this article, we want to work in a 2D geometry in the (e_x, e_y) plane, so we only take into account the 3 systems for which the vectors $n_s(x)$ and $l_s(x)$ belong to the (e_x, e_y) plane (see Fig. 7.1). Knowing the slip systems, one can compute the *orientation tensors* $m_s(x)$, which are defined by

$$m_s(x) = \frac{1}{2}\left(n_s(x) \otimes l_s(x) + l_s(x) \otimes n_s(x)\right). \tag{7.4}$$

The strain rate tensor $\dot{\varepsilon}_\mu$ is the sum of two terms, the *elastic strain rate tensor* $\dot{\varepsilon}^e_\mu$ and the *viscoplastic strain rate tensor* $\dot{\varepsilon}^{vp}_\mu$. The elastic term is given by the linear Hooke law. We do not include in our model any nonlinear elastic effects, for they are small in comparison to the efforts we account for. We suppose that the elastic characteristics are homogeneous and isotropic in the polycrystal. Using the Young modulus E and the Poisson ratio ν, the elastic term reads

$$\varepsilon^e_\mu(x,t) = \frac{1+\nu}{E}\,\sigma_\mu(x,t) - \left(\frac{\nu}{E}\,\mathrm{tr}\,\sigma_\mu(x,t)\right)\,I, \tag{7.5}$$

where I is the identity 3×3 tensor. On the other hand, we assume the viscoplastic term to be of a power-law type

$$\dot{\varepsilon}^{vp}_\mu(x,t) = \sum_{s=1}^{3}\left(\frac{|\,\sigma_\mu(x,t) : m_s(x)\,|}{K_\mu}\right)^n \mathrm{sign}\left(\sigma_\mu(x,t) : m_s(x)\right)\; m_s(x). \tag{7.6}$$

We make the assumption that the parameters n and K_μ of the power-law are the same for all grains. So, as mentioned above, the heterogeneity from one grain to another one just comes from the fact that the orientation tensors $m_s(x)$ are not the same.

So, the constitutive relation inside a grain reads

$$\dot{\varepsilon}_\mu(x,t) = \dot{\varepsilon}^e_\mu(x,t) + \dot{\varepsilon}^{vp}_\mu(x,t). \tag{7.7}$$

Recasting (7.7) in the form of (7.3), we see that, in our case, the microscopic stress potentials (introduced in (7.3)) read

$$U^e_\mu(\dot{\sigma}_\mu) = \frac{1}{2}\,\dot{\sigma}_\mu \cdot \Lambda \cdot \dot{\sigma}_\mu, \tag{7.8}$$

$$U^{vp}_\mu(x,\sigma_\mu) = \frac{1}{n+1}\left(\frac{1}{K_\mu}\right)^n \sum_{s=1}^{3} |\,\sigma_\mu : m_s(x)\,|^{n+1}, \tag{7.9}$$

where the fourth order tensor Λ only depends on E and ν (see (7.5)).

Let Ω be the region occupied by the polycrystal in the reference configuration. Solving the microscopic model consists in searching for the displacement field $u_\mu(x,t)$ solution to the equilibrium equation

$$\forall x \in \Omega,\ \forall t \in [0,T], \qquad \operatorname{div} \sigma_\mu(x,t) = 0, \tag{7.10}$$

along with constitutive laws (7.5 - 7.6 - 7.7), compatibility equation (7.2) and convenient initial and boundary conditions.

Quantitatively, we use the following numerical values :

$$E = 105\,000 \text{ MPa}\,, \qquad \nu = 0.43\,, \qquad K_\mu = 178 \text{ MPa}\,, \qquad n = 6.5.$$

7.3 The homogenization procedure

In Section 7.3.1, we first briefly recall the classical homogenization procedure [199] used in the stationary case when the stress-strain relation can be written by using a single potential. Next, in Section 7.3.2, we use the procedure to determine the analytical expression, up to some parameters, of the effective behaviour of the polycrystal. Henceforth, there are no body forces.

7.3.1 Classical homogenization procedure

Let us consider an elastic material (see Section 7.1) in the stationary case, described by a heterogeneous microscopic stress potential $U_\mu(x,\sigma_\mu)$. The constitutive law is given by (7.1) with $U \equiv U_\mu$. We suppose that U_μ is strictly convex with respect to σ_μ. The microscopic *deformation potential* $W_\mu(x,\varepsilon_\mu)$ is defined as the Legendre transform of U_μ with respect to σ_μ.

We can first work with the displacement as the unknown and define the so-called *effective deformation potential* W_M. For a given symmetric constant tensor ε_M, $W_M(\varepsilon_M)$ is defined by

$$W_M(\varepsilon_M) = \inf\Big\{ \langle\, W_\mu(x,\varepsilon_\mu(x))\,\rangle,\ \varepsilon_\mu(x) \in K(\varepsilon_M) \Big\}, \tag{7.11}$$

where $\langle\,\cdot\,\rangle$ is the average over Ω and the minimization space is defined by

$$K(\varepsilon_M) = \Big\{ \varepsilon_\mu(x);\ \exists\, u_\mu(x) \text{ satisfying (7.2) in } \Omega,\ u_\mu(x) = \varepsilon_M \cdot x \text{ on } \partial\Omega \Big\}.$$

Note that, as a consequence of (7.2), all strain tensors ε_μ in $K(\varepsilon_M)$ satisfy $\langle\, \varepsilon_\mu(x)\,\rangle = \varepsilon_M$. Let $\overline{\varepsilon}_\mu(x)$ be the minimizer of problem (7.11), and let $\overline{\sigma}_\mu(x)$ be the microscopic stress field at equilibrium. The effective strain and stress tensors are defined as the averages over Ω of the microscopic tensors. We have already noticed

that the effective strain tensor is ε_M. We set $\sigma_M = \langle\, \overline{\sigma}_\mu(x)\,\rangle$. One can show that effective tensors and potential are linked by

$$\sigma_M = \frac{\partial W_M}{\partial \varepsilon_M}(\varepsilon_M). \tag{7.12}$$

For completeness, let us mention that there are other ways to define an effective potential. We have so far worked with the deformation potential $W_\mu(x, \varepsilon_\mu)$, we may alternatively work with the stress potential $U_\mu(x, \sigma_\mu)$, the stress field being the unknown. The so-called *effective stress potential* U_M is defined by

$$U_M(\sigma_M) = \inf\Big\{ \langle\, U_\mu(x, \sigma_\mu(x))\,\rangle,\ \sigma_\mu(x) \in S(\sigma_M) \Big\}, \tag{7.13}$$

where σ_M is a given symmetric constant tensor, and where the minimization space is defined by

$$S(\sigma_M) = \Big\{ \sigma_\mu(x);\ \sigma_\mu(x) \cdot n(x) = \sigma_M \cdot n(x) \text{ on } \partial\Omega,\ \operatorname{div} \sigma_\mu = 0 \text{ in } \Omega \Big\}.$$

Let $\overline{\sigma}_\mu(x)$ be the minimizer of problem (7.13), and let $\overline{\varepsilon}_\mu(x)$ be the microscopic strain field at equilibrium, which is related to $\overline{\sigma}_\mu(x)$ by (7.1) where U is replaced by U_μ. Again, effective tensors are defined as averages over Ω of microscopic tensors. All stress tensors $\sigma_\mu(x)$ in $S(\sigma_M)$ satisfy $\langle\, \sigma_\mu(x)\,\rangle = \sigma_M$, so the effective stress tensor is σ_M. We set $\varepsilon_M = \langle\, \overline{\varepsilon}_\mu(x)\,\rangle$. As in the first case, one can show that effective tensors and potential are linked by $\varepsilon_M = \dfrac{\partial U_M}{\partial \sigma_M}(\sigma_M)$.

One says that the material follows an *effective stress-strain relation* if the effective stress potential U_M defined by (7.13) is the Legendre transform, with respect to the macroscopic strain tensor ε_M, of the effective deformation potential W_M defined by (7.11).

Remark *The homogenization procedure we have just recalled is based on calculus of variations, and no quantity depends on time. In the time-dependent case, under the quasistatic approximation, it is also possible to define an effective deformation potential and an effective stress potential, by the same procedure as above.*

7.3.2 Homogenization of the polycrystal law

We now proceed to the homogenization of the polycrystal model presented in Section 7.2. Constitutive laws are (7.5 - 7.6 - 7.7), corresponding potentials are defined by (7.8 - 7.9), and the equilibrium equation is (7.10). When writing this equation, we have neglected the acceleration. As two potentials are involved, and as the constitutive law is time-dependent, we cannot directly use the theory we have just recalled. However, we can apply the theory separately on the elastic stress potential and on the viscoplastic stress potential. Indeed, if we only consider one

potential, we are in the setting detailed in Section 7.3.1. Actually, the procedure is immediate for the elastic potential as elastic properties are homogeneous in the polycrystal. We thus focus on the viscoplastic stress potential. To simplify notation, let $d_\mu = \dot{\varepsilon}_\mu^{vp}$ denote the microscopic viscoplastic strain rate tensor.

7.3.2.1 The viscoplastic term

We first note that tensors m_s are symmetric and satisfy $m_s^{xx} = -m_s^{yy}$ and $m_s^{xz} = m_s^{yz} = m_s^{zz} = 0$ (see (7.4)). Hence, for any symmetric microscopic stress tensor σ_μ, we have $\sigma_\mu(x) : m_s(x) = \alpha_\mu(x)u_s(x) + \beta_\mu(x)v_s(x)$, where we set

$$\alpha_\mu = \sigma_\mu^{xx} - \sigma_\mu^{yy}, \quad \beta_\mu = 2\,\sigma_\mu^{xy}, \quad u_s = m_s^{xx}, \quad v_s = m_s^{xy}.$$

With (7.6), we note that tensors d_μ only depend on two scalar variables, d_μ^{xx} and d_μ^{xy}. So the only variables to consider are d_μ^{xx}, d_μ^{xy}, α_μ and β_μ. The potential $U_\mu^{vp}(x, \sigma_\mu)$ that we introduced in (7.9) is not strictly convex with respect to σ_μ, but if we rewrite it in terms of (α_μ, β_μ),

$$U_\mu^{vp}(x, \alpha_\mu, \beta_\mu) = \frac{1}{n+1}\left(\frac{1}{K_\mu}\right)^n \sum_{s=1}^{3} |\ \alpha_\mu u_s(x) + \beta_\mu v_s(x)\ |^{n+1},$$

it turns out to be a strictly convex function of (α_μ, β_μ), and (7.6) can be recast into

$$d_\mu^{xx} = \frac{\partial U_\mu^{vp}}{\partial \alpha_\mu} \quad \text{and} \quad d_\mu^{xy} = \frac{\partial U_\mu^{vp}}{\partial \beta_\mu}.$$

Let $W_\mu^{vp}(x, d_\mu^{xx}, d_\mu^{xy})$ be the Legendre transform of U_μ^{vp} with respect to (α_μ, β_μ). As U_μ^{vp} is a homogeneous function of degree $n+1$ of the pair (α_μ, β_μ), the potential W_μ^{vp} is a homogeneous function of degree $1 + 1/n$ of the pair (d_μ^{xx}, d_μ^{xy}).

We now turn to the derivation of an effective model. Following the general procedure recalled in Section 7.3.1, we define the effective potential W_M^{vp} by

$$W_M^{vp}(d_M^{xx}, d_M^{xy}) = \inf\left\{ \langle\, W_\mu^{vp}(x, d_\mu^{xx}, d_\mu^{xy})\,\rangle,\ (d_\mu^{xx}, d_\mu^{xy}) \in K(d_M^{xx}, d_M^{xy}) \right\},$$

where $K(d_M^{xx}, d_M^{xy})$ is defined by

$$K(d_M^{xx}, d_M^{xy}) = \left\{ \begin{array}{c} (d_\mu^{xx}(x), d_\mu^{xy}(x));\ \exists\, u_\mu(x) \text{ such that } u_\mu(x) = \gamma(d_M^{xx}, d_M^{xy}) \cdot x \\ \text{on } \partial\Omega \text{ and } \dfrac{1}{2}\left(\nabla u_\mu + {}^t\nabla u_\mu\right) = \gamma(d_\mu^{xx}(x), d_\mu^{xy}(x)) \text{ in } \Omega \end{array} \right\},$$

the function γ being defined by

$$\gamma : (u, v) \in \mathbb{R}^2 \mapsto \begin{pmatrix} u & v & 0 \\ v & -u & 0 \\ 0 & 0 & 0 \end{pmatrix} \in \mathcal{M}_3(\mathbb{R}).$$

Just as (7.12) holds, it holds that

$$\alpha_M \;=\; \sigma_M^{xx} - \sigma_M^{yy} = \frac{\partial W_M^{vp}}{\partial d_M^{xx}}, \qquad \beta_M \;=\; 2\,\sigma_M^{xy} = \frac{\partial W_M^{vp}}{\partial d_M^{xy}}. \tag{7.14}$$

The macroscopic potential W_M^{vp} is a homogeneous function of degree $1 + 1/n$ of (d_M^{xx}, d_M^{xy}). To use this fact, we need to change variables : instead of working with the cartesian variables d_M^{xx} and d_M^{xy}, let us work with the polar coordinates associated to them, the radius

$$R_M = \sqrt{(d_M^{xx})^2 + (d_M^{xy})^2}$$

and the angle θ_M. These variables present the advantage that R_M and θ_M are respectively homogeneous functions of degree 1 and 0 of (d_M^{xx}, d_M^{xy}). So W_M^{vp} reads

$$W_M^{vp}(R_M, \theta_M) = R_M^{1+1/n}\, C(\theta_M),$$

where C is an unknown function.

As this point, we introduce the following simplification. Considering that, first, all grain orientations occur with equal probability, and second, that the geometry of the grains and of the polycrystal is isotropic, we postulate, without any rigorous justification of this fact, that the response of the polycrystal is isotropic, at least when the number of grains is large enough. We therefore simplify the previous expression of W_M^{vp}, setting $C(\theta_M)$ as an (unknown) constant C, for θ_M is an anisotropic variable whereas R_M is an isotropic variable.

Let us define

$$J(\sigma_M) = \sqrt{\frac{3}{2}}\,\sqrt{(\tilde\sigma_M^{xx})^2 + (\tilde\sigma_M^{yy})^2 - \frac{1}{2}\,(\tilde\sigma_M^{zz})^2 + 2\,(\sigma_M^{xy})^2}, \tag{7.15}$$

where $\tilde\sigma_M = \sigma_M - \left(\frac{1}{3}\,\mathrm{tr}\,\sigma_M\right) 1$ is the *deviatoric part* of σ_M. Then equations (7.14) can be written as

$$\dot\varepsilon_M^{vp} = \left(\frac{J(\sigma_M)}{K_M}\right)^n \frac{\partial J}{\partial \sigma_M}, \tag{7.16}$$

where K_M is an unknown parameter (playing the role of the constant C used above) that we will determine by numerical computations in Section 7.4.

7.3.2.2 Postulated macroscopic model for the polycrystal

In the previous part, we have made use of the classical homogenization procedure to obtain separately an elastic effective potential and a viscoplastic effective potential. We postulate, again without any rigorous justification of this fact, that the effective constitutive law for the polycrystal is the sum of the elastic effective term with the viscoplastic effective term. So the effective constitutive law that we use is

$$\dot\varepsilon_M(x,t) = \Lambda : \dot\sigma_M(x,t) + \left(\frac{J(\sigma_M(x,t))}{K_M}\right)^n \frac{\partial J}{\partial \sigma_M}, \tag{7.17}$$

where J is defined by (7.15). Solving the effective model consists in searching for the displacement field $u_M(x,t)$ solution to the equilibrium equation

$$\forall x \in \Omega,\ \forall t \in [0,T], \qquad \operatorname{div} \sigma_M(x,t) = 0, \tag{7.18}$$

along with constitutive law (7.17), compatibility equation (7.2) and convenient initial and boundary conditions.

The whole microscopic constitutive law involves an ODE, and the procedures detailed in Section 7.3.1 do not apply in this case. With the numerical tests described in the following, we check whether this approximation may be sensible.

7.4 Numerical results

In the previous section, working with the deformation potential, we have found an effective model for the polycrystal, *up to the knowledge of the constant* K_M (see (7.17)). In order to determine a value for K_M, we use numerical computations on different polycrystals [155, 156], with several linear displacement boundary conditions. In the following, we check that there exists a single value for K_M such that macroscopic computations agree with microscopic computations for all test problems (that is, macroscopic tensors are equal to the average of microscopic tensors over the polycrystal Ω).

One can also work with the microscopic stress potential to obtain an effective stress potential. One finds the same result as (7.16), with a *a priori* different constant K_M^s. To numerically determine a value for K_M^s, one would follow the same procedure as before, except that one would work with linear surface force boundary conditions. If the value found for K_M^s is the same as the value found for K_M (with linear displacement boundary conditions), then the effective stress potential is the Legendre transform of the effective deformation potential, and the polycrystal actually obeys an effective stress-strain relation (see Section 7.3.1). We did not make this kind of test, since, when one uses surface force boundary conditions, the displacement at equilibrium is only determined up to a rigid body motion.

Finally, a third test is possible : one can use mixed boundary conditions (we impose displacement on some part on the boundary and surface force elsewhere). Results of this kind of test are given in the following. The polycrystal actually obeys an effective stress-strain relation if the value previously found for K_M (using linear displacement boundary conditions) is also valid with these mixed boundary conditions.

We have performed numerical tests with three different polycrystals, one of 30 grains (first with a coarse mesh : 5 to 15 finite elements per grain ; then with a finer mesh : finite element edges two times smaller), and two of 110 grains (the same grain geometry, but with two different orientation samples). We work in generalized plane strain in direction z, that is to say we just simulate a 2D layer of the polycrystal of side surface $\mathcal{S}$, with 3D displacement, strain and stress tensor fields. Shears ε_{xz}

and ε_{yz} are equal to zero, and ε_{zz} is uniform on the whole layer. The value of ε_{zz} is obtained by assuming that the resulting force normal to the layer is zero.

On the side surface $\mathcal{S}$, we choose several different boundary conditions : linear displacement boundary conditions (*tension compression*, thus a strain denoted by a superscript TC ; *shear*, a strain denoted S ; *tension compression shear*, a strain denoted TCS), and also *mixed boundary conditions*, letting two opposite faces force free, imposing zero normal displacement on one face, and imposing a uniform tensile displacement rate on the last face (test denoted T). For displacement boundary conditions, the strain tensors are

$$\varepsilon_M^{TC}(t) = \begin{pmatrix} \alpha\, t & 0 & 0 \\ 0 & -\alpha\, t & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad \varepsilon_M^{S}(t) = \begin{pmatrix} 0 & \alpha\, t & 0 \\ \alpha\, t & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

and

$$\varepsilon_M^{TCS}(t) = \begin{pmatrix} \alpha_1\, t & \alpha_2\, t & 0 \\ \alpha_2\, t & -\alpha_1\, t & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

For brevity, we only detail here one test case, namely that of a polycrystal subjected to shear load. The averaged microscopic strain tensor and the macroscopic strain tensor increase linearly as time increases. One can see on Fig. 7.2 the averaged microscopic stress $\langle\, \sigma_\mu(x,t)\,\rangle$ as a function of time (we have $\langle\, \sigma_\mu^{yy}\,\rangle = -\langle\, \sigma_\mu^{xx}\,\rangle$ and $\sigma_\mu^{zx}(x,t) = \sigma_\mu^{zy}(x,t) = \sigma_\mu^{zz}(x,t) = 0$), and the macroscopic stress $\sigma_M(t)$, which is uniform in this case. We make the assumption that, in the long-time limit, the stress tensors $\sigma_\mu(x,t)$ and $\sigma_M(x,t)$ converge to a limit, which thus corresponds to the stationary regime of (7.7) and (7.17). One can check that the limit $\lim_{t\to\infty} \sigma_M^{xy}(t)$ depends on K_M (for this shear load test, an analytical expression can be found). We choose K_M so that

$$\lim_{t\to\infty} \sigma_M^{xy}(t) = \lim_{t\to\infty} \langle\, \sigma_\mu^{xy}(x,t)\,\rangle,$$

which leads in this case to the numerical value $K_M = 347$ MPa. The previous equation enforces that, in the long-time limit, the effective law is consistent with the microscopic law. The macroscopic stress displayed on Fig. 7.2 has been computed using this value of K_M. We also notice that $\sigma_M^{xx} = 0$, as expected. On the other hand, $\langle\, \sigma_\mu^{xx}(x,t)\,\rangle$ is not zero, however up to a small error.

For the other test problems, the situation is the same as the one we describe here. It is possible to find of value for K_M by adjusting the largest components of stress and strain tensors (in the limit $t \to \infty$), and there is a small error on some components (xx and yy in shear load, xy in tension compression load). The values found for K_M are given in Tab. 7.1. We notice that, up to a 0.4 % error, the value depends neither on the type of boundary conditions, on the number of grains, on the mesh size nor on the orientation distribution sample. Thus the polycrystal obeys an effective constitutive law with $K_M = 346$ MPa.

In order to measure the error of the small components of the tensors with respect to the average value, we define some empiric estimators :

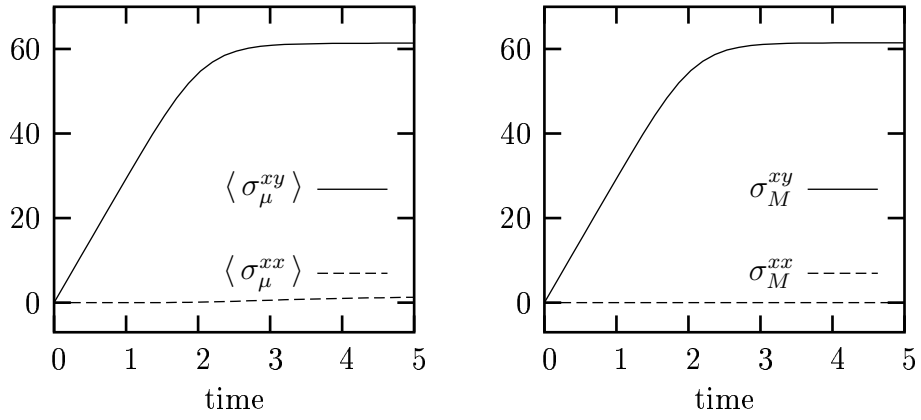

FIG. 7.2 – Shear load on the 30 grain polycrystal : averaged microscopic stress (left), macroscopic stress (right).

	30 grains Coarse mesh	30 grains Fine mesh	110 grains Sample 1	110 grains Sample 2
T	345.6		347.1	
T C	345.6	345.6	347.1	345.3
S	347.25	347.25	344.6	344.6
TCS	346.1	346.0	346.62	345.4

TAB. 7.1 – Values of K_M for different polycrystals with different loadings (the indicated value is the average on different boundary condition values).

- for mixed boundary conditions, $\lim_{t\to\infty} \left(\langle \varepsilon_\mu^{xy} \rangle / \langle \varepsilon_\mu^{yy} \rangle \right)$;
- $\lim_{t\to\infty} \left(\langle \sigma_\mu^{xx} \rangle / \langle \sigma_\mu^{xy} \rangle \right)$ for shear load ;
- for tension-compression load, $\lim_{t\to\infty} \left(\langle \sigma_\mu^{xy} \rangle / \langle \sigma_\mu^{yy} \rangle \right)$;
- $\lim_{t\to\infty} \left(\langle \sigma_\mu^{xx} \pm \sigma_\mu^{xy} \rangle / \sigma_M^{xy} \right)$ for tension-compression-shear load, boundary conditions being so that $\lim_{t\to\infty} \sigma_M^{xx} \pm \sigma_M^{xy} = 0$;

The values found for these estimators are given in Tab. 7.2. One can notice that all errors are small (less than 3%), so the effective law is a good approximation of the microscopic model in most of the situations studied.

It is also interesting to compute averages on grains of stress or strain tensors, and not on the whole polycrystal. We want to know whether these averages are similar from one grain to another one, or very different. Let us focus on the tension-compression-shear load. At each time step, we compute, for each grain, the average over the grain of $(\varepsilon_\mu^{vp})^{yy}$ and of $\tilde{\sigma}_\mu^{yy}$ ($\tilde{\sigma}$ is the deviatoric part of σ). We work with the viscoplastic strain tensor and the deviatoric stress tensor since these are the natural variables for the viscoplastic term of the constitutive law. Results are given in Figs.

	30 grains Coarse mesh	30 grains Fine mesh	110 grains Sample 1	110 grains Sample 2
T	2%		0.9%	
T C	2.5%	2.5%	0.8%	1.2%
S	2.9%	2.9%	1.1%	1%
TCS	0.3%	0.3%	1.4%	1.3%

TAB. 7.2 – Values of error estimators for different polycrystals with different loadings.

7.3 and 7.4. At the beginning, the averages for all grains are the same : even without any yield stress, the viscous flow can be neglected, and due to the uniform elasticity assumption, all grains give the same result. As the viscoplastic term increases, grain responses become heterogeneous. According to classical results in elasto-plasticity [158], the first set of points on Fig. 7.4, corresponding to a very low macroscopic viscoplastic strain (around 0.0002), are aligned along a line, the slope of which is not far from the shear modulus. Like for self-consistent approaches, the slope decreases during the loading (see the three sets of points for a macroscopic viscoplastic strain of 0.0006, 0.0010 and 0.0014), but one can also observe an additional heterogeneity. This kind of curves can be used to calibrate phenomenological models with uniform stress and strain in each phase.

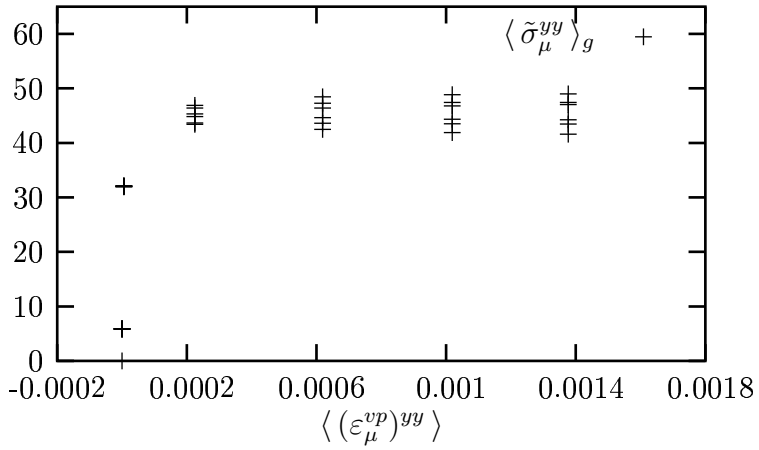

FIG. 7.3 – Evolution of $\langle\, \tilde{\sigma}_\mu^{yy} \,\rangle_g$ as a function of $\langle\, (\varepsilon_\mu^{vp})^{yy} \,\rangle$ for the 30 grain polycrystal, tension-compression-shear load ($\langle\, \cdot \,\rangle_g$ is the average over the grain).

So far, we have just compared responses in the limit $t \to \infty$ (in this limit, the elastic part of the constitutive law cancells). We may also compare responses during the whole load process, to check whether microscopic and effective laws agree only in the viscoplastic limit or also when elastic and viscoplastic terms are of the same

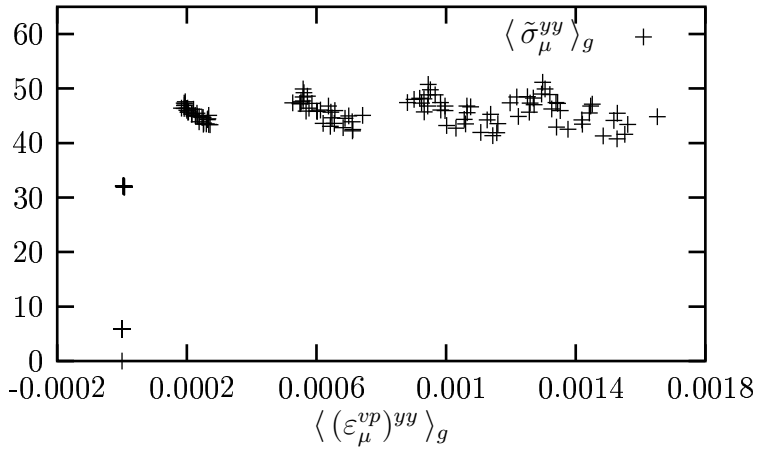

FIG. 7.4 – Evolution of $\langle\, \tilde{\sigma}_\mu^{yy} \,\rangle_g$ as a function of $\langle\, (\varepsilon_\mu^{vp})^{yy} \,\rangle_g$ for the 30 grain polycrystal, tension-compression-shear load ($\langle\, \cdot \,\rangle_g$ is the average over the grain).

order of magnitude. We make such a comparison in Fig. 7.5. For the other test problems, the situation is alike : the effective law is in good agreement with the microscopic law (the difference is smaller than 1%).

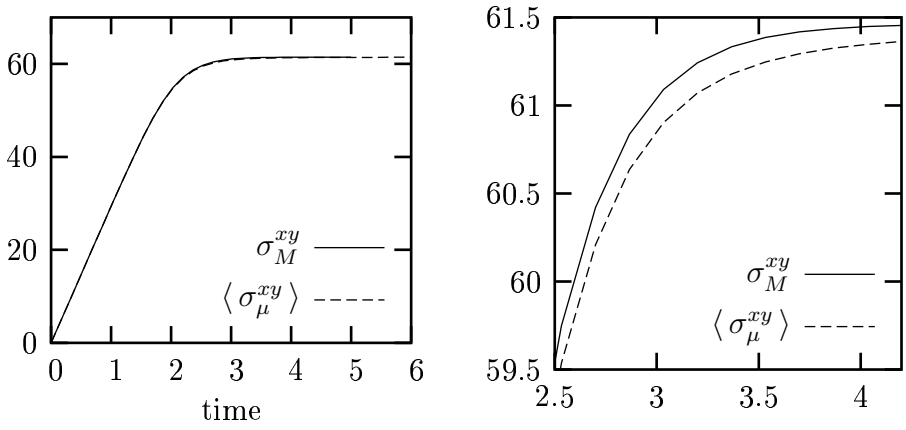

FIG. 7.5 – Transient response of the 30 grain polycrystal, shear load. The effective law is in good agreement with the microscopic law (left). On the right-hand side, a zoom on the region where there are some differences.

This numerical result is very surprising. Starting from a microscopic constitutive law which is time-dependent and involves two potentials, we split it into two terms. We apply separately on each of them a procedure which is based on stationary calculus of variations. We fit K_M on the long-time limit of the system, which corresponds to the viscoplastic regime. The numerical result is that the effective law is in agreement with the microscopic one both in stationary and transient regime! We acknowledge the fact that there is no rigorous reason for this success : we just

observe that the two laws are consistent.

7.5 Conclusions

We have dealt in this article with a simple model of a 2D heterogeneous elasto-viscoplastic polycrystal, for which no theoretical results on the effective law are available. We have succeeded in numerically identifying an effective law. We observe that this effective law is consistent with the microscopic law in both the stationary and transient regime, although it has been obtained by a homogenization procedure designed for stationary problems. We are unfortunately unable to provide any explanation for this fact but are currently working in that direction.

Acknowledgments

I would like to thank Claude Le Bris for many insights incorporated in this article and for his very careful reading of the manuscript, as well as Georges Cailletaud and Renaud Masson for having introduced me to the field and for their very helpful suggestions.

Bibliographie

Références en dynamique moléculaire

[1] Des codes de dynamique moléculaire peuvent être trouvés sur le site `http ://molsim.chem.uva.nl/frenkel_smit`

Ouvrages généraux

[2] M.P. Allen, D.J. Tildesley, *Computer simulation of liquids*, Oxford Science Publications, 1987.

[3] V.I. Arnold, *Mathematical Methods of Classical Mechanics*, Springer-Verlag, 1989.

[4] R. Balian, *From microphysics to macrophysics ; methods and applications of statistical physics*, Springer, 1991.

[5] Ph. Chartier, E. Faou, *Intégration symplectique des systèmes Hamiltoniens intégrables : comportement en temps long*, cours de niveau maîtrise disponible sur le web à l'adresse
`http ://www.irisa.fr/aladin/perso/chartier/`

[6] M.P. Do Carmo, *Riemannian Geometry*, Series *Mathematics : Theory and Applications*, Birkhäuser, Boston, 1992.

[7] D. Frenkel, B. Smit, *Understanding Molecular Simulation, from algorithms to applications*, 2nd ed., Academic Press, 2002.

[8] E. Hairer, *Numerical Geometric Integration*, cours de niveau maîtrise disponible sur le web à l'adresse
`http ://www.unige.ch/math/folks/hairer/polycop.html`

[9] E. Hairer, Ch. Lubich, G. Wanner, *Geometric numerical integration, Structure-preserving algorithms for ordinary differential equations*, Springer Series in Computational Mathematics, Vol. 31, Springer-Verlag, Berlin, 2002.

[10] B. Lapeyre, E. Pardoux, R. Sentis, *Introduction aux méthodes de Monte Carlo*, Collection Mathématiques et Applications, Springer, 1997.

[11] A. Papoulis, *Signal Analysis*, *Electrical and Electronic Engineering Series*, McGraw-Hill, Singapore, 1984.

[12] E. Ramis, C. Deschamps, J. Odoux, *Special course on mathematics : Algebra*, Masson, Paris, 1997.

[13] D.C. Rapaport, *The art of molecular dynamics simulation*, Cambridge University Press, 1995.

[14] R.Y. Rubinstein, *Simulation and the Monte Carlo method*, Wiley, 1981.

[15] J.M. Sanz-Serna, M.P. Calvo, *Numerical Hamiltonian Problems*, Applied Mathematics and Mathematical Computation 7, Chapman & Hall, London, 1994.

[16] T. Schlick, *Molecular Modeling and Simulation, An interdisciplinary guide*, Springer, 2000.

[17] G.F. Simmons, *Topology and Modern Analysis*, McGraw-Hill, 1963.

[18] Y.G. Sinai, *Dynamical systems*, Advanced Series in Nonlinear Dynamics, Vol. 1, World Scientific, 1991.

Articles généraux de mathématiques

[19] A.L. Araujo, A. Murua, J.M. Sanz-Serna, *Symplectic methods based on decompositions*, SIAM J. Numer. Anal. **34**, 5 (1997), pp. 1926-1947.

[20] V.I. Arnold, *Small denominators and problems of stability of motion in classical and celestial mechanics*, Russian Math. Surveys **18** (1963), pp. 85-191.

[21] R.P.K. Chan, A. Murua, *Extrapolation of symplectic methods for Hamiltonian problems*, Appl. Numer. Math. **34**, 2-3 (2000), pp. 189-205.

[22] K. Feng, Z. Shang, *Volume preserving algorithms for source-free dynamical systems*, Numer. Math. **71** (1995), pp. 451-463.

[23] Z. Ge, J.E. Marsden, *Lie-Poisson Hamilton-Jacobi Theory and Lie-Poisson Integrators*, Phys. Lett. A **133** (1988), pp. 134-139.

[24] A.N. Kolmogorov, *On conservation of conditionally periodic motions under small perturbations of the Hamiltonian*, Dokl. Akad. Nauk SSSR **98** (1954), pp. 527-530.

[25] A.N. Kolmogorov, *General theory of dynamical systems and classical mechanics*, Proc. Int. Congr. Math. Amsterdam **1** (1954), pp. 315-333.

[26] W. Krauth, *Les sphères dures en physique statistique*, Pour la science, Octobre / Décembre 2003, pp. 10-15.

[27] C. Le Bris, *Computational chemistry from the perspective of numerical analysis*, Acta Numerica, 2005.

[28] J. Moser, *On invariant curves of area-preserving mappings of an annulus*, Nachr. Akad. Wiss. Göttingen, II. Math.-Phys. K1. 1962, pp. 1-20.

[29] N.N. Nekhoroshev, *An exponential estimate of the time of stability of nearly-integrable Hamiltonian systems*, Russian Math. Surveys **32** (1977), pp. 1-65.

[30] Z. Shang, *KAM theorem of symplectic algorithms for Hamiltonian systems*, Numer. Math. **83** (1999), pp. 477-496.

[31] Z. Shang, *Resonant and Diophantine step sizes in computing invariant tori of Hamiltonian systems*, Nonlinearity **13** (2000), pp. 299-308.

[32] N. Simanyi, *Proof of the Boltzmann-Sinai Ergodic Hypothesis for Typical Hard Disk Systems*, Inventiones Mathematicae **154**, 1 (2003), pp. 123-178.

[33] N. Simanyi, *Proof of the Ergodic Hypothesis for Typical Hard Ball Systems*, Annales Henri Poincaré **5** (2004), pp. 1-31.

[34] Y.G. Sinai, *Dynamical systems with elastic reflections : Ergodic properties of dispersing billiards*, Russian Math. Surveys **25**, 2 (1970), pp. 137-189.

[35] R.D. Skeel, K. Srinivas, *Nonlinear stability analysis of area-preserving integrators*, SIAM J. Num. Ana. **38**, 1 (2000), pp. 129-148.

[36] R.D. Skeel, G. Zhang, T. Schlick, *A family of symplectic integrators : stability, accuracy and molecular dynamics applications*, SIAM J. Sci. Comput. **18**, 1 (1997), pp. 203-222.

[37] M. Suzuki, *General theory of fractal path integrals with applications to many-body theories and statistical physics*, J. Math. Phys. **32**, 2 (1991), pp. 400-407.

[38] M. Suzuki, K. Umeno, *Higher-order decomposition theory of exponential operators and its applications to QMC and nonlinear dynamics*, in Computer Simulation Studies in Condensed-Matter Physics VI, Landau, Mon, Schüttler (eds.), Springer Proceedings in Physics 76 (1993), pp. 74-86.

Articles généraux de physique / chimie

[39] B.J. Alder, T.E. Wainwright, *Phase transition of a hard sphere system*, J. Chem. Phys. **27** (1957), pp. 1208-1209.

[40] E. Barth, K. Kuczera, B. Leimkuhler, R.D. Skeel, *Algorithms for Constrained Molecular Dynamics*, J. Comput. Chem. **16** (1995), pp. 1192-1209.

[41] C. Chipot, B. Maigret, A. Pohorille, *Early events in the folding of an amphipathic peptide. A multi-nanosecond molecular dynamics study*, Proteins : Structure, Function and Genetics **36** (1999), pp. 383-399.

[42] C. Chipot, A. Pohorille, *Conformational equilibria of terminally blocked single amino acids at the water-hexane interface. A molecular dynamics study*, J. Phys. Chem. B **102** (1998), pp. 281-290.

[43] J. Kohanoff, *Phonon spectra from short non-thermally equilibrated molecular dynamics simulations*, Comput. Mater. Sciences **2** (1994), pp. 221-232.

[44] B.J. Leimkuhler, R.D. Skeel, *Symplectic numerical integrators in constrained Hamiltonian systems*, J. Comput. Phys. **112** (1994), pp. 117-125.

[45] D. Levesque, L. Verlet, *Computer "experiments" on classical fluids. III. Time dependent self-correlation functions*, Phys. Rev. A **2** (1970), pp. 2514-2528.

[46] D. Levesque, L. Verlet, J. Kürkijarvi, *Computer "experiments" on classical fluids. IV. Transport properties and time-correlation functions of the Lennard-Jones liquid near its triple point*, Phys. Rev. A **7** (1973), pp. 1690-1700.

[47] M.G. Martin, J.I. Siepmann, *Transferable potentials for phase equilibria. I. United-Atom description of n-alkanes*, J. Phys. Chem. B **102** (1998), pp. 2569-2577.

[48] G.J. Martyna, M.E. Tuckerman, *Symplectic reversible integrators : Predictor-corrector methods*, J. Chem. Phys. **102**, 20 (1995), pp. 8071-8077.

[49] N. Matubayasi, M. Nakahara, *Reversible molecular dynamics for rigid bodies and hybrid Monte Carlo*, J. Chem. Phys. **110**, 7 (1999), pp. 3291-3301.

[50] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, E. Teller, *Equation of state calculations by fast computing machines*, J. Chem. Phys. **21** (1953), pp. 1087-1092.

[51] M. Parrinello, A. Rahman, *Strain fluctuations and elastic constants*, J. Chem. Phys. **76**, 5 (1982), pp. 2662-2666.

[52] G.R.W. Quispel, C.P. Dyt, *Volume preserving integrators have linear error growth*, Phys. Lett. A **242** (1998), pp. 25-30.

[53] T. Schlick, *Computational Molecular Biophysics Today*, J. Comput. Phys. **151** (1999), pp. 1-8.

[54] T. Schlick, E. Barth, M. Mandziuk, *Biomolecular dynamics at long time steps*, Annu. Rev. Biophys. Biomol. Struct. **26** (1997), pp. 181-222.

[55] T. Schlick, R.D. Skeel, A.T. Brunger, L.V. Kalé, J.A. Board, J. Hermans, K. Schulten, *Algorithmic challenges in computational molecular biophysics*, J. Comput. Phys. **151** (1999), pp. 9-48.

[56] M. Tarek, B. Maigret, C. Chipot, *Oriented self-assembly of cyclic peptides in lipid bilayers. A multi-nanosecond molecular dynamics investigation*, Biophys. J. **85** (2003), pp. 2287-2298.

[57] M.E. Tuckerman, G.J. Martyna, *Understanding Molecular Dynamics : Techniques and Applications*, J. Phys. Chem. B **104** (2000), pp. 159-178.

[58] L. Verlet, *Computer "experiments" on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules*, Phys. Rev. **159** (1967), pp. 98-103.

[59] L. Verlet, *Computer "experiments" on classical fluids. II. Equilibrium correlation functions*, Phys. Rev. **165** (1968), pp. 201-214.

[60] H. Yoshida, *Construction of higher order symplectic integrators*, Phys. Lett. A **150** (1990), pp. 262-268.

Dynamique à température ou pression "constantes"

[61] S.D. Bond, B.J. Leimkuhler, B.B. Laird, *The Nosé-Poincaré method for constant temperature molecular dynamics*, J. Comput. Phys. **151** (1999), pp. 114-134.

[62] W.G. Hoover, *Canonical dynamics : Equilibrium phase-space distributions*, Phys. Rev. A **31**, 3 (1985), pp. 1695-1697.

[63] S. Jang, G. Voth, *Simple reversible molecular dynamics algorithms for Nosé-Hoover chain dynamics*, J. Chem. Phys. **107**, 22 (1997), pp. 9514-9526.

[64] S. Jang, G. Voth, *Response to Comment on "Simple reversible molecular dynamics algorithms for Nosé-Hoover chain dynamics"*, J. Chem. Phys. **110**, 7 (1999), pp. 3626-3628.

[65] Y. Liu, M.E. Tuckerman, *Generalized Gaussian moment thermostatting : A new continuous dynamical approach to the canonical ensemble*, J. Chem. Phys. **112** (2000), pp. 1685-1700.

[66] R. Martonak, C. Molteni, M. Parrinello, *A new constant-pressure ab initio/classical molecular dynamics method : simulation of pressure-induced amorphization in a $Si_{35}H_{36}$ cluster*, Comput. Mat. Science **20** (2001), pp. 293-299.

[67] G.J. Martyna, M.L. Klein, M.E. Tuckerman, *Nosé-Hoover chains : The canonical ensemble via continuous dynamics*, J. Chem. Phys. **97**, 4 (1992), pp. 2635-2643.

[68] G.J. Martyna, M.E. Tuckerman, D.J. Tobias, M.L. Klein, *Explicit reversible integrators for extended systems dynamics*, Mol. Phys. **87**, 5 (1996), pp. 1117-1157.

[69] S. Nosé, *A unified formulation of the constant temperature molecular dynamics method*, J. Chem. Phys. **81**, 1 (1985), pp. 511-519.

[70] M.E. Tuckerman, G.J. Martyna, *Comment on "Simple reversible molecular dynamics algorithms for Nosé-Hoover chain dynamics"*, J. Chem. Phys. **110**, 7 (1999), pp. 3623-3625.

[71] M.E. Tuckerman, C.J. Mundy, G.J. Martyna, *On the classical statistical mechanics of non-Hamiltonian systems*, Europhys. Lett. **45**, 2 (1999), pp. 149-155.

Réduction du coût calcul des forces et méthodes multipas

[72] E. Barth, B.J. Leimkuhler, S. Reich, *A time-reversible variable step size integrator for constrained dynamics*, SIAM J. Sci. Comput. **21**, 3 (1999), pp. 1027-1044.

[73] E. Barth, T. Schlick, *Extrapolation versus impulse in multiple-timestepping schemes. II. Linear analysis and applications to Newtonian and Langevin dynamics*, J. Chem. Phys. **109**, 5 (1998), pp. 1633-1642.

[74] P.F. Batcho, D.A. Case, T. Schlick, *Optimized particle-mesh Ewald / multiple-time step integration for molecular dynamics simulations*, J. Chem. Phys. **115**, 9 (2001), pp. 4003-4018.

[75] P. Batcho, T. Schlick, *New splitting formulations for lattice summations*, J. Chem. Phys. **115**, 18 (2001), pp. 8312-8326.

[76] J.J. Biesiadecki, R.D. Skeel, *Dangers of multiple time step methods*, J. Comput. Phys. **109** (1993), pp. 318-328.

[77] T.C. Bishop, R.D. Skeel, K. Schulten, *Difficulties with multiple time stepping and Fast Multipole Algorithm in molecular dynamics*, J. Comp. Chem. **18**, 14 (1997), pp. 1785-1791.

[78] E. Darrigrand, *Couplage Méthodes Multipôles Rapides et Discrétisation Micro-locale pour les Equations Intégrales de l'Electromagnétisme*, thèse du CEA / CESTA, 2002.

[79] E. Darve, *Méthodes multipôles rapides : résolution des équations de Maxwell par formulations intégrales*, thèse de l'Université Paris 6, 1999.

[80] S.W. de Leeuw, J.W. Perram, E.R. Smith, *Simulation of electrostatic systems in periodic boundary conditions. I. Lattice sums and dielectric constant*, Proc. Roy. Soc. London A **373** (1980), pp. 27-56.

[81] U. Essmann, L. Perera, M.L. Berkowitz, T. Darden, H. Lee, L.G. Pedersen, *A smooth particle mesh Ewald method*, J. Chem. Phys. **103**, 19 (1995), pp. 8577-8593.

[82] F. Figueirido, R. Zhou, R. Levy, B.J. Berne, *Large scale simulation of macromolecules in solutions : Combining the periodic FMM with multiple time scales*, J. Chem. Phys. **106**, 23 (1997), pp. 9835-9849.

[83] B. Garcia-Archilla, J.M. Sanz-Serna, R.D. Skeel, *Long time step methods for oscillatory differential equations*, SIAM J. Sci. Comput. **20**, 3 (1998), pp. 930-963.

[84] D.D. Humphreys, R.A. Friesner, B.J. Berne, *A Multiple-Time-Step Molecular Dynamics Algorithm for Macromolecules*, J. Phys. Chem. **98** (1994), pp. 6885-6892.

[85] J.A. Izaguirre, S. Reich, R.D. Skeel, *Longer time steps for Molecular Dynamics*, J. Chem. Phys **110**, 20 (1999), pp. 9853-9864.

[86] T.R. Littell, R.D. Skeel, M. Zhang, *Error analysis of symplectic Multiple Time Stepping*, SIAM J. Num. Ana. **34**, 5 (1997), pp. 1792-1807.

[87] B.A. Luty, I.G. Tironi, W.F. van Gunsteren, *Lattice-sum methods for calculating electrostatic interactions in molecular simulations*, J. Chem. Phys. **103**, 8 (1995), pp. 3014-3021.

[88] M. Mandziuk, T. Schlick, *Resonance in the dynamics of chemical systems simulated by the implicit midpoint scheme*, Chem. Phys. Lett. **237** (1995), pp. 525-535.

[89] H.G. Petersen, *Accuracy and efficiency of the particle mesh Ewald method*, J. Chem. Phys. **103**, 9 (1995), pp. 3668-3679.

[90] P. Procacci, B.J. Berne, *Computer simulation of solid C_{60} using multiple time steps algorithms*, J. Chem. Phys. **101**, 3 (1994), pp. 2421-2431.

[91] P. Procacci, B.J. Berne, *Multiple time scale methods for constant pressure molecular dynamics simulations of molecular systems*, Mol. Phys. **83**, 2 (1994), pp. 255-272.

[92] A. Sandu, T. Schlick, *Masking resonance artifacts in force-splitting methods for biomolecular simulations by extrapolative Langevin dynamics*, J. Comput. Phys. **151** (1999), pp. 74-113.

[93] T. Schlick, M. Mandziuk, R.D. Skeel, K. Srinivas, *Nonlinear resonance artifacts in molecular dynamics*, J. Comp. Phys. **139** (1998), pp. 1-29.

[94] R.D. Skeel, J.J. Biesiadecki, *Symplectic integration with variable stepsize*, Ann. Numer. Math. **1** (1994), pp. 1-9.

[95] S.J. Stuart, R. Zhou, B.J. Berne, *Molecular dynamics with multiple time scales : The selection of efficient reference system propagators*, J. Chem. Phys. **105**, 4 (1996), pp. 1426-1436.

[96] S. Toxvaerd, *Comment on "Reversible multiple time scale molecular dynamics"*, J. Chem. Phys. **99**, 3 (1993), p. 2277.

[97] M.E. Tuckerman, B.J. Berne, *Stochastic molecular dynamics in systems with multiple time scales and memory friction*, J. Chem. Phys. **95**, 6 (1991), pp. 4389-4396.

[98] M.E. Tuckerman, B.J. Berne, *Molecular dynamics in systems with multiple time scales : Systems with stiff and soft degrees of freedom and with short and long range forces*, J. Chem. Phys. **95**, 11 (1991), pp. 8362-8364.

[99] M.E. Tuckerman, B.J. Berne, G.J. Martyna, *Molecular dynamics algorithm for multiple time scales : Systems with long range forces*, J. Chem. Phys. **94**, 10 (1991), pp. 6811-6815.

[100] M.E. Tuckerman, B.J. Berne, G.J. Martyna, *Reversible multiple time scale molecular dynamics*, J. Chem. Phys. **97**, 3 (1992), pp. 1990-2001.

[101] M.E. Tuckerman, B.J. Berne, G.J. Martyna, *Reply to Comment on : "Reversible multiple time scale molecular dynamics"*, J. Chem. Phys. **99**, 3 (1993), pp. 2278-2279.

[102] M.E. Tuckerman, B.J. Berne, A. Rossi, *Molecular dynamics algorithm for multiple time scales : Systems with disparate masses*, J. Chem. Phys. **94**, 2 (1991), pp. 1465-1469.

[103] M.E. Tuckerman, G.J. Martyna, B.J. Berne, *Molecular dynamics algorithm for condensed systems with multiple time scales*, J. Chem. Phys. **93**, 2 (1990), pp. 1287-1291.

[104] R. Zhou, B.J. Berne, *A new Molecular Dynamics method combining the reference system propagator algorithm with a fast multipole method for simulating proteins and other complex systems*, J. Chem. Phys. **103**, 21 (1995), pp. 9444-9459.

[105] R. Zhou, E. Harder, H. Xu, B.J. Berne, *Efficient multiple time steps method for use with Ewald and particle mesh Ewald for large biomolecular systems*, J. Chem. Phys. **115**, 5 (2001), pp. 2348-2358.

Gagner des ordres de grandeur en temps

[106] W. Cai, V.V. Bulatov, J.F. Justo, A.S. Argon, S. Yip, *Intrinsic mobility of a dissociated dislocation in silicon*, Phys. Rev. Lett. **84**, 15 (2000), pp. 3346-3349.

[107] P. Deuflhard, W. Huisinga, A. Fischer, Ch. Schütte, *Identification of almost invariant aggregates in reversible nearly uncoupled Markov chains*, Lin. Algebra and its Applications **315** (2000), pp. 39-59.

[108] P. Deuflhard, M. Weber, *Robust Perron cluster analysis in conformation dynamics*, Konrad-Zuse-Zentrum technical report 03-19, 2003.

[109] W. E, W. Ren, E. Vanden-Eijnden, *String method for the study of rare events*, Phys. Rev. B **66** (2002), article 052301.

[110] G. Henkelman, H. Jonsson, *Improved tangent estimates in the nudged elastic band method for finding minimum energy paths and saddle points*, J. Chem. Phys. **113**, 22 (2000), pp. 9978-9985.

[111] A. Laio, M. Parrinello, *Escaping free energy minima*, Proc. Natl. Acad. Sci. USA **99** (2002), pp. 12562-12566.

[112] F. Montalenti, M.R. Sorensen, A.F. Voter, *Closing the gap between experiments and theory : crystal growth by temperature accelerated dynamics*, Phys. Rev. Lett. **87**, 12 (2001), article 126101.

[113] W. Ren, *Numerical Methods for the Study of Energy Landscapes and Rare Events*, thèse de l'Université de New-York, 2002.

[114] C.F. Sanz-Navarro, R. Smith, *Numerical calculations using the hypermolecular dynamics simulation method*, Comp. Phys. Comm. **137**, 1 (2001) pp. 206-221.

[115] Ch. Schütte, A. Fischer, W. Huisinga, P. Deuflhard, *A direct approach to conformational dynamics based on hybrid Monte Carlo*, J. Comput. Phys. **151** (1999), pp. 146-168.

[116] Ch. Schütte, W. Huisinga, *Biomolecular conformations can be identified as metastable states of molecular dynamics*, in Handbook of Numerical Analysis, Vol. X, Special volume : Computational chemistry, North-Holland, 2003, pp. 699-744.

[117] M.R. Sorensen, A.F. Voter, *Temperature accelerated dynamics for simulation of infrequent events*, J. Chem. Phys. **112**, 21 (2000), pp. 9599-9606.

[118] A.F. Voter, *A method for accelerating the molecular dynamics simulation of infrequent events*, J. Chem. Phys. **106**, 11 (1997), pp. 4665-4677.

[119] A.F. Voter, *Parallel replica method for dynamics of infrequent events*, Phys. Rev. B **57**, 22 (1998), pp. 13985-13988.

[120] Z. Zhu, M.E. Tuckerman, S.O. Samuelson, G.J. Martyna, *Using novel variable transformations to enhance conformational sampling in molecular dynamics*, Phys. Rev. Lett. **88** (2002), article 100201.

Mise en oeuvre des techniques multipas ou de travail à pression constante en dynamique moléculaire *ab initio*

[121] J. Cao, G.J. Martyna, *Adiabatic path integral molecular dynamics methods. 2. Algorithms*, J. Chem. Phys. **104**, 5 (1996), pp. 2028-2035.

[122] G.J. Martyna, A. Hughes, M.E. Tuckerman, *Molecular dynamics algorithms for path integrals at constant pressure*, J. Chem. Phys. **110**, 7 (1999), pp. 3275-3290.

[123] M.E. Tuckerman, M. Parrinello, *Integrating the Car-Parrinello equations. II. Multiple time scale techniques*, J. Chem. Phys. **101**, 2 (1994), pp. 1316-1329.

Références en mécanique et sur les problèmes multiéchelles

Ouvrages généraux

[124] G. Allaire, *Analyse numérique et optimisation*, Cours de l'Ecole Polytechnique, 2002.

[125] Y. Bamberger, *Mécanique de l'ingénieur*, Hermann, 1981 et 1997.

[126] J. Besson, G. Cailletaud, J.-L. Chaboche, S. Forest, *Mécanique non linéaire des matériaux*, Etudes en mécanique des matériaux et des structures, Hermès, Paris, 2001.

[127] S.C. Brenner, L.R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer, 1991.

[128] "Multiscale Modelling of Materials" symposium proceedings, V.V. Bulatov, T.D. de la Rubia, R. Phillips, E. Kaxiras, N. Ghoniem eds., Materials Research Society, 1999.

[129] M. Chipot, *Elements of nonlinear analysis*, Birkhäuser, 2000.

[130] P.G. Ciarlet, *Mathematical Elasticity, Vol. I : Three-Dimensional Elasticity*, Studies in Mathematics and its Applications, North Holland, 1988.

[131] D. Kinderlehrer, G. Stampacchia, *Introduction to variational inequalities and their applications*, Academic Press, 1980.

[132] B. Larrouturou, P.-L. Lions, *Méthodes mathématiques pour les sciences de l'ingénieur : Optimisation et analyse numérique*, Cours de l'Ecole Polytechnique, 1996.

[133] C. Le Bris, *Systèmes multi-échelles : modélisation et simulation*, Cours de l'Ecole Polytechnique, 2003.

[134] D. Leguillon, E. Sanchez-Palencia, *Computation of singular solutions in elliptic problems and elasticity*, Masson, Paris, 1987.

[135] J.E. Marsden, T.J.R. Hugues, *Mathematical foundations of Elasticity*, Dover, 1994.

[136] A. Quarteroni, A. Valli, *Numerical Approximation of Partial Differential Equations*, Springer, 1997.

[137] J. Salençon, *Mécanique des milieux continus*, Cours de l'Ecole Polytechnique, 1996.

[138] J. Sanchez-Hubert, E. Sanchez-Palencia, *Coques élastiques minces : propriétés asymptotiques*, Masson, Paris, 1997.

[139] C. Truesdell, W. Noll, *The nonlinear field theories of mechanics theory of elasticity*, Handbuch der Physik, III/3, Springer Berlin, 1965, pp. 1-602.

Articles de mathématiques

[140] G. Allaire, *Homogenization and two-scale convergence*, SIAM J. Math. Anal. **23**, 6 (1992), pp. 1482-1518.

[141] S. Bartels, A. Prohl, *Multiscale resolution in the computation of crystalline microstructure*, Numer. Math. **96** (2004), pp. 641-660.

[142] X. Blanc, C. Le Bris, P.-L. Lions, *From molecular models to continuum mechanics*, Arch. Ration. Mech. Anal. **164** (2002), pp. 341-381.

[143] A. Braides, G. Dal Maso, A. Garroni, *Variational Formulation of Softening Phenomena in Fracture Mechanics : the One-Dimensional Case*, Arch. Ration. Mech. Anal. **146** (1999), pp. 23-58.

[144] C. Carstensen, T. Roubicek, *Numerical approximation of young measures in nonconvex variational problems*, Numer. Math. **84** (2000), pp. 395-415.

[145] P.G. Ciarlet, *An $O(h^2)$ method for a non-smooth boundary value problem*, Aequationes Math. **2** (1968), pp. 39-49.

[146] P.G. Ciarlet, *Basic Error Estimates for Elliptic Problems*, in Handbook of Numerical Analysis, Vol. II, P.G. Ciarlet, J.-L. Lions, eds., North-Holland, 1991, pp. 17-351.

[147] T. Hou, X. Wu, *A multiscale finite element method for elliptic problems in composite materials and porous media*, J. Comput. Phys. **134** (1997), pp. 169-189.

[148] P. Le Tallec, *Numerical Methods for nonlinear three-dimensional elasticity*, in Handbook of Numerical Analysis, Vol. III, P.G. Ciarlet, J.-L. Lions, eds., North-Holland, 1994, pp. 465-622.

[149] M. Luskin, *On the computation of crystalline microstructure*, Acta Numerica (1996), pp. 191-257.

[150] P. Marcellini, *Periodic solutions and homogenization of nonlinear variational problems*, Ann. Mat. Pura Appl. **4**, 117 (1978), pp. 139-152.

[151] A.-M. Matache, Ch. Schwab, *Two-scale FEM for Homogenization Problems*, Seminar for Applied Mathematics research report 2001-06, 2001.

[152] E. Sanchez-Palencia, *Passage à la limite de l'élasticité tridimensionnelle à la théorie asymptotique des coques minces*, C. R. Acad. Sci. Paris, Série II, **311** (1990), pp. 909-916.

[153] Ch. Schwab, A.-M. Matache, *Generalized FEM for Homogenization Problems*, Seminar for Applied Mathematics research report 2001-03, 2001.

Articles de mécanique

[154] F. Barbe, L. Decker, D. Jeulin, G. Cailletaud, *Intergranular and intragranular behaviour of polycrystalline aggregates, Part 1 : F.E. model*, Int. J. Plasticity **17** (2001), pp. 513-536.

[155] F. Barbe, S. Forest, G. Cailletaud, *Intergranular and intragranular behaviour of polycrystalline aggregates, Part 2 : Results*, Int. J. Plasticity **17** (2001), pp. 537-563.

[156] J.-L. Chaboche, G. Cailletaud, *Integration methods for complex plastic constitutive equations*, Comput. Methods Appl. Mech. Engrg. **133** (1996), pp. 125-155.

[157] J.D. Eshelby, *The determination of the elastic field of an ellipsoidal inclusion and related problems*, Proc. R. Soc. Lond. A **421** (1957), pp. 376-396.

[158] E. Kröner, *Zur plastischen Verformung des Vielkristalls (On the plastic deformation of polycrystals)*, Acta. Met. **9** (1961), pp. 155-161.

La méthode "QuasiContinuum Method" (cf. la section 1.2.1.1)

[159] J. Knap, M. Ortiz, *An Analysis of the QuasiContinuum Method*, J. Mech. Phys. Solids **49** (2001), pp. 1899-1923.

[160] R.E. Miller, E.B. Tadmor, *The Quasicontinuum Method : Overview, Applications and Current Directions*, J. of Computer-Aided Materials Design **9**, 3 (2002), pp. 203-239.

[161] R. Miller, E.B. Tadmor, R. Phillips, M. Ortiz, *Quasicontinuum simulation of fracture at the atomic scale*, Modelling Simul. Mater. Sci. Eng. **6** (1998), pp. 607-638.

[162] V.B. Shenoy, R. Miller, E.B. Tadmor, R. Phillips, M. Ortiz, *Quasicontinuum Models of Interfacial Structure and Deformation*, Phys. Rev. Lett. **80** (1998), pp. 742-745.

[163] V.B. Shenoy, R. Miller, E.B. Tadmor, D. Rodney, R. Phillips, M. Ortiz, *An adaptative finite element approach to atomic-scale mechanics - the QuasiContinuum Method*, J. Mech. Phys. Solids **47** (1999), pp. 611-642.

[164] E.B. Tadmor, *The quasicontinuum method*, thèse de l'Université de Brown, 1996.

[165] E.B. Tadmor, M. Ortiz, R. Phillips, *QuasiContinuum analysis of defects in solids*, Phil. Mag. A **73** (1996), pp. 1529-1563.

[166] E.B. Tadmor, R. Phillips, *Mixed Atomistic and Continuum Models of Deformation in Solids*, Langmuir **12** (1996), pp. 4529-4534.

[167] E.B. Tadmor, G.S. Smith, N. Bernstein, E. Kaxiras, *Mixed finite element and atomistic formulation for complex crystals*, Phys. Rev. B **59** (1999), pp. 235-245.

Méthode fondée sur la dynamique Hamiltonienne (cf. la section 1.2.1.3)

[168] F.F. Abraham, J.Q. Broughton, N. Bernstein, E. Kaxiras, *Spanning the continuum to quantum length scales in a dynamic simulation of brittle fracture*, Europhys. Lett. **44**, 6 (1998), pp. 783-787.

[169] F.F. Abraham, J.Q. Broughton, N. Bernstein, E. Kaxiras, *Spanning the Length Scales in Dynamic Simulation*, Computers in Physics **12**, 6 (1998), pp. 538-546.

[170] T. Belytschko, S.P. Xiao, *Coupling methods for continuum model with molecular model*, Int. J. for Multiscale Computational Engineering **1**, 1 (2003), pp. 115-126.

[171] J.Q. Broughton, F.F. Abraham, N. Bernstein, E. Kaxiras, *Concurrent coupling of length scales : Methodology and application*, Phys. Rev. B **60** (1999), pp. 2391-2403.

[172] W. E, Z. Huang, *Matching Conditions in Atomistic-Continuum Modeling of Materials*, Phys. Rev. Lett. **87**, 13 (2001), article 135501.

[173] W. E, Z. Huang, *A Dynamic Atomistic-Continuum Method for the Simulation of Crystalline Materials*, J. Comput. Phys. **182**, 1 (2002), pp. 234-261.

[174] E. Lidorikis, M.E. Bachlechner, R.K. Kalia, A. Nakano, P. Vashishta, *Coupling Length Scales for Multiscale Atomistic-Continuum Simulations : Atomistically Induced Stress Distribution in Si/Si_3N_4 Nanopixels*, Phys. Rev. Lett. **87**, 8 (2001), article 086104.

[175] A. Nakano, M.E. Bachlechner, R.K. Kalia, E. Lidorikis, P. Vashishta, G.Z. Voyiadjis, *Multiscale simulation of nanosystems*, Computing in Science & Engineering, July/August 2001, pp. 56-66.

[176] A. Nakano, T.J. Campbell, R.K. Kalia, S. Kodiyalam, S. Ogata, F. Shimojo, X. Su, P. Vashishta, *Scalable multiresolution algorithms for classical and quantum molecular dynamics of nanosystems*, in Handbook of Numerical Analysis, Vol. X, Special volume : Computational chemistry, North-Holland, 2003, pp. 639-666.

[177] A. Nakano, R.K. Kalia, P. Vashishta, T. Campbell, S. Ogata, F. Shimojo, S. Saini, *Scalable Atomistic Simulation Algorithms for Materials Research*, Scientific Programming **10**, 4 (2002), pp. 263-270.

[178] S. Ogata, E. Lidorikis, F. Shimojo, A. Nakano, P. Vashishta, R.K. Kalia, *Hybrid finite-element / molecular-dynamics / electronic density-functional approach to materials simulations on parallel computers*, Comput. Phys. Comm. **138** (2001), pp. 143-154.

[179] R.E. Rudd, J.Q. Broughton, *Coarse-grained molecular dynamics and the atomic limit of finite elements*, Phys. Rev. B **58**, 10 (1998), pp. 5893-5896.

[180] R.E. Rudd, J.Q. Broughton, *Concurrent coupling of length scales in Solid State Systems*, Phys. Stat. Sol. B **217** (2000), pp. 251-291.

[181] F. Shimojo, R.K. Kalia, A. Nakano, P. Vashishta, *Linear-scaling density-functional-theory calculations of electronic structure based on real-space grids : design, analysis, and scalability test of parallel algorithms*, Comput. Phys. Comm. **140** (2001), pp. 303-314.

D'autres méthodes multi-échelles

[182] S. Curtarolo, G. Ceder, *Dynamics of a non homogeneously coarse grained system*, Phys. Rev. Lett. **88** (2002), article 255504.

[183] W. E, P. Ming, private communication.

[184] H. Gao, B. Ji, *Modeling fracture in nano-materials via a virtual internal bond method*, Engineering Fracture Mechanics **70**, 14 (2003), pp. 1777-1791.

[185] N.G. Hadjiconstantinou, *Hybrid Atomistic-Continuum Formulations and the Moving Contact-Line Problem*, J. Comput. Phys. **154** (1999), pp. 245-265.

[186] N.G. Hadjiconstantinou, *Combining Atomistic and Continuum Simulations of Contact-Line Motion*, Phys. Rev. E **59**, 2 (1999), pp. 2475-2478.

[187] B. Ji, H. Gao, *A study of fracture mechanisms in biological nano-composites via the virtual internal bond model*, Materials Science & Engineering A **366** (2004), pp. 96-103.

[188] P.A. Klein, H. Gao, *Study of Crack Dynamics Using the Virtual Internal Bond Method*, in Multiscale Deformation and Fracture in Materials and Structures-

The James R. Rice 60th Anniversary Volume, T.-J. Chuang, J. W. Rudnicki eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 2000, pp. 275-309.

[189] N. Triantafyllidis, S. Bardenhagen, *On higher order gradient continuum theories in 1-D nonlinear elasticity. Derivation from and comparison to the corresponding discrete models*, J. Elasticity **33** (1993), pp. 259-293.

[190] G.J. Wagner, W.K. Liu, *Coupling of atomistic and continuum simulations using a bridging scale decomposition*, J. Comput. Phys. **190** (2003), pp. 249-274.

[191] P. Zhang, P.A. Klein, Y. Huang, H. Gao, P.D. Wu, *Numerical simulation of cohesive fracture by the virtual-internal-bond model*, Computer Modeling in Engineering and Sciences **3** (2002), pp. 263-277.

Homogénéisation en mécanique

[192] M. Bornert, P. Ponte Castañeda, *Second-order estimates of the self-consistent type for viscoplastic polycrystals*, Proc. R. Soc. Lond. A **454** (1998), pp. 3035-3045.

[193] F. Feyel, J.-L. Chaboche, *FE^2 multiscale approach for modelling the elasto-viscoplastic behaviour of long fiber SiC / Ti composite materials*, Comput. Meth. in Appl. Mech. Engng. **183** (2000), pp. 417-455.

[194] R.V. Kohn, T.D. Little, *Some model problems of polycrystal plasticity with deficient basic crystals*, SIAM J. Appl. Math **59**, 1 (1998), pp. 172-197.

[195] V. Kouznetsova, M.G.D. Geers, W.A.M. Brekelmans, *Multiscale constitutive modelling of heterogeneous materials with a gradient-enhanced computational homogenization scheme*, Int. J. Num. Eng. **54**, 8 (2002), pp. 1235-1260.

[196] P. Ladeveze, A. Nouy, *Une stratégie de calcul multiéchelle avec homogénéisation en espace et en temps*, C. R. Acad. Sci. Paris, Mécanique, **330** (2002), pp. 683-689.

[197] S. Moorthy, S. Ghosh, *Adaptivity and convergence in the Voronoi cell finite element model for analyzing heterogeneous materials*, Comp. Meth. Appl. Mech. Engng. **185** (2000), pp. 37-74.

[198] P. Ponte Castañeda, *Exact second-order estimates for the effective mechanical properties of nonlinear composite materials*, J. Mech. Phys. Solids **44**, 6 (1996), pp. 827-862.

[199] P. Ponte Castañeda, P. Suquet, *Nonlinear composites*, Advances in Applied Mechanics **34** (1998), pp. 171-302.

[200] P. Ponte Castañeda, J.R. Willis, *Variational second-order estimates for nonlinear composites*, Proc. R. Soc. Lond. A **455** (1999), pp. 1799-1811.

Méthodes des fonctions de Green (cf. le chapitre 5)

[201] S. Rao, C. Hernandez, J.P. Simmons, T.A. Parthasarathy, C. Woodward, *Green's function boundary conditions in two-dimensional and three-dimensional atomistic simulations of dislocations*, Phil. Mag. A **77**, 1 (1998), pp. 231-256.

[202] S. Rao, T.A. Parthasarathy, C. Woodward, *Atomistic simulation of cross-slip processes in model fcc structures*, Phil. Mag. A **79**, 5 (1999), pp. 1167-1192.

[203] S. Rao, C. Woodward, *Atomistic simulation of* $(a/2)\langle 111\rangle$ *screw dislocations in bcc Mo using a modified generalized pseudopotential theory potential*, Phil. Mag. A **81**, 5 (2001), pp. 1317-1327.

[204] J.E. Sinclair, P.C. Gehlen, R.G. Hoagland, J.P. Hirth, *Flexible boundary conditions and nonlinear geometric effects in atomic dislocation modeling*, J. Appl. Phys. **49**, 7 (1978), pp. 3890-3897.

[205] C. Woodward, S.I. Rao, Ab-initio *simulation of isolated screw dislocations in bcc Mo and Ta*, Phil. Mag. A **81**, 5 (2001), pp. 1305-1316.

Sur le calcul de fractures ou d'instabilités mécaniques (par modèle atomistique ou de continuum)

[206] T. Belytschko, S.P. Xiao, G.C. Schatz, R.S. Ruoff, *Atomistic simulations of nanotube fracture*, Phys Rev B **65**, 25 (2002), article 235430.

[207] D. Leguillon, *Strength or toughness ? A criterion for crack onset at a notch*, Eur. J. Mech A / Solids **21** (2002), pp. 61-72.

[208] M. Marder, *Molecular dynamics of cracks*, Computing in Science & Engineering, Septembre / Octobre 1999, pp. 48-55.

[209] N. Sukumar, N. Moës, B. Moran, T. Belytschko, *Extended Finite Element Method for Three-Dimensional Crack Modeling*, Int. J. Num. Meth. Engng **48**, 11 (2000), pp.1549-1570.

[210] L. Truskinovsky, *Fracture as a phase transformation*, in Contemporary research in mechanics and mathematics of materials, Ericksen's Symposium, R. Batra, M. Beatty, eds., CIMNE, Barcelona, 1996, pp. 322-332.

[211] J. Wang, J. Li, S. Yip, *Mechanical instabilities of homogeneous crystals*, Phys. Rev. B **52**, 17 (1995), pp. 12627-12635.

[212] J. Wang, J. Li, S. Yip, D. Wolf, S. Phillpot, *Unifying two criteria of Born : elastic instability and melting of homogeneous crystals*, Physica A **240** (1997), pp. 396-403.

Divers

[213] G. Alberti, C. Mantegazza, *A Note on the Theory of SBV Functions*, Bollettino U.M.I. Sez. B **7** (1997), pp. 375-382.

[214] G. Bal, Y. Maday, *A "parareal" time discretization for non-linear PDE's with application to the pricing of an American put*, "Workshop on domain decomposition" proceedings, Zürich, LNCSE Series, Springer Verlag, 2001.

[215] C.S. Gardner, C. Radin, *The infinite-volume ground state of the Lennard-Jones potential*, J. Stat. Phys. **20**, 6 (1979), pp. 719-724.

[216] Z. Li, *A mesh transformation method for computing microstructures*, Numer. Math. **89** (2001), pp. 511-533.

[217] J.-L. Lions, Y. Maday, G. Turinici, *Résolution d'EDP par un schéma en temps "pararéel"*, C. R. Acad. Sci. Paris, Série I, **332**, 7 (2001), pp. 661-668.

[218] K.J. Van Vliet, J. Li, T. Zhu, S. Yip, S. Suresh, *Quantifying the early stages of plasticity through nanoscale experiments and simulations*, Phys. Rev. B **67** (2003), article 104105.