Université Paris-Est

École doctorale MSTIC

Spécialité Mathématiques Appliquées

Thèse de doctorat

# Simulations numériques pour la prédiction de l'évolution microstructurale d'alliages ferritiques. Une étude de la dynamique d'amas.

Pierre TERRIER

*Thèse dirigée par* Gabriel STOLTZ *et* Manuel ATHÈNES

Thèse soutenue le 19 décembre 2018 devant un jury composé de

| | | |
|---|---|---|
| Rapporteur | François CASTELLA | Université Rennes 1 |
| Rapporteur | Julien SALOMON | Inria Paris |
| Examinatrice | Olga MULA | Université Paris Dauphine |
| Examinateur | Charles-Édouard BRÉHIER | Université Lyon 1 |
| Directeur de thèse | Gabriel STOLTZ | École des Ponts et Chaussées |
| Encadrant | Manuel ATHÈNES | CEA |
| Invité | Thomas JOURDAN | CEA |

**Pierre TERRIER**
*Simulations numériques pour la prédiction de l'évolution microstructurale d'alliages ferritiques. Une étude de la dynamique d'amas.*
Thèse de doctorat, 19 décembre 2018
Rapporteurs : François CASTELLA and Julien SALOMON
Directeur de thèse : Gabriel STOLTZ

**Université Paris-Est**
École doctorale MSTIC
Spécialité Mathématiques Appliquées
6-8 avenue Blaise Pascal
77420 and Champs-sur-Marne

*À mes parents et mon frère.*

# Remerciements

*La thèse, c'est comme l'éternité, c'est long, surtout vers la fin.*

— D'après une citation de Woody Allen

Il y a des temps particuliers dans une vie qui imposent de s'arrêter et de se retourner. La fin d'une thèse est de ces moments là. On regarde en arrière, on perçoit le chemin parcouru, on revoit toutes les personnes qui nous ont accompagnés ces trois dernières années. Des personnes sans qui la thèse aurait été probablement différente, sans doute plus difficile, assurément beaucoup moins amusante.

Ce sont toutes ces personnes que j'aimerais remercier, à commencer par les membres du jury, François Castella et Julien Salomon, rapporteurs, ainsi que Olga Mula et Charles-Édouard Bréhier, examinateurs. Vous apportez un point final à ce travail et je vous remercie grandement pour l'intérêt que vous avez manifesté. Je tiens en particulier à remercier Charles-Édouard que j'ai rencontré à plusieurs reprises en conférences et qui m'a été d'une aide précieuse dans la compréhension de certains points techniques.

Mes remerciements vont ensuite à Gabriel Stoltz, mon directeur de thèse, qui a d'abord été un professeur exigeant et un excellent encadrant de stage. Ta rigueur et ta grande discipline n'ont jamais détrompé tes origines alsaciennes, mais ton implication totale et ta disponibilité inégalable (même avec 5 doctorants et un pied dans le plâtre !) m'ont toujours permis de surmonter les difficultés et d'aller un peu plus loin. Je te remercie pour ta confiance depuis toutes ces années, ce fut un réel plaisir de travailler avec toi.

Les deux autres personnes avec qui j'ai eu la joie de travailler sont, selon le terme consacré, mes « encadrants » du CEA, l'un officiel, Manuel Athènes, l'autre officieux, Thomas Jourdan. L'utilisation des guillemets signifie simplement qu'ils ont été, et sont toujours, bien plus que cela. J'ai débuté au SRMP avec Manuel lors de mon stage de M2 et, sans aucune hésitation, j'ai rempilé pour la thèse. Manuel, tu as toujours fait preuve d'une passion et d'un enthousiasme communicatifs. Tes idées, souvent foisonnantes, toujours brillantes, m'ont beaucoup aidé et je te remercie pour ton implication, j'apprécie énormément nos échanges qui sont pour moi une source de stimulation intellectuelle. Quant à Thomas, je dois te remercier — à minima — infiniment pour ton aide précieuse dans la mise en œuvre de l'algorithme dans CRESCENDO. J'ai adoré coder à quatre mains à tes côtés, malgré des bugs et des résultats parfois incompréhensibles, ces moments étaient parfois très drôles et toujours sympathiques. Un grand merci à tous les deux.

Je remercie aussi EDF R&D pour m'avoir donné l'opportunité de me pencher sur des sujets passionnants, plus particulièrement Gilles Adjanor avec qui j'ai eu plaisir de travailler.

Et puis la thèse ne serait rien sans son cadre de travail, qui fut double dans mon cas. Il y a d'abord l'École des Ponts et le CERMICS qui m'ont accueilli. Je remercie au passage le corps des ingénieurs des ponts, des eaux et des forêts qui m'a permis de réaliser cette thèse. Le CERMICS ne saurait fonctionner sans Isabelle et Fatna, que je remercie pour leur aide dans l'organisation des déplacements, pour tous les aspects administratifs, ce fut toujours un plaisir de passer vous voir au secrétariat. Parmi les chercheurs du laboratoire, sans pouvoir être exhaustif, j'aimerais remercier Tony, Frédéric et Benjamin. Tony, tu as le don de rendre simples des objets complexes et quand tu m'expliques des concepts, j'ai toujours l'impression d'être un peu plus intelligent que je ne le suis. Frédéric, je te remercie pour le temps que tu m'as consacré, et même si nous nous sommes finalement peu croisés, j'apprécie beaucoup ta sollicitude. Benjamin, il a fallu que nous nous croisions à Los Angeles pour que je te parle de mes difficultés stochastiques, mais ton aide m'a été précieuse. Je remercie également Éric, à qui la tâche de directeur du CERMICS incombe provisoirement. La recherche et l'ambiance d'un laboratoire seraient tout autre sans les jeunes recrues, les forçats de la recherche, doctorant.e.s et post-doctorant.e.s. Je tiens en particulier à remercier Laura et Florent avec qui j'ai passé de superbes moments à Los Angeles. Je remercie également Julien, Grégoire, Marc, Frédéric, Pierre-Loïk, Adrien, Sami et Boris. Que ce soit en conférences ou ailleurs, ce fut un réel plaisir de vous connaître et de partager des instants sympathiques et plein d'humour.

Il y a ensuite le CEA et le SRMP. C'est sur le plateau de Saclay que j'ai passé le plus de mon temps et malgré des réveils matinaux souvent difficiles, c'était toujours une joie d'aller y retrouver des personnes incroyables. Je pense d'abord aux « permanent.e.s » qui font la marque du SRMP, merci à toutes et à tous, les rencontres ont été belles. J'aimerais en particulier remercier Jean-Paul, toujours prêt à défendre la veuve et l'orphelin, ou au moins les salarié.e.s. Ton engagement couplé à un humour sans faille font que les pauses café sont avec toi toujours de grands moments. Je pense également à Cosmin, Emmanuel et Fabien, avec qui les échanges sont toujours rafraîchissants, merci à vous. Je souhaite également remercier Jean-Luc, chef du SRMP, pour son accueil au sein du service, ainsi que Patricia, toujours présente pour assurer la bonne vie du service. *Last but not least*, je n'oublie pas Claire, la dernière arrivée parmi les permanent.e.s. Merci pour toutes ces discussions à table et en pause, et un grand merci de m'avoir évité une carence en chocolat. Et puis il y a les « jeunes », cet ensemble n'étant pas forcément disjoint de celui des permanents ! Merci à toutes et à tous pour ces petits et grands moments partagés lors d'une pause, d'un repas, d'une sortie ou d'une conférence. Je tiens en particulier à remercier Olivier, animateur de pauses café, source de savoirs et toujours plein d'énergie. Un grand merci à Lisa, pour ta bonne humeur débordante, je pense que ton rire raisonnera encore longtemps dans les couloirs du 520, à Denise, pour ton attention et ta gentillesse et à Marie pour toutes ces belles courses à pied. Je souhaite également remercier Aurélien (vive la Switch !), Camille (x2 !), Elric, Émile, Bérengère (merci de m'avoir introduit à la Guiness !), Adrien (qui sera toujours là pour boire des bières !), Anne-Hélène, Thomas. Je n'oublie pas Arthur, mon co-bureau, ami et encore voisin au Ministère de la transition écologique et solidaire, l'homme qui a voulu faire du photovoltaïque à la Direction de l'énergie nucléaire !

Évidemment, la thèse déborde largement du cadre de travail, c'est une période qui implique souvent bien malgré eux les ami.e.s et la famille. Je tiens d'abord à évoquer une

personne qui m'a accompagné une partie de la thèse et bien avant, nos chemins s'étant séparés depuis. C'est un merci à la saveur douce-amer que je t'adresse, j'imagine que tu comprendras.

Lors de ces années de thèse, j'ai découvert le théâtre d'improvisation et j'y ai rencontré des personnes fantastiques. Je tiens à vous remercier pour ces moments de rigolade et de pur bonheur.

Je souhaite également remercier Antoine et Romain, deux acolytes des Ponts avec qui j'ai partagé les plus belles et longues soirées de geekeries. C'est une vraie chance de vous avoir rencontrés et de vous avoir comme amis.

Raihere, ton amitié et ton sens de l'humour ne m'ont jamais épargné, et je te remercie pour nos conversation parfois tardives, généralement accompagnées d'un bon verre, qui m'ont toujours instruites et qui remettent souvent les choses en perspective. Ta compagnie depuis toutes ces années m'est très chère. Je te remercie pour tout, et ce qu'il reste encore à venir.

Laurent, je pense que ce petit paragraphe ne saurait dire l'affection que j'ai pour toi, ni la chance que j'ai de t'avoir comme ami. En tout cas, je te suis sincèrement reconnaissant pour ton soutien sans faille depuis tant d'années. Tu as toujours été à l'écoute et tu m'amènes constamment à être une meilleure version de moi-même. Merci.

Enfin, j'aimerais conclure en remerciant les êtres qui me sont le plus chers, mon petit frère et mes parents. Sans vous je ne serais pas arrivé jusque là, et je sais que vous serez toujours présents pour me soutenir dans tous les moments de la vie. Surtout, vous m'avez donné ce qui compte le plus, votre amour. Et puisque les écrits restent, je tiens à le dire ici : merci pour tout, je vous aime profondément.

*
* * *

**Épilogue.** Un épilogue dans des remerciements de thèse ? Bah oui, la vie continue ! Il y a bien une vie après la thèse, et la mienne continue au beau ministère de la transition écologique et solidaire. J'en profite donc pour remercier les personnes qui m'ont accordé leur confiance et qui ont vu plus loin que le doctorant ! En tout cas j'espère pouvoir accomplir de belles choses au service de l'État et je me donne rendez-vous dans dix ans ;-)

# Simulations numériques pour la prédiction de l'évolution microstructurale d'alliages ferritiques. Une étude de la dynamique d'amas.


## Résumé

Cette thèse s'intéresse au vieillissement des métaux au niveau microstructural. On étudie en particulier les défauts (amas de lacunes, interstitiels ou solutés) via un modèle de dynamique d'amas (DA), qui permet de prédire l'évolution des concentrations de défauts sur des temps longs (plusieurs dizaines d'années). Ce modèle est décrit par un système d'équations différentielles ordinaires (EDOs) de très grande taille, pouvant excéder la centaine de milliards d'équations. Les méthodes numériques classiques de simulation d'EDOs ne sont alors pas efficaces pour de tels systèmes.

On montre dans un premier temps que la DA est bien posée et qu'elle vérifie certaines bonnes propriétés physiques comme la conservation de la quantité de matière et la positivité de la solution. On s'intéresse également à une approximation de la DA, qui prend la forme d'une équation aux dérivées partielles, de type Fokker–Planck. On caractérise en particulier l'erreur d'approximation entre la DA et cette approximation.

Dans un second temps, on introduit un algorithme de simulation de la DA. Cet algorithme est basé sur un splitting de la dynamique ainsi que sur une interpretation probabiliste des équations de la DA (sous la forme d'un processus de saut) ou de son approximation de Fokker–Planck (sous la forme d'un processus de Langevin). Le but est de réduire le nombre d'équations à résoudre et d'accélérer par conséquent les simulations.

On utilise enfin cet algorithme de simulation à différents modèles physiques. On confirme l'intérêt de ce nouvel algorithme pour des modèles complexes. On montre également que cet algorithme permet d'enrichir le modèle de dynamique d'amas à moindre coût.



## Mots clés

Dynamique d'amas, Équations différentielles ordinaires, Fokker–Planck, Processus de Langevin, Équations différentielles stochastiques, Processus de sauts, Matériaux.

# Numerical simulations for predicting the microstructural evolution of ferritic alloys. A study of Cluster Dynamics.


## Abstract

We study ageing of materials at a microstructural level. In particular, defects such as vacancies, interstitials and solute atoms are described by a model called Cluster Dynamics (CD), which characterize the evolution of the concentrations of such defects, on period of times as long as decades. CD is a set of ordinary differential equations (ODEs), which might contain up to hundred of billions of equations. Therefore, classical methods used for solving system of ODEs are not suited in term of efficiency.

We first show that CD is well-posed and that physical properties such as the conservation of matter and the preservation of the sign of the solution are verified. We also study an approximation of CD, namely the Fokker–Planck approximation, which is a partial differential equation. We quantify the error between CD and its approximation.

We then introduce an algorithm for simulating CD. The algorithm is based on a splitting of the dynamics and couples a deterministic and a stochastic approach of CD. The stochastic approach interprets directly CD as a jump process or its approximation as a Langevin process. The aim is to reduce the number of equations to solve, hence reducing the computation time.

We finally apply this algorithm to physical models. The interest of this approach is validated on complex models. Moreover, we show that CD can be efficiently improved thanks to the versatility of the algorithm.



## Keywords

Cluster Dynamics, Ordinary Differential Equations, Fokker–Planck, Langevin process, Stochastic Differential Equations, Jump process, Materials.

# Publications et communications

## Publications

P. TERRIER, M. ATHÈNES, T. JOURDAN, G. ADJANOR et G. STOLTZ. Cluster dynamics modelling of materials: A new hybrid deterministic/stochastic coupling approach. *J. Comput. Phys.*, 350:280–295, 2017.

G. STOLTZ et P. TERRIER. A mathematical analysis of the Fokker–Planck approximation for Cluster Dynamics. ArXiv:1810.01462.

D. CARPENTIER, T. JOURDAN, P. TERRIER, M. ATHÈNES et Y. LE BOUAR. Effect of sink strength dispersion on particle size distributions simulated by cluster dynamics. *En préparation*.

### Antérieures aux travaux de thèse

P. TERRIER, C. M. MARINICA et M. ATHÈNES. Using Bayes formula to estimate rates of rare events in transition path sampling simulations. *J. Chem. Phys.*, 143(13):134121, 2015.

M. ATHÈNES et P. TERRIER. Estimating thermodynamic expectations and free energies in expanded ensemble simulations: Systematic variance reduction through conditioning. *J. Chem. Phys.*, 146(19):194101, 2017.

## Communications scientifiques

### Communications orales

Multiscale Materials Modeling, Dijon, octobre 2016.

The MRS Spring Meeting & Exhibit, Phoenix, avril 2017.

TMS Annual Meeting & Exhibit, Phoenix, mars 2018.

Séminaire des doctorants du LAMFA, Amiens, décembre 2017.

Congrès d'analyse numérique (CANUM), Cap d'Agde, juin 2018

### Poster

Congrès de la SMAI, La Tremblade, juin 2017,

# Table des matières

**Remerciements** v

**Résumé/Abstract** ix

**Liste des publications** xiii

**1 Étude du vieillissement par des modèles cinétiques** 3
1.1 Modèles cinétiques sur réseaux . . . 3
1.1.1 Présentation du modèle Monte Carlo cinétique . . . 4
1.1.2 Algorithme Monte Carlo cinétique . . . 5
1.1.3 Limites du modèle pour des simulations en temps long . . . 6
1.2 Un premier modèle en champ moyen, l'équation pilote chimique . . . 6
1.2.1 Présentation du modèle . . . 6
1.2.2 Validité du modèle et extension aux matériaux irradiés . . . 7
1.3 De la CME vers la dynamique d'amas . . . 8
1.3.1 Présentation du modèle . . . 8
1.3.2 Dynamique d'amas et équations de Becker–Döring, un même problème bien posé . . . 10
1.3.3 Approximations et simulations . . . 11
1.3.4 La limite Lifshitz-Slyozov-Wagner en temps long . . . 13
1.3.5 Nouvelles approches hybrides . . . 14
1.4 Contributions principales . . . 15
1.4.1 Analyse mathématique de la dynamique d'amas et de l'approximation de Fokker–Planck . . . 15
1.4.2 Présentation d'un nouvel algorithme hybride . . . 16
1.4.3 Présentation de résultats numériques pour des modèles physiques et amélioration du modèle de dynamique d'amas . . . 17

**2 Mathematical analysis** 19
2.1 Introduction . . . 19
2.2 Well-posedness of Cluster Dynamics . . . 21
2.2.1 Full Cluster Dynamics . . . 21
2.2.2 Splitting of the dynamics and qualitative properties . . . 24
2.3 The Fokker–Planck approximation in the linear case . . . 25
2.3.1 Heuristic derivation of the Fokker–Planck approximation . . . 26
2.3.2 Heuristic reformulation as a diffusion equation . . . 27
2.3.3 Decay estimates of the solution of the diffusion equation . . . 31
2.3.4 Relating Cluster Dynamics with its Fokker–Planck approximation . . . 34
2.A Proofs for the well-posedness of CD . . . 35
2.A.1 Regularized Cluster Dynamics . . . 35
2.A.2 Proof of the existence of a global-in-time solution . . . 40
2.A.3 Proof of the uniqueness of a global-in-time solution . . . 42

2.B Proofs related to the splitting of the dynamics . . . 45
2.B.1 Estimates on the second subdynamics . . . 45
2.B.2 Some estimates on elements of $\mathcal{Q}$ . . . 47
2.B.3 Proof of the convergence of the splitting . . . 49
2.C Proofs for the decay estimates of the solution to the diffusion equation . . . 55
2.C.1 Proof of the Theorem . . . 55
2.C.2 Some technical results on $\varphi$, $\Psi$, $\eta$ and their derivatives . . . 58
2.C.3 Proofs on the relation between Cluster dynamics and its Fokker–Planck approximation . . . 62

**3 A new hybrid deterministic/stochastic coupling approach 67**
3.1 Introduction . . . 67
3.2 Model description and main algorithm . . . 68
3.2.1 Rate equations . . . 68
3.2.2 Splitting of the dynamics . . . 69
3.2.3 Main algorithm . . . 72
3.3 Discretization of the large size cluster subdynamics . . . 73
3.3.1 The Jump process approach . . . 74
3.3.2 The Langevin process approach . . . 77
3.4 Approximating the dynamics of $C_{\mathrm{vac}}$ . . . 79
3.4.1 Decomposition into elementary integrable ODEs . . . 79
3.4.2 Quasi-stationary limit . . . 80
3.4.3 Mass conservation . . . 80
3.5 Numerical analysis . . . 80
3.5.1 Deterministic splitting . . . 80
3.5.2 Stochastic error . . . 81
3.5.3 Coupling of errors . . . 82
3.6 Results . . . 83
3.6.1 On the quasi-stationary assumption . . . 83
3.6.2 Accuracy of the splitting algorithm for thermal ageing . . . 85
3.6.3 Parallelization . . . 87
3.7 Conclusion . . . 87

**4 Some applications of the coupling algorithm 89**
4.1 Introduction . . . 89
4.2 Implementation of the coupling algorithm in CRESCENDO . . . 89
4.2.1 The CD code CRESCENDO . . . 90
4.2.2 Modifications of the hybrid coupling algorithm for CRESCENDO . . . 92
4.3 Using the coupling algorithm in CRESCENDO for Fe and FeHe under irradiation . . . 94
4.3.1 Fe under irradiation . . . 94
4.3.2 FeHe under irradiation . . . 97
4.4 Improving Cluster Dynamics with stochastic coefficients . . . 99
4.4.1 Introducing sink strength dispersion in cluster dynamics simulations 99

**5 Conclusion and perspectives 105**

**Bibliography 107**

# Introduction

L'étude du vieillissement des matériaux du nucléaire (acier de cuve et éléments internes en particulier) est cruciale, d'un point de vue environnemental, économique et social. La compréhension de leur évolution, sous des contraintes bien particulières induites par l'irradiation, est nécessaire pour assurer le bon fonctionnement des centrales nucléaires. L'irradiation que subissent les matériaux vient en effet modifier différentes propriétés, comme des propriétés mécaniques telles que la ténacité (*i.e.* la capacité d'un matériau à résister à la propagation d'une fissure) ou la ductilité (*i.e.* la capacité d'un matériau à se déformer sans se rompre), et suscite différents phénomènes physiques, comme le durcissement (*i.e.* une diminution de la ductilité) ou le gonflement (*i.e.* une augmentation du volume du matériau). Ces phénomènes macroscopiques s'expliquent par une modification de la micro-structure, qu'il convient donc d'étudier précisément. On comprend toutefois que le choix d'une approche, qu'elle soit micro-, méso- ou macroscopique, s'intègre dans une vision multi-échelle du phénomène de vieillissement.

La compréhension globale du phénomène de vieillissement sous irradiation sur des temps physiques de l'ordre de dizaines d'années — typiquement plus de 40 ans, le temps d'exploitation d'un réacteur nucléaire — nécessite donc des méthodes adaptées aux différents phénomènes qui entrent en jeu. Ainsi on ne peut pas comprendre le phénomène de gonflement sans comprendre l'évolution de la microstructure et des amas lacunaires du matériau, et par ailleurs, les paramètres utilisés aux échelles de la microstructure nécessitent parfois d'avoir recours à des méthodes à l'échelle atomique pour obtenir un accord avec l'expérimentation.

Des projets d'intégration multi-échelle sont donc fondamentaux et l'on peut citer le projet PERFECT (*Prediction of Irradiation Damage Effects in Reactor Components*), projet européen financé par le sixième programme-cadre pour la recherche et le développement technologique et coordonné par EDF, entre 2004 et 2008. Ce projet illustre la complexité de combiner des phénomènes aux différentes échelles.

Sans prétendre à l'exhaustivité, on peut illustrer le long enchaînement des méthodes qui conduit des dommages causés par l'irradiation jusqu'au contraintes mécaniques responsables de fractures. Ainsi on part de la mécanique quantique et de la dynamique moléculaire pour simuler les premières cascades de déplacements d'atomes liés aux irradiations. Ces cascades sont responsables de la production de défauts (lacunes, interstitiels) dont l'évolution va être prédite par des méthodes cinétiques (Monte Carlo cinétique, dynamique d'amas). Ces défauts vont interagir avec des dislocations responsables de la plasticité du matériau, et dont les interactions seront étudiées à une échelle mésoscopique *via* la dynamique des dislocations. Enfin, les résultats obtenus par dynamique des dislocations,

éventuellement intégrés à d'autres échelles intermédiaires, fourniront des paramètres comme les tenseurs de contraintes pour des modèles macroscopiques généralement simulés à l'aide de méthodes par éléments finis.

Ce complexe système multi-échelle révèle les limites de chaque méthode, ces limites étant liées à notre capacité à simuler de tels phénomènes. Ainsi, les méthodes dites *ab-initio*, c'est-à-dire basées sur les équations de la mécanique quantique, ne permettent de simuler qu'un tout petit nombre d'atomes (de l'ordre de la centaine au maximum) et sur des temps de l'ordre de la nanoseconde. La dynamique moléculaire, basée sur les lois de la physique classique, parfois paramétrée par des résultats *ab-initio*, permet de simuler un plus grand nombre d'atomes (jusqu'à un million d'atomes sur de courtes simulations de l'ordre de la nanoseconde) et d'atteindre des temps de l'ordre de la milliseconde. Il faut aller vers les modèles cinétiques, comme l'AkMC (Atomistic kinetic Monte Carlo) pour atteindre des temps plus longs allant de la minute à l'année.

Pour étudier efficacement les phénomènes liés à l'évolution de la micro-stucture, en particulier l'agrégation de défauts, les méthodes cinétiques Monte Carlo sur Objets (OkMC) ou sur événements (EkMC) ainsi que les modèles en champs moyen de type dynamique d'amas sont les plus adaptés. Ils permettent d'atteindre des temps longs de l'ordre de l'année tout en produisant des résultats proches de l'expérience.

Le Chapitre 1 est l'occasion de présenter plus en détail ces différentes approches et de passer en revue les différents verrous qui existent encore et qui limitent nos capacités à simuler des problèmes complexes. La suite porte plus particulièrement sur l'étude de la dynamique d'amas. D'abord d'un point de vue mathématique, on montre dans le Chapitre 2 que les outils employés par la communauté des sciences des matériaux sont bien définis et nous présentons quelques résultats qui viendront valider la méthode de simulation numérique du Chapitre 3. En effet, dans ce chapitre on présente un nouvel algorithme pour la dynamique d'amas, dont le but est de réduire les temps de simulation tout permettant la prise en compte d'amas de grande taille. Le Chapitre 4 présente l'ensemble des résultats obtenus avec le nouvel algorithme sur des cas tests réalistes et d'intérêts pour le CEA et EDF.

# 1 Étude du vieillissement par des modèles cinétiques

Les phénomènes responsables du vieillissement des matériaux, et particulièrement ceux du nucléaire, se caractérisent par leur échelles, spatiale — la microstructure (quelques nanomètres) — et temporelle — les temps considérés sont longs du fait de leur nature industrielle, le rapport entre les phénomènes atomiques et les années d'exploitation d'une centrale étant alors considérable. Ces contraintes empêchent l'utilisation de la dynamique moléculaire, dont les pas de temps sont de l'ordre de $10^{-15}$ secondes, et qui permet de simuler des phénomènes sur quelques microsecondes tout au plus. Cette limitation a suscité le développement de méthodes cinétiques qui viennent décrire l'évolution d'événements à l'échelle atomique (migration de défauts, réactions chimiques, etc) sur des temps plus longs, depuis les années 1960 [YE66 ; Gil76]. La pertinence et l'efficacité de ces modèles cinétiques dans de nombreux domaines (chimie, physique, biologie) a conduit à une littérature foisonnante et au développement de nombreuses classes de méthodes. Nous en distinguons ici deux grandes classes. La première contient les modèles cinétiques sur réseaux, comme les méthodes Monte Carlo cinétique atomistique (AkMC pour *Atomistic kinetic Monte Carlo*) [YE66 ; Bor+75 ; Soi+10] ou encore Monte Carlo cinétique sur objet (OkMC pour *Object kinetic Monte Carlo*) [Bec+10] ou sur événement [Lan74]. Ces méthodes conservent une vision spatiale de la microstructure et sont introduites en Section 1.1. La seconde contient des approches en champ moyen, qui s'affranchissent de cette description spatiale. On présentera les modèles d'équations pilotes chimiques (CME pour *Chemical Master Equation*) [Gil76], en Section 1.2 ou encore la dynamique d'amas (RECD pour *Rate Equation Cluster Dynamics*) [Goo64 ; Wol+77] en Section 1.3. Enfin, les contributions de ce travail de thèse sont résumées en Section 1.4.

## 1.1 Modèles cinétiques sur réseaux

Les méthodes AkMC et OkMC sont très proches dans leur mise en œuvre mais ont des spécificités qui orientent dans le choix d'une méthode ou de l'autre en fonction des phénomènes étudiés. Ainsi les méthodes AkMC sont utilisées pour avoir un niveau de détail assez fin des phénomènes étudiés avec une modélisation explicite de chaque atome du réseau et de l'ensemble des interactions supposées. Au contraire, les méthodes OkMC tendent à regrouper des éléments de plusieurs atomes (amas de lacunes, interstitiels, etc.) en un seul objet et à construire des interactions entre ces différents objets. Pour une introduction générale aux méthodes de Monte Carlo cinétique, on renvoie le lecteur à la très bonne revue de Arthur Voter sur le sujet [Vot07]. Nous présentons ici les grands principes de la méthode. En Section 1.1.1, nous décrivons les équations du modèle ainsi que les limites fondamentales à une simulation déterministe. Un algorithme de simulation stochastique est alors présenté en Section 1.1.2, avant de décrire les limites du modèle pour des simulations en temps longs (Section 1.1.3).

### 1.1.1 Présentation du modèle Monte Carlo cinétique

L'introduction des méthodes kMC part du constat que les événements importants dans l'évolution de la microstructure sont des événements peu fréquents du fait de la métastabilité du système. Ainsi une simulation en dynamique moléculaire serait peu efficace puisque la majorité du temps de la simulation, on observerait une oscillation d'un objet — généralement des atomes — autour de sa position d'équilibre mécanique local pendant longtemps avant qu'il ne s'échappe et oscille dans un autre bassin d'attraction. En considérant que le système « oublie » la façon dont il est arrivé dans un bassin d'énergie, puisqu'il y reste longtemps [DG+16], on suppose alors qu'il suffit de le caractériser par des probabilités de transition entre différents états métastables. On définit donc des taux de transitions, notés $k_{ij}$, qui décrivent la probabilité de passer d'un état $i$ à un état $j$. À partir de ces probabilités de transition, on peut alors formuler un système d'équations gouvernant l'évolution du vecteur probabilité du système, c'est-à-dire définir un problème d'évolution appelé équation pilote (ou *Master Equation* en anglais).

Supposons que le système est composé de $N > 0$ configurations possibles et considérons donc $P$ le vecteur probabilité du système, de taille $N$ (l'élément $P_i$ représentant la probabilité d'être dans l'état $i$) et la matrice de transition $M \in \mathbb{R}^{N\times N}$ telle que

$$\forall\, i,j \in \{1,\cdots,N\}\,, \quad \begin{cases} M_{ij} = k_{ij}, & i \neq j, \\ M_{ii} = -\displaystyle\sum_{\ell\neq i} k_{i\ell}. \end{cases}$$

Alors l'évolution du vecteur probabilité (ou aussi vecteur aléatoire à densité) $P$ est donné par l'équation pilote suivante :

$$\forall t \geqslant 0, \qquad \dot{P}(t) = MP(t). \tag{1.1}$$

La solution analytique d'un tel système s'écrit

$$\forall t \geqslant 0, \qquad P(t) = \exp(Mt)P(0), \tag{1.2}$$

$P(0)$ correspondant à l'état initial du système. On note par ailleurs que, pour tout $1 \leqslant i \leqslant N$, la probabilité $P_i$ d'être dans l'état $i$ satisfait

$$\dot{P}_i(t) = \sum_{1\leqslant j\leqslant N} M_{ij}P_j(t) = \sum_{j\neq i} \big(M_{ij}P_j(t) - M_{ji}P_i(t)\big). \tag{1.3}$$

Le premier problème que rencontre une telle approche est la taille de cette matrice $M$. En effet, en considérant un modèle d'un réseau 2D de taille $L \times L$ où chaque site peut avoir 2 états différents (par exemple un modèle d'Ising [Isi25]), le nombre de configurations $N$ est $N = 2^{L^2}$ et la matrice $M$ sera donc de taille $N^2$. Un simple maillage avec $L = 10$ contient donc plus de $1.26 \times 10^{30}$ configurations. Il est évident qu'il est impossible de stocker une telle matrice sur un ordinateur actuel et encore moins de calculer analytiquement la solution (1.2). Bien sûr, le nombre de chemins possibles pour passer d'une configuration à une autre est limité, et un grand nombre de taux de transitions $k_{ij}$ sont nuls, de telle sorte que $N_{\text{possibles}} \ll N$, où $N_{\text{possible}}$ est le nombre de taux non nuls. Ainsi, la matrice $M$ est creuse et dans certains cas très simples, il est possible d'obtenir des informations sur le système sans avoir à utiliser un algorithme de type kMC présenté dans la section suivante [Red01].

### 1.1.2 Algorithme Monte Carlo cinétique

La littérature concernant les méthodes kMC est foisonnante, et de nombreuses méthodes ont été développées au cours des années [Bor+75 ; Met+53 ; Lan74 ; Vot86]. Nous donnons ici un algorithme naïf basé sur une vision mathématique de l'équation pilote. En effet, l'équation (1.1) est l'équation de Kolmogorov première d'un processus de Markov [VK92], dont les taux de transition sont indépendants du temps. Ce processus de Markov $(X_t)_{t\geqslant 0}$ est tel que

$$\mathbb{P}\left(X_{t+h} = i \,|\, X_t = j\right) = \delta_{ij} + hM_{ij} + \mathrm{o}(h),$$

où $h$ tend vers $0$, $\delta_{ij}$ est le symbole de Kronecker, et $\mathbb{P}\left(X_{t+h} = i | X_t = j\right)$ représente la probabilité que $X$ soit dans l'état $i$ à l'instant $t+h$ sachant qu'il se trouvait dans l'état $j$ à l'instant $t$. En notant $P_i(t) = \mathbb{P}(X(t) = i)$ la probabilité que le processus se trouve dans l'état $i$ à l'instant $t$, le vecteur $P = (P_i)_{1\leqslant i\leqslant N}$ satisfait (1.1). Une définition équivalente [Nor97] d'un tel processus correspond à considérer un processus de saut $(Y_i)_{i\geqslant 1}$, qui est une chaîne de Markov sur les états $1, \cdots, N$ et des variables $(S_i)_{i\geqslant 1}$ qui décrivent le temps passé dans chaque état, les variables $S_i$ étant indépendantes, de loi exponentielle de paramètre $-M_{Y_iY_i}$. En notant que sous la condition $Y_n = i$, la variable aléatoire $Y_{n+1}$ suit une loi de Bernoulli généralisée $\left(\pi^i_j = M_{ij}/(-M_{ii})\right)_{1\leqslant j\leqslant N}$, un algorithme très simple permet de simuler des trajectoires statistiquement compatibles avec l'équation pilote (1.1). L'estimation de la solution est alors obtenue par des moyennes de chemin.

Considérant que le système est dans un état $i$ à l'instant initial $t_0 = 0$, on tire un premier temps de saut $\tau_i$ selon une loi exponentielle de paramètre $-M_{ii}$ et on incrémente le temps de la simulation à $t_1 = t_0 + \tau_i$. Ensuite on tire un état $j$ selon la loi de Bernoulli généralisée $\pi^i$ et on passe le système à l'état $j$. On itère ainsi le processus en partant du temps $t_1$ et de l'état $j$.

Une particularité de la loi exponentielle permet d'adopter un autre point de vue sur les méthodes de Monte Carlo cinétique. Considérons donc la variable aléatoire $S_i$ suivant une loi exponentielle de paramètre $-M_{ii} = \sum_j k_{ij}$ et considérons également les variables aléatoires indépendantes $(\mathcal{S}^j_i)_{1\leqslant j\leqslant N}$ de loi exponentielle de paramètre $(k_{ij})_{1\leqslant j\leqslant N}$. Alors $S_i$ a la même loi que $\min\left(\mathcal{S}^1_i, \cdots, \mathcal{S}^N_i\right)$. En effet, en notant que $\mathbb{P}\left(\min\left(\mathcal{S}^1_i, \cdots, \mathcal{S}^N_i\right) \leqslant s\right) = 1 - \mathbb{P}\left(\mathcal{S}^1_i \geqslant s, \cdots, \mathcal{S}^N_i \geqslant s\right)$, on a, par indépendance,

$$\begin{aligned}\mathbb{P}\left(\min\left(\mathcal{S}^1_i, \cdots, \mathcal{S}^N_i\right) \leqslant s\right) &= 1 - \prod_{j=1}^{N} \mathbb{P}\left(\mathcal{S}^j_i \geqslant s\right) = 1 - \prod_{j=1}^{N} \exp(-k_{ij}s) \\ &= 1 - \exp\left(-\sum_{j=1}^{N} k_{ij}s\right) = \mathbb{P}(S_i \leqslant s).\end{aligned}$$

Ainsi, tirer un temps selon la loi de $S_i$ est équivalent à choisir le premier événement parmi tous les événements tirés selon les lois $(\mathcal{S}^j_i)_{1\leqslant j\leqslant N}$. Cette propriété de la loi exponentielle est en particulier utilisée pour justifier l'algorithme EkMC (*Event kinetic Monte Carlo*) [Lan74]. Dans cet algorithme, les temps des événements sont tirés de manière indépendante en utilisant les lois exponentielles correspondantes, puis le premier événement est réalisé. Après chaque événement, il est nécessaire d'actualiser la liste des événements permis

en supprimant ceux qui sont devenus impossibles et en ajoutant ceux qui deviennent possibles.

### 1.1.3 Limites du modèle pour des simulations en temps long

L'algorithme de Monte Carlo cinétique que nous venons de présenter illustre bien la philosophie des méthodes cinétiques sur réseau. Considérant le système dans une certaine configuration, on cherche le moment où le système va changer d'état parmi tous les chemins possibles. Toutefois, compte tenu du grand nombre d'événements possibles, une des premières limitations des méthodes kMC concerne les temps de sauts très courts entre chaque événement. Ceci s'illustre mathématiquement avec le fait que l'espérance d'une loi exponentielle de paramètre $\lambda_i = \sum_j k_{ij}$ est égale à $\lambda_i^{-1}$. Le temps tiré selon cette loi exponentielle sera généralement très petit, ce qui ne permet pas de simuler des phénomènes sur des temps longs. Les temps de simulations se limitent généralement à quelques dizaines de secondes [Vot07].

Pour pallier ce problème tout en conservant une description spatiale du système, différentes méthodes, dont l'OkMC, ont été développées. L'OkMC ne tient plus compte de l'ensemble des atomes du système, mais considère seulement des objets qui interagissent entre eux, comme des amas de lacunes ou des boucles de dislocation. Ainsi ces objets se déplacent seulement en fonction de leur diffusivité, ce qui évite de calculer les nombreuses étapes de déplacement d'atomes individuels conduisant à un déplacement global. Ainsi, les méthodes OkMC permettent d'atteindre des temps de simulations bien plus longs, de l'ordre de la seconde, voire de conditions réelles de réacteurs sous pression sur 30 ans, pour des petites boites de simulation [Dom+04].

Un second problème, qui apparaît davantage avec les méthodes « *coarsed grained* » de type OkMC, concerne la modélisation même des phénomènes physiques. Les différentes transitions sont en effet tirées d'une liste d'événements déterminés *a priori*. Il est alors possible que certaines transitions soient omises, car inconnues ou contre-intuitives, et pourtant importantes dans certains processus. Cette difficulté est discutée dans la revue [Vot07].

## 1.2 Un premier modèle en champ moyen, l'équation pilote chimique

Les méthodes cinétiques sur réseaux sont limitées par leur vision spatiale des réactions. Les approches en champ moyen s'absolvent de telles contraintes en considérant uniquement les réactions entre éléments d'un matériau ou d'un système chimique ou biologique. C'est d'ailleurs dans le cadre de systèmes chimiques complexes impliquant plusieurs éléments et réactions que l'algorithme SSA (*Stochastic Simulation Algorithm*) et son équation associée, l'équation pilote chimique (CME), ont été introduits [Gil76 ; McQ67]. On présente le modèle en Section 1.2.1, tout en notant le formalisme liant la CME aux méthodes kMC. Les limites du modèle sont discutées en Section 1.2.2

### 1.2.1 Présentation du modèle

L'algorithme SSA est conçu pour simuler un système de $L$ éléments chimiques interagissants *via* $M$ réactions chimiques $(R_1, \cdots, R_M)$ dans un volume $V$ fixé. On décrit un tel système par un vecteur d'état $X(t) = (X_1(t), \cdots, X_L(t))$ où $X_i(t)$ représente le nombre

de molécules de type $i$ à un instant $t$. Gillespie [Gil07] caractérise l'évolution du système par un vecteur de changement d'état $\nu_j = (\nu_{1,j}, \cdots, \nu_{L,j})$ pour chaque réaction $R_j$ où $\nu_{i,j}$ définit le changement d'état dans la population $X_i$ pour une réaction $R_j$. Il définit également la probabilité $a_j(x)h$ qu'une réaction $R_j$ survienne lors d'un intervalle de temps $[t, t+h]$ sachant que $X(t) = x$. En introduisant

$$P(x, t \,|\, x_0, t_0) = \mathbb{P}\left(X(t) = x \,|\, X(t_0) = x_0\right),$$

il montre que $P$ vérifie l'équation pilote chimique (CME)

$$\frac{dP(x, t \,|\, x_0, t_0)}{dt} = \sum_{j=1}^{M} \left(a_j(x - \nu_j)P(x - \nu_j, t \,|\, x_0, t_0) - a_j(x)P(x, t \,|\, x_0, t_0)\right). \tag{1.4}$$

Le premier terme du membre de droite représente alors la production d'espèce $x$ *via* les réactions à partir des espèces $x - \nu_j$ tandis que le second terme représente la disparition d'espèce $x$ *via* les réactions produisant des espèces de type $x - \nu_j$. On reconnaît en fait le formalisme présenté lors du modèle de Monte Carlo cinétique *via* l'équation (1.3). Soit $N$ le nombre d'états possibles du système (déterminé *a priori* puisque le vecteur $X$ représente le nombre de molécules de chaque type). Dans le modèle de Gillespie, ils sont caractérisés par des multi-indices $x = (x_1, \cdots, x_L) \in \mathbb{N}^L$. On peut les ordonner, et définir une bijection $f$ de l'ensemble des multi-indices du système dans $\{1, \cdots, N\}$. Soit donc $P$ le vecteur probabilité, de taille $N$, tel que $P_i(t) = \mathbb{P}\left(X(t) = f^{-1}(i)\right)$, alors $P$ vérifie

$$\forall t \geqslant t_0, \qquad \dot{P}(t) = M^{\mathrm{SSA}} P(t),$$

où, pour $1 \leqslant i \neq j \leqslant N$ et $x = f^{-1}(i)$, on a $M^{\mathrm{SSA}}_{ij} = a_j(x - \nu_j)$ lorsque $\nu_j$ est différent de zéro, et $M^{\mathrm{SSA}}_{ii} = -\sum_{i \neq j} M^{\mathrm{SSA}}_{ij}$.

On retrouve donc une équation pilote de la forme (1.1). On peut alors y associer un processus de Markov et en déduire un algorithme pour simuler un tel processus (voir la Section 1.1.2). Enfin, notons que les probabilités de réactions $a_j(x)h$ dépendent potentiellement de la quantité d'un ou plusieurs éléments $x_i$, ce qui peut conduire à des non-linéarités dans la CME. Par exemple, si $R_j$ caractérise une réaction bi-moléculaire entre des éléments de type $k$ et $\ell$, des arguments de théorie cinétique [Gil07] donnent $a_j(x) \propto x_k x_\ell$.

### 1.2.2 Validité du modèle et extension aux matériaux irradiés

La méthode en champ moyen introduite par Gillespie s'applique essentiellement aux systèmes chimiques et biologiques [GB00 ; MA97 ; Sam+05]. En effet, la CME est valide pour des systèmes bien mélangés et à l'équilibre thermique [Gil92], ce qui est généralement le cas des systèmes chimiques et biologiques contrôlés par des réactions dans des milieux gazeux de réactants. Le problème des matériaux et plus généralement de la matière condensée est que les réactions sont généralement contrôlées par un phénomène de diffusion lente d'espèces qui réagissent lorsqu'elles entrent en collision [MB11]. La pertinence d'une telle extension aux matériaux est un peu discutée [LR91]. Toutefois, une telle discussion est hors de propos et de nombreuses comparaison à des modèles de types Monte Carlo cinétiques montrent la validité de l'approche en champ moyen.

Dans un récent papier [MB11], Marian et Bulatov montrent l'intérêt d'utiliser ces approches

stochastiques pour la simulation de matériaux irradiés. Leur problème est de déterminer correctement les taux de réaction $a_j(x)$ qui apparaît comme le principe fondamental pour caractériser la méthode SSA. Pour cela, ils ont recours à la dynamique d'amas, second modèle en champ moyen, mais fondamentalement lié à la CME.

## 1.3 De la CME vers la dynamique d'amas

La dynamique d'amas, et plus généralement les équations d'états de réaction (RRE pour *Reaction-Rate Equation*) sont issues d'une approximation de la CME (1.4). Considérons la moyenne des états $\mathbb{E}[X(t)] = \sum xP(x,t \mid x_0,t_0)$, où la somme est prise sur l'ensemble des états possibles. Alors

$$\frac{d}{dt}\mathbb{E}[X(t)] = \sum_{j=1}^{M} \nu_j \mathbb{E}[a_j(X(t))].$$

En effet, on a

$$\begin{aligned}
\frac{d}{dt}\mathbb{E}[X(t)] &= \sum_{x\in\mathbb{N}^L} x \sum_{j=1}^{M} (a_j(x-\nu_j)P(x-\nu_j,t \mid x_0,t_0) - a_j(x)P(x,t \mid x_0,t_0)) \\
&= \sum_{j=1}^{M} \sum_{x\in\mathbb{N}^L} x\,(a_j(x-\nu_j)P(x-\nu_j,t \mid x_0,t_0) - a_j(x)P(x,t \mid x_0,t_0)) \\
&= \sum_{j=1}^{M} \sum_{y\in\mathbb{N}^L} (y+\nu_j)a_j(y)P(y,t \mid x_0,t_0) - \sum_{j=1}^{M} \sum_{x\in\mathbb{N}^L} a_j(x)P(x,t \mid x_0,t_0)
\end{aligned}$$

avec le changement de variable $y = x - \nu_j$ ce qui permet de conclure. On peut montrer par ailleurs [Gil07] que dans la limite $V \to +\infty$, $\mathbb{E}[a_j(X(t))] = a_j(\mathbb{E}[X(t)])$, ce qui donne une équation sur les concentrations $C_i = \lim_{V\to+\infty} \mathbb{E}(X_i)/V$ de la forme

$$\frac{dC_i}{dt} = f_i(C_1,\cdots,C_L), \qquad 1 \leqslant i \leqslant L,$$

où les fonctions $f_i$ dépendent en particulier des taux de réactions $a_j$, pour $1 \leqslant j \leqslant M$ et sont caractérisées par les différentes réactions $R_j$ possibles. On obtient donc un système d'équations différentielles ordinaires d'ordre 1 qui décrit l'évolution des concentrations au cours du temps. On présente plus précisément les équations de la dynamique d'amas en Section 1.3.1. On discute ensuite des différentes stratégies qui ont été développées afin de simuler les équations de la dynamique d'amas (Section 1.3.3) avant d'évoquer plus particulièrement le développement récent d'approches hybrides (Section 1.3.5).

### 1.3.1 Présentation du modèle

Dans les cas d'applications aux systèmes chimiques et biologiques, les réactions sont généralement au plus d'ordre 2, puisqu'on considère qu'une réaction faisant intervenir 3 éléments correspond à une succession de 2 réactions. C'est également le cas dans l'étude de l'évolution des concentrations de défauts dans les matériaux. Plus précisément, on aura

— des réactions d'ordre 0, sous la forme de constantes $G_i$ décrivant le taux de défauts de type $i$ créé au cours du temps. On appelle généralement $G_i$ un terme source. Il est

lié à l'irradiation que subit le matériaux et est obtenu *via* la simulation des premières cascades de déplacements [DLR+97 ; RT74 ; Nor+98 ; JC12] ;

— des réactions d'ordre 1, sous la forme $\alpha_{i,j}C_i$ décrivant le phénomène d'émission d'un amas de défaut de type $j$ par un amas de défaut de type $i$ ;

— des réactions d'ordre 2, sous la forme $\beta_{i,j}C_iC_j$ décrivant le phénomène d'absorption d'un amas de type $j$ par un amas de type $i$.

Les phénomènes d'absorption et d'émission donnent en général toute la thermodynamique du système. On peut parfois enrichir le modèle avec d'autres termes d'ordre 1 (*cf.* Chapitre 4), représentant des forces de puits caractérisant l'absorption d'amas par des dislocations ou des joints de grain.

### Modèle 1D

On présente un premier modèle assez simple de croissance d'amas de lacunes dans du fer sous irradiation, le fer étant le principal composant des aciers de cuve. Cet exemple est typique de la dynamique d'amas puisqu'il considère les principaux phénomènes d'absorption et d'émission. La dynamique des amas immobiles de taille $n$ s'écrit alors

$$\frac{dC_n}{dt} = \beta_{n-1}C_{n-1}C_1 - (\beta_n C_1 + \alpha_n)C_n + \alpha_{n+1}C_{n+1}, \tag{1.5}$$

avec $\alpha_n$ et $\beta_n$ de la forme [Ovc+03]

$$\begin{cases} \beta_n = \beta_0 n^{1/3}, \quad n \geqslant 1 \\ \alpha_n = \alpha_0 n^{1/3} \exp\left(-\dfrac{E^{\mathrm{b}}_{\mathrm{vac}}(n)}{k_{\mathrm{B}}T}\right), \end{cases}$$

où $\alpha_0 = \beta_0 = (48\pi^2/V_{\mathrm{at}}^2)^{1/3}D_{\mathrm{vac}}$, avec $V_{\mathrm{at}}$ le volume atomique $D_{\mathrm{vac}}$ le coefficient de diffusion des lacunes. Le terme $E^{\mathrm{b}}_{\mathrm{vac}}(n)$ représente l'énergie de liaison d'une lacune avec un amas de taille $n$ :

$$E^{\mathrm{b}}_{\mathrm{vac}}(n) = E^{\mathrm{f}}_{\mathrm{vac}} - \frac{2\gamma V_{\mathrm{at}}}{r(n)},$$

où $\gamma$ est une énergie de surface, $r$ est le rayon de la cavité, supposée sphérique, donné par $r(n) = (3nV_{\mathrm{at}}/4\pi)^{1/3}$ et $E^{\mathrm{f}}_{\mathrm{vac}}$ est l'énergie de formation d'une lacune, obtenue par des calculs *ab-initio*. Les lacunes ($n = 1$) sont les seuls amas mobiles du modèle considéré. La dynamique d'évolution de la concentration de lacunes $C_1$ est telle que

$$\frac{dC_1}{dt} = G_1 - 2\beta_1 C_1^2 - \sum_{n\geqslant 2} \beta_n C_n C_1 + \sum_{n\geqslant 2} \alpha_n C_n + \alpha_2 C_2. \tag{1.6}$$

Cette équation traduit une équation de bilan sur les lacunes. Elle implique en particulier une équation de conservation sur la quantité de matière totale $\mathcal{Q}_{\mathrm{tot}}$ du système simulé :

$$\frac{d\mathcal{Q}_{\mathrm{tot}}}{dt} = \frac{d}{dt}\left(C_1 + \sum_{n\geqslant 2} nC_n\right) = G_1.$$

Dans les Chapitres 2 et 3, nous étudions le cas $G_1 = 0$, c'est-à-dire sans irradiation.

#### Généralisation

On présente également un modèle plus complet décrivant l'évolution des concentrations de défauts d'un matériau irradié, comprenant des amas de lacunes, des amas interstitiels ainsi que des solutés. De façon générale, un amas est identifié par un $k$-tuple $\nu = (n, p_1, \cdots, p_{k-1})$ où $n \in \mathbb{Z}$ caractérise le nombre de lacunes ($n < 0$) ou d'interstitiels ($n > 0$) et les $p_j \in \mathbb{N}$ représentent le nombre d'atomes de soluté de type $j$. En notant $\mathcal{M}$ l'ensemble des amas mobiles et $\Omega$ l'ensemble des amas, la dynamique sur les amas immobiles s'écrit, pour $\nu \in \Omega \backslash \mathcal{M}$,

$$\frac{dC_\nu}{dt} = G_\nu + \sum_{\mu \in \mathcal{M}} \beta_{\nu,\mu} C_{\nu-\mu} C_\mu - (\beta_{\nu,\mu} C_\mu + \alpha_{\nu,\mu}) C_\nu + \alpha_{\nu,\mu} C_{\nu+\mu}, \tag{1.7}$$

où $G_\nu$ correspond au taux de création d'un amas de type $\nu$, les coefficients $\alpha_{\nu,\mu}$ et $\beta_{\nu,\mu}$ représentent respectivement les coefficients d'émission et d'absorption d'un amas mobile de type $\mu$ par un amas de type $\nu$. Notons que $\beta_{\nu,\mu} = 0$ si l'amas de type $\nu - \mu$ n'est pas défini (par exemple si l'indice d'un soluté est strictement négatif). La dynamique des amas mobiles s'écrit quant à elle

$$\begin{aligned} \frac{dC_\nu}{dt} = G_\nu + \sum_{\mu \in \mathcal{M}} \beta_{\nu,\mu} C_{\nu-\mu} C_\mu - (\beta_{\nu,\mu} C_\mu + \alpha_{\nu,\mu}) C_\nu + \alpha_{\nu+\mu,\mu} C_{\nu+\mu} \\ - \sum_{\mu \in \Omega} \beta_{\nu,\mu} C_\nu C_\mu - \alpha_{\nu+\mu,\mu} C_{\nu+\mu} - \sum_{i=1}^{N_i} k_{i,\nu}^2 D_\nu \left(C_\nu - C_\nu^{\mathrm{eq}}\right), \end{aligned} \tag{1.8}$$

où les coefficients $k_{i,\nu}^2$ représentent les forces de puits de type $1 \leqslant i \leqslant N_i$, $D_\nu$ le coefficient de diffusion des amas de type $\nu$ et $C_\nu^{\mathrm{eq}}$ les concentrations d'équilibre de tels amas. Les forces de puits sont en fait dues à l'interaction des amas avec des objets comme les dislocations ou les joints de grain. Ces objets peuvent capter des amas de défauts mais n'en émettent pas.

### 1.3.2 Dynamique d'amas et équations de Becker–Döring, un même problème bien posé

L'approche présentée ici consiste à aborder les problèmes de vieillissement des matériaux irradiés *via* des modèles cinétiques. Ces modèles cinétiques présentent la même structure mathématique, une équation pilote ou *Master Equation*. Leur différence relève du degré de détails inclus dans chaque modèle, allant du plus précis (AkMC) au plus général (CME). La dynamique d'amas est alors une approche limite de la CME, décrivant une évolution moyenne des états, en l'occurence des concentrations de défauts, plutôt que leur nombre, dans la limite d'un volume infini.

Une autre approche, née dans la communauté physicienne travaillant sur des phénomènes de germination, construit au contraire directement les équations cinétiques de la dynamique d'amas en se basant sur les mécanismes d'émission/absorption : il s'agit de l'approche de Becker–Döring [BD35 ; HY17].

Les équations de Becker–Döring apparaissent en 1935 afin d'étudier la germination dans des milieux sursaturés [BD35]. Les équations ont ensuite été utilisées et popularisées dans

les années 1970 pour décrire des phénomènes de condensation [Bur77]. Aujourd'hui, le modèle de Becker–Döring est utilisé pour décrire des phénomènes de germination, de transition de phase en physique et trouve également des applications dites de coagulation-fragmentation en biologie (de nombreux exemples sont cités dans la revue de Hingant et Yvinec [HY17]).

Les années 1970 ont connu un développement parallèle autour des mêmes équations cinétiques (1.5)–(1.6) mais dans des communautés différentes et qui ont chacune adoptées des terminologies propres. Ainsi, la communauté des matériaux du nucléaire s'est intéressée à l'évolution d'amas de grandes tailles, faisant intervenir des phénomènes de croissance qui allaient au-delà des phénomènes de germination pour lesquels les équations de Becker–Döring ont été introduites. Cet intérêt à des phénomènes à des échelles plus larges, comme la croissance et la coalescence, ont laissé place à ce que Kiritani [Kir73], Ghoniem [GS80] puis d'autres ont appelé la dynamique d'amas. Par ailleurs, les équations de la dynamique d'amas (1.5)–(1.6) correspondent aux équations Becker–Döring avec un forçage, du fait de la présence du terme source $G_1$ caractéristique des phénomènes d'irradiation. Ainsi, les équations de Becker–Döring, telles qu'elles sont considérées dans la littérature, s'écrivent

$$\begin{aligned}\frac{dC_n}{dt} &= \beta_{n-1}C_{n-1}C_1 - (\beta_n C_1 + \alpha_n)C_n + \alpha_{n+1}C_{n+1},\\ \frac{dC_1}{dt} &= -2\beta_1 C_1^2 - \sum_{n\geqslant 2}\beta_n C_n C_1 + \sum_{n\geqslant 2}\alpha_n C_n + \alpha_2 C_2.\end{aligned}$$

Enfin, si les équations générales de la dynamique d'amas (1.7)–(1.8) n'ont jamais été étudiées mathématiquement, les équations de Becker–Döring ont au contraire été abordées par de nombreux mathématiciens. En particulier, leur caractère bien posé est établi dès 1986 par Ball, Carr et Penrose [Bal+86] sous certaines conditions sur les coefficients d'émission et d'absorption, résultat qui a par la suite été généralisé par Laurençot et Mischler [LM02] avec des hypothèses moins contraignantes.

### 1.3.3 Approximations et simulations

Une approche numérique pour résoudre les équations de la dynamique d'amas est d'abord de fixer la taille maximale $N_{\max}$ des amas dans la simulation, ce qui revient à résoudre un système d'ODEs de taille $N_{\max}$. Dans la pratique, pour des matériaux complexes (par exemple du fer avec des amas de défauts comprenant interstitiels, lacunes ou hélium), l'utilisateur fait face à une explosion combinatoire du modèle, le nombre d'équations pouvant dépasser $N_{\max} = 10^{11}$. Différentes stratégies et méthodes ont été développées afin de simuler efficacement l'évolution de tels systèmes, que nous passons en revue ici.

#### Méthode dite de *Grouping*

La méthode de *Grouping* a d'abord été introduite par Kiritani [Kir73] avant d'être enrichie et améliorée au cours des années [Gol+01 ; Ovc+03 ; KW16]. Cette méthode consiste à grouper des amas en différentes classes, chaque classe étant pilotée par une seule équation. De nombreuses difficultés apparaissent lors de simulations utilisant le *Grouping*, comme le choix des différentes classes, souvent ajusté de façon à conserver la masse et certaines moyennes, ou encore l'apparition de concentrations négatives. Pour ces raisons, nous n'avons pas étudié davantage cette méthode.

### *Stochastic Cluster Dynamics* (SCD)

Une approche récente développée par Marian et Bulatov [MB11] tente de contourner l'explosion combinatoire de la dynamique d'amas en retournant à une vision purement stochastique de type CME. Le formalisme permettant d'obtenir les probabilités de réaction est simple, puisqu'il suffit de multiplier les équations de la dynamique d'amas par un volume $V$ fixé et d'écrire les équations sur les quantités de défauts $X_n = C_n V$ pour $n \geqslant 1$. Cette méthode a pour principal avantage de limiter les coûts de simulation définis *a priori* par le volume de la simulation. En plus de la réduction du coût de simulation, les auteurs justifient l'intérêt de la méthode par l'introduction de fluctuations stochastiques reflétant des phénomènes physiques et prenant mieux en compte l'aspect aléatoire de l'irradiation. Toutefois, les modèles de dynamique d'amas sont souvent des problèmes raides où certains phénomènes sont caractérisés par des échelles de temps très variées. Les simulations peuvent être considérablement ralenties par des événements fréquents. La méthode SCD trouve alors sa pertinence pour les simulations de matériaux complexes où l'explosion combinatoire limite considérablement les simulations déterministes.

### L'approximation de Fokker–Planck

La dernière classe d'approches est basée sur l'approximation de Fokker–Planck pour les grandes tailles d'amas [Goo64 ; Wol+77 ; GS80 ; Gho99 ; Jou+14]. L'approximation de Fokker–Planck consiste à décrire l'évolution des concentrations d'amas de grande taille par une seule équation aux dérivées partielles. Sous l'hypothèse que les concentrations varient lentement en fonction de la taille d'amas, on suppose qu'il existe une fonction $\mathscr{C}$ telle que $C_n(t) = \mathscr{C}(t, n)$ et qui vérifie l'équation aux dérivées partielles (EDP) d'advection-diffusion

$$\frac{\partial \mathscr{C}}{\partial t} = -\frac{\partial (F\mathscr{C})}{\partial x} + \frac{1}{2}\frac{\partial^2 (D\mathscr{C})}{\partial x^2}, \tag{1.9}$$

où $F(x) = \beta(x)C_1 - \alpha(x)$ et $D(x) = \beta(x)C_1 + \alpha(x)$. Les fonctions $\alpha$ et $\beta$ sont en fait une version continue des coefficients $\alpha_n$ et $\beta_n$. Il a été montré que l'approximation de Fokker–Planck, utilisée pour simuler l'évolution des concentrations d'amas de grande taille, couplée avec les équations (1.5) et (1.6), produit des résultats en très bon accord avec les solutions obtenues par une simulation du système d'EDOs (1.5)–(1.6) complet [Jou+16]. Par ailleurs, l'approximation permet une simulation efficace et rapide de systèmes physiques dans des conditions réelles [Jou+14].

Si une telle approximation permet de réduire le nombre d'équations à résoudre en utilisant un maillage astucieux pour simuler l'équation de Fokker–Planck, elle n'échappe pas à l'explosion combinatoire due à la complexité de certains systèmes. Ainsi, si les défauts sont de différents types (amas de lacunes, interstitiels, solutés), la dimension de l'EDP augmente en espace ($x$ est par exemple un vecteur de dimension $k$ dans le cas du modèle (1.7)–(1.8)) et devient rapidement impossible à simuler dans des temps raisonnables pour des matériaux avec 3 espèces de défauts ou plus. Par ailleurs, le couplage entre les EDOs (1.7)–(1.8) et l'EDP de Fokker–Planck est difficile à mettre en œuvre dans le cas d'amas mobiles mixtes, c'est-à-dire contenant des défauts de deux espèces différentes (par exemples amas de lacunes et solutés), ce qui limite le champ d'application de la méthode. Enfin, le choix du schéma de discrétisation de l'EDP de Fokker–Planck est crucial dans la mise en œuvre de la méthode, certains schémas permettant une exécution très rapide de

l'algorithme, mais créant artificiellement de la diffusion, quand d'autres, plus précis sont plus coûteux [Jou+16].

### 1.3.4 La limite Lifshitz-Slyozov-Wagner en temps long

L'approximation de Fokker–Planck (1.9) est particulièrement utilisée dans la communauté de la dynamique d'amas. Toutefois un autre modèle limite apparaît naturellement lorsqu'on étudie le comportement en temps longs des équations de Becker–Döring, il s'agit de la limite dite Lifshitz-Slyozov-Wagner ou LSW.

#### La limite LSW : une limite hyperbolique

La limite LSW est obtenue par un scaling espace-temps hyperbolique. En considérant un petit paramètre $\varepsilon > 0$ et la fonction $\mathscr{C}(t,x) \simeq C_{n/\varepsilon}(t/\varepsilon)$, la limite $\mathscr{C}$ est solution de l'équation hyperbolique

$$\frac{\partial \mathscr{C}}{\partial t} + \frac{\partial (F\mathscr{C})}{\partial x} = 0. \tag{1.10}$$

On reconnaît en fait la partie advective de l'equation de Fokker–Planck (1.9) où $F(x) = \beta(x)C_1 - \alpha(x)$. Mathématiquement, et plus précisément, deux approches ont été développées pour faire le lien entre l'équation limite LSW (1.10) et les équations de Becker–Döring. La première, basée sur un rescaling hyperbolique, prouve que les solutions des équations de Becker–Döring convergent au sens des distributions vers une solution de LSW sur un intervalle de temps donné [LM02 ; Vas+02]. La deuxième approche montre que certaines solutions de Becker–Döring sont proches des solutions de LSW en temps long [Pen+78 ; Pen97 ; Nie03 ; Nie04]. Le point essentiel à retenir est la correspondance entre Becker–Döring et LSW pour des grandes tailles d'amas à des temps suffisamment longs.

#### Comparaison des problèmes limites : Fokker–Planck et LSW

La limite LSW répond au besoin de décrire le phénomène universel de coalescence, aussi appelé mûrissement d'Ostwald, dans la limite des temps longs et des amas de grande taille. Cette théorie considère uniquement des systèmes dynamiques conservatifs. Autrement dit, la fraction volumique de la deuxième phase (la concentration totale de lacunes par exemple) est conservée, ce qui signifie que le terme source doit être absent. Jusqu'à présent toutefois, la distribution d'amas correspondant au régime asymptotique LSW n'a jamais été observée à l'échelle des temps d'expérimentation ou d'exploitation des matériaux, pour les problèmes de vieillissement à la fois hors et sous irradiation. Les distributions observées expérimentalement sont systématiquement plus étalées. L'origine de ce désaccord est attribuée à la distribution spatiale des amas et à la dispersion des forces de puits qui en résulte. Cette dispersion est prise en compte dans le Chapitre 4. L'enjeu est davantage de décrire la germination ainsi que le régime de croissance des amas de défauts. Ce régime intermédiaire est insuffisamment décrit par LSW, comme l'illustre le travail de Berthier [Ber+11]. L'approximation de Fokker–Planck prend au contraire tout son sens dans un tel régime.

Sous irradiation, le phénomène est notable. La comparaison entre LSW, Fokker–Planck et les équations de la dynamique d'amas en Figure 1.1 montre l'importance du terme de diffusion qui est de second ordre et négligé dans la limite LSW.

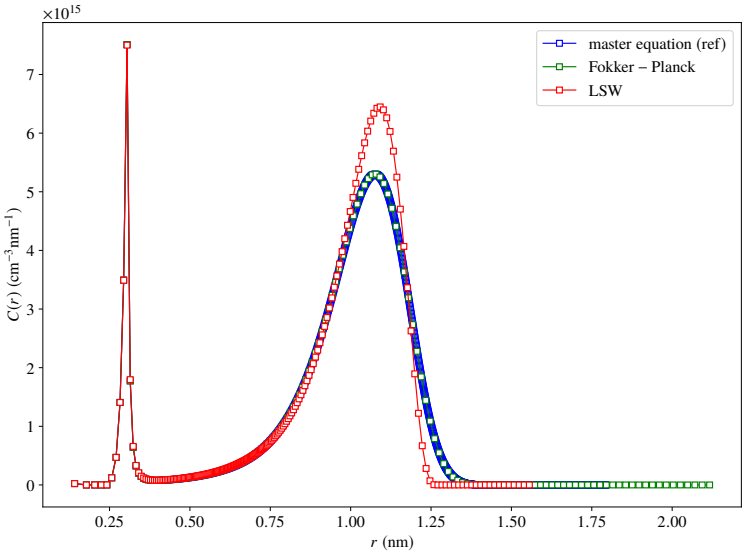



**Figure 1.1:** Simulation sous irradiation de croissance d'amas de lacunes, pour $t = 10^5$ s, à partir des équations de la dynamique d'amas, de l'approximation de Fokker–Planck et de la pseudo-limite LSW. Cette pseudo-limite est obtenue au moyen d'une simulation en supprimant le terme de diffusion d'ordre 2, mais en gardant le terme d'advection et le terme source. La vraie limite LSW ne prend pas en compte ce terme source et a une forme analytique simple.

### 1.3.5 Nouvelles approches hybrides

Les approches purement stochastiques évitent l'explosion combinatoire pour les systèmes complexes mais sont limitées par des événements fréquents tandis que les approches purement déterministes sont très efficaces mais rapidement limitées par l'explosion combinatoire. L'idée de coupler les avantages des deux méthodes a donc été développée ces dernières années [Ghe+12 ; Sur+04].

Dans le travail [Sur+04], l'idée est d'utiliser les équations de la dynamique d'amas (1.7)–(1.8) pour simuler l'évolution des concentrations des amas de petites tailles et d'utiliser l'approximation de Fokker–Planck (1.9) pour les amas de grande taille *via* une représentation probabiliste sous la forme d'un processus de Langevin. Ainsi, à partir d'une certaine taille d'amas, et à chaque pas de temps de la simulation, les concentrations d'une taille d'amas limite sont transformées en une particule stochastique qui évolue selon le processus de Langevin. Ainsi, la précision de la méthode est en particulier limitée par le nombre de particules stochastiques qui sont émises et qui correspondent au nombre de pas de temps au cours d'une simulation. Le choix de pas de temps petits est alors limitant pour atteindre des échelles de temps longs.

Un travail plus récent [Ghe+12] se base sur une approche de type CME et intègre l'évolution des petits amas en considérant l'évolution de leur concentration de manière déterministe. Cette méthode permet ainsi d'éviter le calcul d'événements fréquents dus à l'évolution des petits amas. Elle est toutefois limitée par le fait qu'une approche purement CME/SSA des gros amas est moins efficace qu'une approche Langevin basée sur l'approximation de Fokker–Planck à partir d'une certaine taille. Par ailleurs, dans [Ghe+12], la

partie stochastique est basée sur un algorithme séquentiel. A chaque cycle SSA, on doit déterminer quel amas a réagi. Cette façon de procéder est moins efficace qu'une approche parallèle que nous allons adopter par la suite.

## 1.4 Contributions principales

Si la dynamique d'amas (1.5)–(1.6) est maintenant utilisée depuis une soixantaine d'année dans la communauté des sciences des matériaux et particulièrement des matériaux irradiés, son étude mathématique a uniquement été conduite sous l'angle des équations de Becker-Döring. En particulier, une analyse numérique permettant de coupler la dynamique d'amas et ses approximations aux grandes taille n'a jamais été étudiée à ma connaissance. Par ailleurs, l'approximation de Fokker–Planck est encore parfois controversée [KW16]. Mon travail de thèse a été l'occasion de mener une étude transversale allant des propriétés mathématiques fines de la dynamique d'amas jusqu'aux applications physiques pouvant enrichir les connaissances de la communauté physicienne. Ce travail est organisé en trois parties, résumées dans les trois sections suivantes.

### 1.4.1 Analyse mathématique de la dynamique d'amas et de l'approximation de Fokker–Planck

Dans une première partie (Chapitre 2), on étudie le caractère bien posé de la dynamique d'amas ainsi que la pertinence de l'approximation de Fokker–Planck. Plus particulièrement, on précise d'abord le cadre d'existence de solutions classiques en temps longs pour la dynamique (1.5)–(1.6). On prouve en particulier que les solutions de la dynamique existent dans un sous-ensemble $\mathcal{Q}$ de l'espace des suites de carrés sommables $\ell^2(\mathbb{N}^*, \mathbb{R})$, tel que

$$\mathcal{Q} = \left\{ u \in \ell^2(\mathbb{N}^*, \mathbb{R}) \;\middle|\; \forall n \geqslant 1, \quad u_n \geqslant 0 \quad \text{and} \quad \sum_{n \geqslant 1} n u_n < +\infty \right\}.$$

On introduit ensuite une décomposition de la dynamique, motivé par deux raisons. La décomposition est d'abord un élément crucial dans l'introduction d'un algorithme de simulation numérique hybride déterministe/stochastique puisqu'il permet de découpler la non-linéarité de la dynamique sur les lacunes (1.6), du reste de la dynamique (1.5) sur les concentrations d'amas de taille plus grande que 2, qui est alors linéaire. Cette décomposition permet également d'exploiter la linéarité du reste de la dynamique afin d'étudier l'approximation de Fokker–Planck dans ce cadre.

L'étude de l'approximation de Fokker–Planck fait l'objet de la seconde partie du Chapitre 2. On rappelle que cette approximation découle d'une dérivation heuristique qui demande une étude mathématique pour être rendue rigoureuse. Plus précisément, comme cette approximation est obtenue par un développement de Taylor à l'ordre 2 pour un pas d'espace $\Delta x = 1$, il est nécessaire de contrôler la taille des termes de reste. On prouve que la dynamique d'amas et son approximation de Fokker–Planck (1.9) sont liées à un changement de variable près à une équation de diffusion pure

$$\frac{\partial \mathfrak{C}}{\partial t} = \frac{1}{2}\sigma^2(q)\frac{\partial^2 \mathfrak{C}}{\partial q^2}.$$

La particularité de cette équation de diffusion est que le coefficient de diffusion $\sigma^2$ est

non-homogène en espace et non-borné, mais de croissance sous-linéaire. On peut alors, à l'aide d'outils stochastiques et de formules de représentations probabilistes, contrôler la décroissance des dérivées des solutions de cette équation de diffusion. Ces estimées de décroissance nous permettent ensuite de contrôler les termes d'erreurs qui relient cette équation de diffusion à la dynamique d'amas et son approximation de Fokker–Planck. Une étude numérique vient ensuite compléter l'analyse théorique et confirmer la validité de l'approximation.

### 1.4.2 Présentation d'un nouvel algorithme hybride déterministe/stochastique

On présente au Chapitre 3 un nouvel algorithme de simulation pour la dynamique d'amas. Cet algorithme combine simulations déterministes et simulations stochastiques dans le but de pallier les différents inconvénients des approches purement déterministes ou purement stochastiques. La méthode est basée sur le splitting introduit au Chapitre 2 et utilise ensuite la linéarité de la dynamique (1.5) à $C_1$ fixé pour décomposer l'évolution des amas de petite taille de ceux de grande taille. Cette décomposition permet de traiter chaque classe d'amas avec des méthodes dédiées. Ainsi, l'évolution des concentrations des petits amas sera faite de manière déterministe en simulant les équations (1.5). Pour les grands amas, des méthodes stochastiques sont utilisées.

On présente en particulier deux approches stochastiques. La première consiste à considérer la dynamique, dorénavant linéaire, sur les amas de grande taille, comme une équation de Kolmogorov première d'un processus de Markov. C'est en fait le formalisme sous-jacent aux méthodes cinétiques (kMC, CME) et plus particulièrement à la vision adoptée dans [MB11]. La deuxième approche est quant à elle basée sur l'approximation de Fokker–Planck utilisée dans les méthodes déterministes [Jou+14]. L'EDP (1.9) est en effet liée à un processus stochastique $(X_t)_{t\geqslant 0}$, dit processus de Langevin, tel que

$$dX_t = F(X_t)dt + \sqrt{D(X_t)}dW_t,$$

où $(W_t)_{t\geqslant 0}$ est un mouvement brownien standard. La loi d'un tel processus est en fait solution de l'équation de Fokker–Planck. L'intérêt des deux approches stochastiques est double. D'une part, le passage en stochastique limite naturellement le problème d'explosion combinatoire puisqu'on contrôle dorénavant la complexité du modèle *via* le nombre de particules stochastiques qui sont simulées. D'autre part, du fait de la linéarité des équations à $C_1$ fixé (ce qui apparaît naturellement avec le splitting de la dynamique), la parallélisation des méthodes fondée sur des réalisations indépendantes des processus est immédiate et particulièrement efficace.

L'utilisation du couplage déterministe/stochastique permet par ailleurs de limiter les inconvénients d'approches purement déterministes ou stochastiques. En effet, la limitation des approches déterministes réside principalement dans l'explosion combinatoire qui est *de facto* supprimée dans une telle approche hybride. Par ailleurs, la limitation liée aux événements fréquents des approches purement stochastiques est également contournée *via* la résolution déterministe de l'évolution des concentrations des petits amas. En effet, on observe que les événements fréquents sont principalement liés au comportement des petits amas dont les fréquences de saut sont très élevées. Un choix de frontière déterministe/stochastique judicieux est donc nécessaire pour trouver un équilibre entre la

réduction du nombre d'EDOs et la limitation des approches stochastiques aux événements peu fréquents.

On illustre sur un cas de croissance d'amas de lacunes les résultats de cet algorithme hybride. On observe ainsi que les distributions de concentrations obtenues par notre algorithme de couplage sont très proches de celles obtenues par une résolution numérique du système d'EDOs (1.5)–(1.6). En particulier l'erreur liée au splitting est d'ordre 1 en le pas de temps quand celle liée à l'approche stochastique décroit en $1/\sqrt{N_{\text{part}}}$ où $N_{\text{part}}$ est le nombre de particules utilisées pour les simulations stochastiques. Enfin, on observe une bonne scalabilité de la parallélisation de la partie stochastique.

### 1.4.3 Présentation de résultats numériques pour des modèles physiques et amélioration du modèle de dynamique d'amas

Dans une dernière partie (Chapitre 4), on présente les applications de l'algorithme hybride à des modèles d'intérêts physiques et industriels. Le Commissariat à l'énergie atomique et aux énergies alternatives (CEA) et EDF ont en effet développé un code de dynamique d'amas — appelé CRESCENDO [Jou+14] — dans le but de prédire l'évolution des défauts dans des matériaux d'intérêts soumis à l'irradiation. Ce code est basé sur l'approximation de Fokker–Planck [Jou+14] et est donc limité dans ses applications. On a alors implémenté l'algorithme hybride avec quelques modifications dues à des contraintes techniques particulières et on a également apporté différentes améliorations au cours du développement.

On étudie dans un premier temps deux exemples de matériaux irradiés. Le premier est du fer irradié aux neutrons dans des conditions de fonctionnement de réacteurs et sur des temps longs (de l'ordre de la dizaine d'années). Le deuxième matériau est du fer irradié aux ions hélium, ce qui vient créer des défauts d'amas de deux espèces différentes et illustre un cas d'application en dimension 2. On montre en particulier que si l'approche purement déterministe est très efficace en dimension 1, l'approche hybride devient compétitive en dimension 2.

Enfin, nous montrons comment améliorer l'accord entre les simulations de dynamique d'amas et mes simulations OkMC qui tiennent compte de l'hétérogénéité spatiale et qui en particulier prennent mieux en compte la dispersion des distances entre amas. Il a été observé numériquement en OkMC une dispersion des coefficients effectifs d'absorption et d'émission qui explique les désaccords avec la dynamique d'amas. Nous montrons comment prendre en compte une telle dispersion dans la dynamique d'amas en utilisant l'approche hybride déterministe/stochastique. Cette approche permet de réaliser efficacement des simulations qui prendraient beaucoup plus de temps *via* une approche purement déterministe. Les résultats présentés sont plus fidèles aux modèles aux échelles inférieures de type OkMC.

# 2 A mathematical analysis of the Fokker–Planck approximation for Cluster Dynamics

*This chapter is mainly based on the article* "A mathematical analysis of the Fokker–Planck approximation for Cluster Dynamics" [arxiv:1810:01462].

## 2.1 Introduction

Simulating the ageing of materials over a long period of time remains a challenge in the materials science community. Purely atomistic approaches, such as molecular dynamics or kinetic Monte Carlo [Bor+75; Isi25; Soi+10; Vot07; YE66] do not allow to reach times as long as years of ageing. To achieve this goal, mean-field models have been developed. One model, called Cluster Dynamics (CD), has been considered in the community of nuclear materials [BC07; GS80; Jou+14] in order to study the evolution of defects under irradiation. It consists in simulating the evolution of concentration of clusters of defects such as vacancies, other self defects or solute gas. From a mathematical viewpoint, CD is an infinite set of ordinary differential equations (ODEs), one for each type of defect. We focus in this work on a simple but paradigmatic example of vacancy clustering.

Let us present the equations used to describe the time evolution of concentrations. Denote by $C_n$ the concentration of clusters composed of $n$ vacancies. The evolution of $C_n$ is given by

$$\frac{dC_n}{dt} = \beta_{n-1}C_{n-1}C_1 - (\beta_n C_1 + \alpha_n)C_n + \alpha_{n+1}C_{n+1}, \tag{2.1}$$

where $\alpha_n$ and $\beta_n$ are respectively called emission and absorption coefficients, while $C_1$ represents the concentration of clusters composed of only one vacancy. Equation (2.1) describes a simple process where clusters of defects can either emit or absorb a single vacancy. More precisely, the term $\beta_{n-1}C_{n-1}C_1$ describes the increase in the population of clusters of size $n$ coming from clusters of size $n-1$ absorbing a single vacancy, while the term $\alpha_{n+1}C_{n+1}$ describes the increase in the population of clusters of size $n$ coming from clusters of size $n+1$ emitting a single vacancy. Finally, the term $-(\beta_n C_1 + \alpha_n)C_n$ encodes the rate of decrease of clusters of size $n$ arising from their transformation into clusters of sizes $n-1$ or $n+1$. Single vacancies are considered as mobile clusters and their evolution is therefore related to the evolution of all other clusters as follows:

$$\frac{dC_1}{dt} = -2\beta_1 C_1^2 - \sum_{n\geqslant 2} \beta_n C_n C_1 + \sum_{n\geqslant 2} \alpha_n C_n + \alpha_2 C_2. \tag{2.2}$$

The latter equation is determined by the requirement that the total quantity of matter is conserved, namely

$$\frac{d}{dt}\left(C_1 + \sum_{n\geqslant 2} nC_n\right) = 0.$$

The reasons such a model is a simplification of complex phenomena occurring in real materials are twofold:

1. First, mobile clusters can be of size greater than one. Equation (2.1) can be enriched with terms describing the absorption or emission of clusters of sizes $m \geqslant 1$, with equations similar to (2.2) describing the evolution of concentrations of sizes $m \geqslant 1$.
2. Second, clusters can be made of different types of defects, *e.g.* vacancies and helium atoms in iron. Therefore, defect concentrations are in general indexed by $k$-tuples, where $k$ is the number of types of defects.

A numerical approach to solve CD is to fix the maximal size $N_{\max}$ of the clusters in the simulation, which amounts to solving a system of $N_{\max}$ ODEs. In practice, the ODE system is stiff so that dedicated solvers are required [Jou+14]. It can however become computationally impossible to solve the ODEs. For example, a system with various type of defects such as vacancies (V) and helium atoms (He), might contain up to $N_{\max} = N_{\max,\mathrm{V}} N_{\max,\mathrm{He}} = 10^6 \times 10^5$ equations. This motivated the development of an approximate model based on a Fokker–Planck approximation. This approximation was first presented in [Goo64] and further developed and used in more recent works [Jou+14; Jou+16; Ter+17]. Assume that the concentrations vary slowly with time and cluster sizes, so that $C_n(t) \simeq \mathscr{C}(t,n)$ for some smooth function $(t,x) \mapsto \mathscr{C}(t,x)$ depending on the physical time $t$ and a spatial variable $x$. In fact, $\partial_x \mathscr{C}(t,n) \simeq C_{n+1}(t) - C_n(t)$. Then, for large cluster sizes, the system of ODEs reduces to a single partial differential equation (PDE), of advection-diffusion type (we recall the heuristic derivation of this equation in Section 2.3.1):

$$\frac{\partial \mathscr{C}}{\partial t} = -\frac{\partial (F\mathscr{C})}{\partial x} + \frac{1}{2}\frac{\partial^2 (D\mathscr{C})}{\partial x^2}. \tag{2.3}$$

The Fokker–Planck equation (2.3) is characterized by the drift $F$ and the diffusion $D$, both coefficients depending on the coefficients of absorption and emission $\beta$ and $\alpha$. While this approximation gives accurate results in practice (when compared to the solution of the full ODE system), the consistency of this approach has never been rigorously proven, and the approximation error never been quantified in function of the minimal size of the clusters for which it is used.

A difficulty in making the heuristic argument of [Goo64] rigorous is that the approximation on which the derivation relies is based on a Taylor expansion of order 2 for a mesh with fixed spacing 1. Therefore, in order to prove the validity of this approach, we rely on another equation, namely a diffusion equation of the form

$$\frac{\partial \mathfrak{C}}{\partial t} = \frac{1}{2}\sigma^2(q)\frac{\partial^2 \mathfrak{C}}{\partial q^2}, \tag{2.4}$$

where the diffusion coefficient $\sigma^2$ depends on the coefficients $\alpha$ and $\beta$. The interest of the diffusion equation is that it is possible to make precise the decay of the derivatives of $\mathfrak{C}$. The PDE (2.4) allows us to relate CD and its Fokker–Planck approximation. Due to the fact that CD is inherently nonlinear, we also introduce a splitting of the dynamics in order to restrict the nonlinearity to the evolution (2.2) of the single vacancies. This splitting allows us to work in a simpler framework and to prove rigorously the link between the Fokker–Planck approximation and CD. Finally, this splitting is also of interest for numerical simulations [Ter+17].

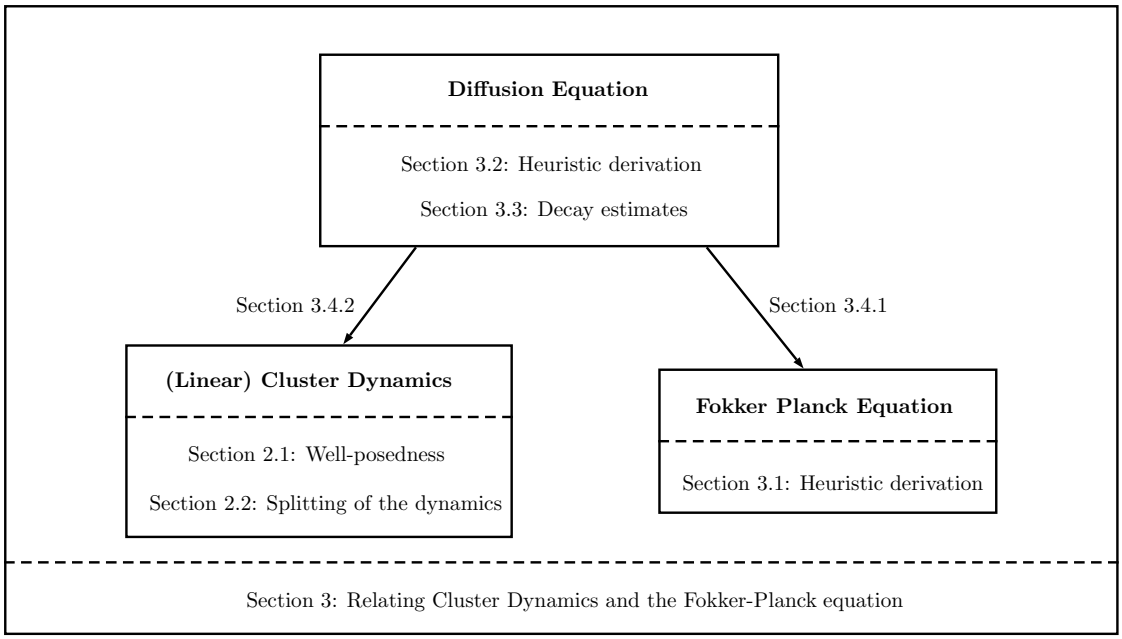



**Fig. 2.1:** Structure of this Chapter. The diffusion equation and the decay estimates of the spatial derivatives of its solutions (proved in Section 2.3) are crucial to relate Cluster Dynamics and its Fokker–Planck approximation (Section 2.3). The well-posedness of Cluster Dynamics is presented in Section 2.2.

This article is organized as follows. We first present results concerning CD in Section 2.2. We start by proving in Section 2.2.1 that the nonlinear CD is well-posed from a mathematical viewpoint using a regularized version of CD, standard techniques from the theory of semigroups of operators and fixed point theorems. We next discuss in Section 2.2.2 the convergence of the splitted dynamics. The proofs of the results given in Section 2.2 are postponed to Appendix 2.A and 2.B. We then focus our attention on the Fokker–Planck approximation in Section 2.3. After a heuristic derivation of the Fokker–Planck approximation as well as the diffusion equation (2.4), we state decay results on the solutions of this equation, which allows us to relate the diffusion equation and both the Fokker–Planck equation (2.3) and Cluster Dynamics (2.1). The proofs of the technical results of Section 2.3 are gathered in Appendix 2.C. Figure 2.1 summarizes the organization of this work and highlights that the diffusion equation (2.4) is a key feature in this article.

## 2.2 Well-posedness of Cluster Dynamics

We make precise in Section 2.2.1 the mathematical framework in which CD is well posed. We next introduce in Section 2.2.2 a splitting of the dynamics and prove that it is consistent of order 1.

### 2.2.1 Full Cluster Dynamics

We consider the full CD (*i.e.* Equations (2.1)–(2.2)), which is a nonlinear dynamics. Its well-posedness can be proved with the approach described in [Paz12, Chapter 6] for some regularized dynamics, and an appropriate passage to the limit. We work on the Hilbert space $\mathcal{H} = \ell^2(\mathbb{N}^*, \mathbb{R})$, endowed with its natural norm $\|\cdot\|$ and inner product $\langle \cdot, \cdot \rangle$. Let us note that such a problem has already been studied by Ball, Carr and Penrose [Bal+86], with

further complements by Laurençot and Mischler [LM02]. The proofs given in Section 2.A serve a pedagogical purpose.

**Remark 1.** *Unless stated otherwise, the norm $\|\cdot\|$ is the natural norm of $\mathcal{H}$ or the norm of bounded operators on $\mathcal{H}$, depending on the context.*

Consider $u = (u_1, u_2, \cdots, u_n, \cdots) \in \mathcal{H}$ and denote by $(e_n)_{n\in\mathbb{N}^*}$ the orthonormal basis of $\mathcal{H}$ defined by $(e_n)_i = \delta_{ni}$, where $\delta_{ni}$ is the usual Kronecker symbol. In particular, $\langle u, \mathrm{e}_i\rangle = u_i$. The full CD can be written with this notation as the following Cauchy problem in $\mathcal{H}$:

$$\begin{cases} \dfrac{du}{dt} = A(u_1)u, \\ u(0) = u^0, \end{cases} \tag{CD}$$

where the quasi-linear operator $A$ is defined, for all $v \in \mathcal{H}$, by:

$$\begin{aligned} A(v_1)e_1 &= -2\beta_1 v_1 e_1 + \beta_1 v_1 e_2, \\ A(v_1)e_n &= (\alpha_n - \beta_n v_1)e_1 + \alpha_n e_{n-1} - (\beta_n v_1 + \alpha_n)e_n + \beta_n v_1 e_{n+1}, \qquad n \geqslant 2. \end{aligned}$$

Alternatively, $A$ can be written as the following infinite matrix:

$$A(v_1) = \begin{pmatrix} -2\beta_1 v_1 & 2\alpha_2 - \beta_2 v_1 & \alpha_3 - \beta_3 v_1 & \alpha_4 - \beta_4 v_1 & \cdots \\ \beta_1 v_1 & -(\beta_2 v_1 + \alpha_2) & \alpha_3 & 0 & \cdots \\ 0 & \beta_2 v_1 & -(\beta_3 v_1 + \alpha_3) & \alpha_4 & \cdots \\ 0 & 0 & \beta_3 v_1 & -(\beta_4 v_1 + \alpha_4) & \ddots \\ \vdots & \vdots & \vdots & \ddots & \ddots \end{pmatrix}.$$

The main difficulty of the problem (CD) comes from the unboundedness of the coefficients $\alpha_n$ and $\beta_n$. As will be made clear below, it is nonetheless possible to obtain existence and uniqueness results for coefficients which do not grow too fast. In fact, physically,

$$\alpha_n, \beta_n = \mathrm{O}(n^\gamma),$$

with $\gamma = 1/3$ for vacancies and solutes which generate three-dimensional objects such as bubbles, and $\gamma = 1/2$ for interstitials which generate two-dimensional objects such as loops (see Example 1 below). Nevertheless, in order to prove the existence and uniqueness of solutions, we only require the following assumptions on $\alpha$ and $\beta$.

**Assumption 1.** *The sequences $\alpha = (\alpha_n)_{n\geqslant 1}$ and $\beta = (\beta_n)_{n\geqslant 1}$ are two sequences of positive real numbers. Moreover, defining $R_n^\alpha = \alpha_{n+1} - \alpha_n$ and $R_n^\beta = \beta_{n+1} - \beta_n$, there exists $B \in \mathbb{R}_+$ such that*

$$\forall\, n \geqslant 1, \qquad |R_n^\alpha|, \left|R_n^\beta\right| \leqslant B.$$

Let us give a specific example, which we will use throughout this work to illustrate the relevance of our assumptions.

**Example 1.** *In many physical models [BC07; GS80; Jou+14; Ovc+03], the expression of $\alpha_n$*

*and $\beta_n$ are chosen as follows:*

$$\beta_n = \left(\frac{48\pi^2}{V_{\mathrm{at}}^2}\right)^{\gamma} D r_n, \tag{2.5}$$

*and*

$$\alpha_n = \left(\frac{48\pi^2}{V_{\mathrm{at}}^2}\right)^{\gamma} D r_n \exp\left(-\frac{E_{\mathrm{v}}^{\mathrm{f}}}{k_{\mathrm{B}}T}\right) \exp\left(\frac{\omega}{r_n}\right), \tag{2.6}$$

*where $D$ is the diffusion coefficient of mobile clusters, $V_{\mathrm{at}}$ the atomic volume, $E_{\mathrm{v}}^{\mathrm{f}}$ the formation energy of a vacancy, $k_{\mathrm{B}}$ the Boltzmann constant, $T$ the temperature, $\omega$ a parameter related to the type of clusters (vacancies or interstitials) and $r_n = n^{\gamma}$, with $\gamma \in \{\frac{1}{3}, \frac{1}{2}\}$. It is easy to see that the sequences $\alpha$ and $\beta$ indeed satisfy Assumption 1.*

Let us now define the linear operators $A_{\mathrm{L}}^{\alpha}$ and $A_{\mathrm{L}}^{\beta}$ (where L stands for "linear") as:

$$A_{\mathrm{L}}^{\alpha} e_1 = 0, \qquad A_{\mathrm{L}}^{\alpha} e_n = \alpha_n (e_1 + e_{n-1} - e_n), \tag{2.7}$$

and

$$A_{\mathrm{L}}^{\beta} e_1 = -\beta_1 (2e_1 - e_2), \qquad A_{\mathrm{L}}^{\beta} e_n = -\beta_n (e_1 + e_n - e_{n+1}). \tag{2.8}$$

Note that $A(v_1) = A_{\mathrm{L}}^{\alpha} + v_1 A_{\mathrm{L}}^{\beta}$. In order to prove the global-in-time well-posedness, we introduce the following convex subset of $\mathcal{H}$:

$$\mathcal{Q} = \left\{ u \in \mathcal{H}_+ \;\middle|\; \sum_{n \geqslant 1} n u_n < +\infty \right\}, \tag{2.9}$$

where $\mathcal{H}_+ = \{u \in \mathcal{H} \,|\, \forall\, n \geqslant 1,\ u_n \geqslant 0\}$ is the subset of elements of $\mathcal{H}$ whose components are non-negative. The condition $\sum n u_n < +\infty$ translates the physical fact that the total quantity of matter is finite. In fact, as we will see, it is conserved by the CD dynamics (2.1)–(2.2). For any element $u \in \mathcal{Q}$, we define

$$Q(u) = \sum_{n \geqslant 1} n u_n.$$

**Remark 2.** *In view of* (2.9) *and Assumption 1, an element $u \in \mathcal{Q}$ satisfies in particular*

$$0 \leqslant \sum_{n \geqslant 1} \alpha_n u_n < +\infty \qquad \textit{and} \qquad 0 \leqslant \sum_{n \geqslant 1} \beta_n u_n < +\infty.$$

*This means that the sequences $(\alpha_n u_n)_{n \geqslant 1}$ and $(\beta_n u_n)_{n \geqslant 1}$ are in $\ell^1(\mathbb{N}^*, \mathbb{R})$ when $u \in \mathcal{Q}$.*

The main result of this section is the following.

**Theorem 3.** *Fix an initial condition $u^0 \in \mathcal{Q}$ such that $u_1^0 > 0$ and suppose that Assumption 1 holds true. Then, there exists a unique global-in-time classical solution $u \in \mathcal{C}^0(\mathbb{R}_+, \mathcal{Q}) \cap \mathcal{C}^1(\mathbb{R}_+, \mathcal{H})$ of the problem* (CD)*. Moreover,*

$$\forall\, t \geqslant 0, \qquad Q(u(t)) = Q\left(u^0\right) \qquad \textit{and} \qquad \|u(t)\| \leqslant \frac{\pi Q(u^0)}{\sqrt{6}}.$$

The proof of this result can be read in Appendix 2.A. In order to prove it, we first consider a regularized version of (CD) and prove the existence and uniqueness of a solution to such a problem using a standard fixed point argument on the integral form of the equation (see Appendix 2.A.1). We then use an argument based on the Arzelà-Ascoli theorem to prove

the existence of a solution to (CD) in Appendix 2.A.2. Finally we prove the uniqueness of such a solution. The argument we use requires in particular to prove that, for any given non-negative function $b \in \mathcal{C}^0(\mathbb{R}_+)$, the family of linear operators $(A(b(t)))_{t\geqslant 0}$ is a family of infinitesimal generators of strongly continuous semigroups. This is made precise in Appendices 2.A.3.

### 2.2.2 Splitting of the dynamics and qualitative properties

We discuss in this section some properties of the dynamics obtained by splitting the nonlinear dynamics (CD) into two sub-dynamics, one on the first concentration only and another one on the remaining concentrations. The motivations for considering the properties of this splitting are twofold:

1. It allows to restrict the nonlinearity to one equation, while the remainder of the dynamics becomes linear. It is one of the key features we used in [Ter+17] for an efficient numerical integration of cluster dynamics. The algorithm we developed indeed relies on the linearity of the sub-dynamics describing the evolution of clusters of sizes greater than 2.
2. Moreover, the validity of the Fokker–Planck approximation is proved only for linear dynamics. The proof of such an approximation in Section 2.3 is performed for the linear sub-dynamics of clusters of larger sizes.

Note that the splitting introduced in this section can be generalized to a splitting on a first dynamics on small clusters from sizes 1 to $M$, and on a second dynamics on larger clusters of sizes greater than $M+1$ for some $M \geqslant 1$.

The splitting we consider here is a simple Lie-Trotter splitting [Tro59]. We prove that the associated dynamics is consistent with the full dynamics (CD). This however requires strengthening Assumption 1 in order to control the first and second derivatives of the solution, a property which is crucial for our estimates.

**Assumption 2.** *There exist* $0 \leqslant \gamma \leqslant 1/2$ *and* $K \in \mathbb{R}_+^*$ *such that*

$$\forall\, n \geqslant 1, \qquad 0 \leqslant \alpha_n, \beta_n \leqslant K n^\gamma.$$

Note that Assumption 2 clearly holds true for Example 1 (see (2.5)–(2.6)). We now consider the following splitted dynamics. The sub-dynamics for the first concentration only reads

$$\begin{cases} \dfrac{du_1}{dt} = -2\beta_1 u_1^2 - \left(\displaystyle\sum_{n\geqslant 2} \beta_n u_n\right) u_1 + \left(\sum_{n\geqslant 2} \alpha_n u_n + \alpha_2 u_2\right), \\ \dfrac{du_n}{dt} = 0, \qquad \forall n \geqslant 2, \end{cases} \tag{2.10}$$

*i.e.* $u_n$ is fixed for $n \geqslant 2$. We denote by $\varphi_t^{(u_2,\cdots)}$ the flow of this dynamics, or simply $\varphi_t$ when the dependence is clear. Note that this sub-dynamics is well-posed (see (2.45) in Appendix 2.B.3). The second sub-dynamics, for the remaining concentrations, reads

$$\begin{cases} \dfrac{du_1}{dt} = 0, \\ \dfrac{du_n}{dt} = \beta_{n-1} u_1 u_{n-1} - (\beta_n u_1 + \alpha_n) u_n + \alpha_{n+1} u_{n+1}, \qquad \forall n \geqslant 2, \end{cases} \tag{2.11}$$

*i.e.* $u_1$ is fixed. We denote by $\chi_t^{u_1}$ the flow of this dynamics, or simply $\chi_t$ when the dependence is clear. This sub-dynamics is well-posed (see Propsition 29 in Appendix 2.B.1). One step of the splitted dynamics is encoded by the mapping $u^{k+1} = S_{\Delta t}\left(u^k\right)$ for a given time step $\Delta t > 0$, defined as

1. update the first concentration as $u_1^{k+1} = \varphi_{\Delta t}^{(u_2^k, \cdots)}(u_1^k)$,
2. update the remaining concentrations as $\left(u_2^{k+1}, \cdots\right) = \chi_{\Delta t}^{u_1^{k+1}}\left(u_2^k, \cdots\right)$.

The iterates defined as $u^k = S_{\Delta t}(u^{k-1})$ for $k \geqslant 1$ and some initial condition $u^0$ are an approximation of $u(k\Delta t)$, the solution of (CD) with initial condition $u^0$ at time $k\Delta t$. The following proposition states that the Lie-Trotter splitting is consistent of order 1.

**Proposition 4.** *Fix an initial condition $u^0 \in \mathcal{Q}$ with $u_1^0 > 0$ and a time $\tau > 0$. Suppose that Assumptions 1 and 2 hold true. Suppose also that there exists $k \geqslant 2$ such that $u_k^0 > 0$. Then, there exist a constant $K(\tau, u^0) \in \mathbb{R}_+$ and a time step $\Delta t_* > 0$ such that*

$$\forall\, \Delta t \in (0, \Delta t_*], \qquad \forall\, 0 \leqslant n \leqslant \frac{\tau}{\Delta t}, \qquad \|u(n\Delta t) - u^n\| \leqslant K\left(\tau, u^0\right)\Delta t.$$

The assumption that the initial condition is non-zero in the sub-domain $[2, +\infty)$ ensures that the subdynamics (2.10) are well posed (see Lemma 33). In order to prove Proposition 4, we need some control over the first and second derivatives of the solution of (CD).

**Proposition 5.** *Suppose that Assumptions 1 and 2 hold true. Let $u$ be the classical solution of problem* (CD) *with initial condition $u^0 \in \mathcal{Q}$ and $u_1^0 > 0$. Fix a time $\tau > 0$, a constant $Q_* \in \mathbb{R}_+$ and suppose that $Q(u^0) \leqslant Q_*$. Then, $u \in \mathcal{C}^2([0, \tau], \mathcal{H})$ and there exists $R(Q_*) \in \mathbb{R}_+$ such that, for all $0 \leqslant t \leqslant \tau$,*

$$\left\|\frac{du}{dt}(t)\right\|, \left\|\frac{d^2u}{dt^2}(t)\right\| \leqslant R(Q_*).$$

In the following section, we will prove that the linear problem (2.11) is related, up to a small error, to a diffusion equation, which is the key equation to relate the Fokker–Planck approximation and the Cluster Dynamics.

## 2.3 The Fokker–Planck approximation in the linear case

The Fokker–Planck approximation (2.3) is widely used in the materials science community to approximate the dynamics of large size clusters [Goo64; Wol+77]. This approximation gives accurate results in very good agreement with CD when sufficiently precise numerical schemes are used for the simulation [Jou+16]. It proves to be efficient and speeds up the simulations for complex systems [Jou+14]. We also report a very good agreement between the solution of the exact CD and a coupling approach solving the ODEs for small size clusters and the Fokker–Planck PDE for large size ones [Ter+17]. Nevertheless, the agreement between CD and its Fokker–Planck approximation has never been quantified to our knowledge. We provide in this section a proof of the correctness of the Fokker–Planck limit using stochastic techniques, and quantify the approximation error.

In the whole section, the concentration $C_1$ of single vacancies is supposed to be fixed. The section is organized as follows. We first present a formal derivation of the Fokker–Planck approximation in Section 2.3.1. We next heuristically derive a reformulation of the Fokker–Planck approximation in the form of a diffusion equation (see Section 2.3.2), for which we state a key result on the decay of the spatial derivatives of the solution in Section 2.3.3. Using this result allows us to rigorously establish the link between the Fokker–Planck approximation and Cluster Dynamics, and to quantify errors as a function of the minimal cluster size (see Section 2.3.4).

### 2.3.1 Heuristic derivation of the Fokker–Planck approximation

We describe here the derivation of the Fokker–Planck approximation as presented in the materials science community, pointing out the parts of the argument which require a more rigorous mathematical analysis. Let us emphasize that all computations presented here are formal.

Define the regular mesh $(x_n)_{n\geqslant 0}$ of $[0,+\infty)$ by $x_n = n$. The mesh size is $\Delta x = 1$. Let us assume that there exist smooth functions $\alpha, \beta \in \mathcal{C}^\infty(\mathbb{R}_+, \mathbb{R}_+)$ and $\mathscr{C} \in \mathcal{C}^\infty(\mathbb{R}_+ \times \mathbb{R}_+, \mathbb{R}_+)$ such that, for all $n \in \mathbb{N}^*$ and $t \geqslant 0$,

$$\mathscr{C}(t, x_n) = C_n(t), \qquad \alpha(x_n) = \alpha_n, \qquad \beta(x_n) = \beta_n.$$

When $C_1$ is fixed, Equation (2.1) can be written, for $n \geqslant 2$, as

$$\frac{\partial \mathscr{C}}{\partial t}(t, x_n) = C_1[\beta(x_{n-1})\mathscr{C}(t, x_{n-1}) - \beta(x_n)\mathscr{C}(t, x_n)] + \alpha(x_{n+1})\mathscr{C}(t, x_{n+1}) - \alpha(x_n)\mathscr{C}(t, x_n).$$

Let us emphasize that fixing $C_1$ is crucial for the argument. In practice, this arises through the splitting described in Section 2.2.2. By a Taylor expansion at order 2,

$$\beta(x_{n-1})\mathscr{C}(t, x_{n-1}) \simeq \beta(x_n)\mathscr{C}(t, x_n) - \Delta x \frac{\partial}{\partial x}(\beta\mathscr{C})(t, x_n) + \frac{1}{2}(\Delta x)^2 \frac{\partial^2}{\partial x^2}(\beta\mathscr{C})(t, x_n).$$

Similarly,

$$\alpha(x_{n+1})\mathscr{C}(t, x_{n+1}) \simeq \alpha(x_n)\mathscr{C}(t, x_n) + \Delta x \frac{\partial}{\partial x}(\alpha\mathscr{C})(t, x_n) + \frac{1}{2}(\Delta x)^2 \frac{\partial^2}{\partial x^2}(\alpha\mathscr{C})(t, x_n).$$

Since $\Delta x = 1$, we finally obtain

$$\frac{\partial \mathscr{C}}{\partial t}(t, x) \simeq -\frac{\partial (F\mathscr{C})}{\partial x}(t, x) + \frac{1}{2}\frac{\partial^2 (D\mathscr{C})}{\partial x^2}(t, x), \tag{FP}$$

where $F$ and $D$ are given by

$$F(x) = \beta(x)C_1 - \alpha(x), \qquad D(x) = \beta(x)C_1 + \alpha(x).$$

The main problem with this derivation is to control the remainders of the Taylor expansions since the mesh size $\Delta x$ is fixed. This amounts to controlling the third derivatives of $\alpha\mathscr{C}$ and $\beta\mathscr{C}$. Since $\alpha$ and $\beta$ are known, we actually only need to control the third derivative of $\mathscr{C}$. However, classical tools from the analysis of PDEs are usually used to produce *a priori* regularity estimates of the solution, and sometimes of its derivatives, but rarely to

state decay estimates [Eva98; Fri08]. Moreover, the cases where such decay estimates are stated correspond to the situations where the diffusion term is bounded, which is not the case for Cluster Dynamics as $D(x) = \mathrm{O}(x^\gamma)$.

**Remark 6.** *Even though we do not study the well-posedness of the Fokker–Planck equation* (2.3)*, there exist classical solutions to Cauchy problems associated to such an equation [LBL08].*

The aim of the next subsection is to reformulate the Fokker–Planck equation as another equation for which we can characterize the decay of the derivatives.

### 2.3.2 Heuristic reformulation as a diffusion equation

In order to control the decay of the derivatives of the solution to (FP), we use a change of variables to reformulate the Fokker–Planck approximation with fixed $C_1$ as a diffusion equation without advection. The decay of the spatial derivatives of the solution of the diffusion equation can then be made precise (see Theorem 10).

#### Main assumptions

Let us first state the assumptions we need on the coefficients $\alpha$ and $\beta$ for this analysis.

**Assumption 3.** *The functions $\alpha$ and $\beta$ are smooth, non-decreasing and non-negative. Moreover, there exists $0 \leqslant \gamma \leqslant 1/2$ such that, as $x \to +\infty$,*

$$\alpha(x) = \mathrm{O}(x^\gamma) \qquad \text{and} \qquad \beta(x) = \mathrm{O}(x^\gamma),$$

*as well as*

$$\alpha'(x) = \mathrm{O}\Big(x^{\gamma-1}\Big) \qquad \text{and} \qquad \beta'(x) = \mathrm{O}\Big(x^{\gamma-1}\Big).$$

*and*

$$\alpha''(x) = \mathrm{O}\Big(x^{\gamma-2}\Big) \qquad \text{and} \qquad \beta''(x) = \mathrm{O}\Big(x^{\gamma-2}\Big).$$

*Finally, we assume that there exist $M, K_+, K_- > 0$ (which depend on $C_1$), such that, for all $x \geqslant M$, the function $F = \beta C_1 - \alpha$ is positive and increasing and*

$$\forall\, x \geqslant M, \qquad K_- x^\gamma \leqslant F(x) \leqslant K_+ x^\gamma. \tag{2.12}$$

A consequence of this assumption is that the functions $F$ and $D$ are smooth. Moreover,

$$F(x) = \mathrm{O}(x^\gamma), \qquad F'(x) = \mathrm{O}\Big(x^{\gamma-1}\Big), \qquad F''(x) = \mathrm{O}\Big(x^{\gamma-2}\Big),$$

and the same estimates hold for $D$ and its derivatives.

**Example 2.** *The functions $\alpha$ and $\beta$ associated with the coefficients of Example 1 read, for $x \geqslant 1$,*

$$\alpha(x) = \alpha_0 (x-1)^\gamma \exp\left(\frac{\omega}{x^\gamma}\right), \qquad \beta(x) = \beta_0 x^\gamma,$$

*where $\alpha_0 = \left(48\pi^2 V_{\mathrm{at}}^{-2}\right)^\gamma D \exp(-E_{\mathrm{v}}^{\mathrm{f}}/(k_{\mathrm{B}}T))$, $\beta_0 = \left(48\pi^2 V_{\mathrm{at}}^{-2}\right)^\gamma D$ and $\gamma \in \left\{\frac{1}{3}, \frac{1}{2}\right\}$. Assumption 3 therefore holds true for $M > \max\left(1, \omega^{1/\gamma} \ln(\beta_0 C_1/\alpha_0)^{-1/\gamma}\right)$.*

### Introducing a change of variable

Define the following functions for $x \geqslant M$:

$$g(x) = \frac{1}{F(x)}, \qquad G(x) = \int_M^x g(y)\, dy.$$

Both are well defined for $x \geqslant M$, and smooth. Moreover, in view of (2.12) in Assumption 3, $G$ is a non-negative increasing function such that $G(x) \to +\infty$ as $x \to +\infty$. We denote by $G^{-1}$ the inverse function of $G$, well defined on $[0, +\infty)$ and with values in $[M, +\infty)$. Introduce the domain

$$Z_M = \{(t, x) \in \mathbb{R}_+ \times [M, +\infty) \mid t \leqslant G(x)\}, \tag{2.13}$$

illustrated in Figure 2.2, and the function

$$\forall\, (t, x) \in Z_M, \qquad Q(t, x) = G^{-1}(G(x) - t). \tag{2.14}$$

Then, for a given function $\mathscr{C}^0_{\mathrm{adv}} \in \mathcal{C}^1([M, +\infty))$, the function defined for all $(t, x) \in Z_M$ by

$$\mathscr{C}_{\mathrm{adv}}(t, x) = \mathscr{C}^0_{\mathrm{adv}}[Q(t, x)],$$

satisfies (by the methods of characteristics, see *e.g.* [Sar03]), for all $(t, x) \in Z_M$,

$$\frac{\partial \mathscr{C}_{\mathrm{adv}}}{\partial t}(t, x) = -F(x) \frac{\partial \mathscr{C}_{\mathrm{adv}}}{\partial x}(t, x).$$

The above equation corresponds to the dominant "advection" part of the Fokker–Planck equation (FP) (see the discussion at the end of this section, in particular the estimate (2.21)). Let us notice that, for $t \geqslant 0$ and $q \geqslant M$, it holds $G(q) + t \geqslant 0$. We can therefore introduce the functions

$$\forall\, (t, q) \in \mathbb{R}_+ \times [M, +\infty), \qquad X(t, q) = G^{-1}(G(q) + t),$$

and

$$\forall\, (t, q) \in \mathbb{R}_+ \times [M, +\infty), \qquad \mathfrak{C}(t, q) = \mathscr{C}[t, X(t, q)], \tag{2.15}$$

where $\mathscr{C}$ is the solution of (2.3). Note that $X$ is the inverse function of the characteristic $Q(t, x)$ appearing in (2.14), *i.e.* $X(t, Q(t, x)) = x$ and $Q(t, X(t, q)) = q$. Using the function $X$ will allow us to suppress the advection part in (FP). Before we make this precise, let us state some useful estimates on the functions $X$ and $Q$.

**Lemma 7.** *Suppose that Assumption 3 holds true. Then, there exist $\rho_1, \rho_2 > 0$ such that*

$$\forall\, x \geqslant M, \qquad 0 \leqslant G(x) \leqslant \rho_1 x^{1-\gamma}, \qquad \textit{and} \qquad \forall\, x \geqslant 1, \qquad M \leqslant G^{-1}(x) \leqslant \rho_2 x^{\frac{1}{1-\gamma}}.$$

*Proof.* In view of (2.12), it holds

$$\forall\, x \geqslant M, \qquad 0 \leqslant G(x) \leqslant \frac{1}{K_-(1-\gamma)} \Big( x^{1-\gamma} - M^{1-\gamma} \Big),$$

from which the first estimate follows. Moreover, by definition, $\big(G^{-1}\big)' = F \circ G^{-1}$. Since $G^{-1}(x) \geqslant M$ for all $x \geqslant 0$, it holds $\big(G^{-1}\big)'(x) \leqslant K_+ \big(G^{-1}(x)\big)^\gamma$. Using a nonlinear generalisation of Gronwall's inequality [Dra03], we obtain $G^{-1}(x) \leqslant \big(M^{1-\gamma} + (1-\gamma) K_+ x\big)^{\frac{1}{1-\gamma}}$,

from which the second estimate follows. □

**Lemma 8.** *Fix a time $t \geqslant 0$. Suppose that Assumption 3 holds true. Then,*

$$\lim_{q\to+\infty} \frac{X(t,q)}{q} = 1 \qquad \textit{and} \qquad \lim_{x\to+\infty} \frac{Q(t,x)}{x} = 1. \tag{2.16}$$

*Proof.* Since $G(x) \to +\infty$ as $x \to +\infty$, there is $x^*$ such that $t + 1 \leqslant G(x^*)$. Therefore, $Q(t,x)$ is well-defined for all $x \geqslant x^*$. For any $x \geqslant x^*$, there exists $t_x \in [0,t]$ such that

$$Q(t,x) = Q(0,x) - t\partial_t Q(t_x,x) = x - t\Big(F \circ G^{-1}\Big)(G(x) - t_x). \tag{2.17}$$

Moreover, $1 \leqslant G(x) - t_x \leqslant G(x)$ and since $G^{-1}$ is increasing, $M \leqslant G^{-1}(G(x) - t_x) \leqslant x$. Therefore, since $F$ is increasing by Assumption 3,

$$F(M) \leqslant \Big(F \circ G^{-1}\Big)(G(x) - t_x) \leqslant K_+ x^\gamma, \tag{2.18}$$

so that, in view of (2.17), it holds $x - tK_+x^\gamma \leqslant Q(t,x) \leqslant x - tF(M)$. The conclusion follows from a squeeze theorem since $0 \leqslant \gamma \leqslant 1/2$. A similar reasoning can be used to prove the limit of $X(t,q)/q$ as $q \to +\infty$. □

**Lemma 9.** *Define the functions $R_{F,1}$, $R_{F,2}$ and $R_{D,1}$ as*

$$\forall\,(t,x) \in Z_M, \quad R_{F,1}(t,x) = \frac{F(Q(t,x))}{F(x)} - 1, \quad R_{F,2}(t,x) = \frac{F^2(Q(t,x))}{F^2(x)} - 1,$$

$$R_{D,1}(t,x) = \frac{D(Q(t,x))}{D(x)} - 1.$$

*Suppose that Assumption 3 holds true. Then, there exists a non-negative function $K \in \mathcal{C}^0(\mathbb{R}_+)$ such that*

$$\forall\,(t,x) \in Z_M, \qquad |R_{F,1}(t,x)|, |R_{F,2}(t,x)|, |R_{D,1}(t,x)| \leqslant K(t)x^{\gamma-1}.$$

*Proof.* In view of (2.17), it holds $Q(t,x) = x - R(t,x)$ for any $(t,x) \in Z_M$, where $R(t,x) = t(F \circ G^{-1})(G(x) - t_x)$. Fix $(t,x) \in Z_M$. Using a Taylor expansion, there exists $\zeta_{F,1}, \zeta_{D,1} \in [x - R(t,x), x]$ such that $F(Q(t,x)) = F(x) - R(t,x)F'(\zeta_{F,1})$ and $D(Q(t,x)) = D(x) - R(t,x)D'(\zeta_{D,1})$. Moreover, there exists $\zeta_2 \in [x - R(t,x), x]$ such that $F^2(Q(t,x)) = F^2(x) - 2R(t,x)F(\zeta_2)F'(\zeta_2)$. The conclusion then follows from (2.12) and (2.18). □

**Example 3.** *Fix $C_1 > 0$ and consider the following functions $\alpha$ and $\beta$:*

$$\alpha(x) = \alpha_0 x^\gamma \qquad \textit{and} \qquad \beta(x) = \beta_0 x^\gamma.$$

*These functions are asymptotically equivalent to these of Example 2. The simplicity of their expressions allows us to give analytic expressions for the functions introduced in this section. Defining $\lambda_0 = \beta_0 C_1 - \alpha_0$, it holds, for any $M > 0$,*

$$\forall\, x \geqslant M, \qquad G(x) = \frac{1}{\lambda_0(1-\gamma)}\Big[x^{1-\gamma} - M^{1-\gamma}\Big],$$

*and*

$$\forall\, y \geqslant 0, \qquad G^{-1}(y) = \Big[M^{1-\gamma} + \lambda_0(1-\gamma)y\Big]^{\frac{1}{1-\gamma}}.$$

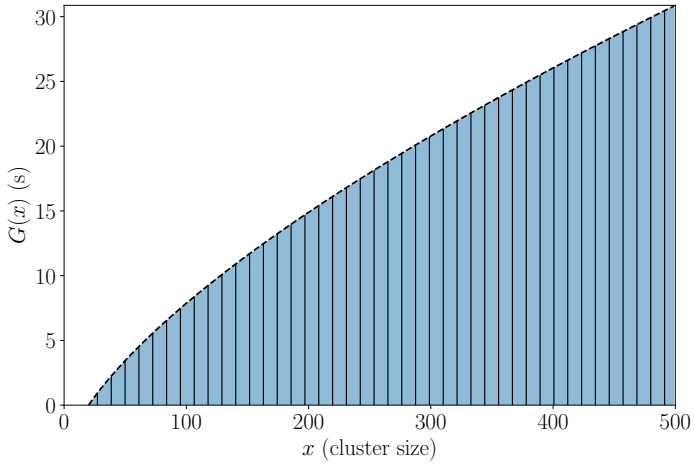



**Fig. 2.2:** The dashed line represents the function $G$ so that $Z_M$ is the hatched domain for the parameters of Example 3.

*Therefore,*

$$\forall\,(t,x)\in Z_M,\qquad Q(t,x)=\left[x^{1-\gamma}-\lambda_0(1-\gamma)t\right]^{\frac{1}{1-\gamma}},$$

*and*

$$\forall\,(t,q)\in\mathbb{R}_+\times[M,+\infty),\qquad X(t,q)=\left[q^{1-\gamma}+\lambda_0(1-\gamma)t\right]^{\frac{1}{1-\gamma}}.$$

*The domain $Z_M$ is illustrated in Figure 2.2, for the parameters used in [Ovc+03], namely $\gamma=1/3$, $D=1.83\times10^{-13}\,m^2/s$, $V_{\rm at}=1.205\times10^{-29}\,m^3$, $T=823\,K$, $E_{\rm v}^{\rm f}=1.7\,eV$ and $M=20$.*

### Heuristic reformulation of (FP)

Let us now reformulate the Fokker–Planck equation as a diffusion equation without advection with the change of variable introduced in (2.15). In order to obtain such an equation, we calculate the partial derivatives of $\mathfrak{C}$. Let us first notice that

$$\forall\,(t,q)\in\mathbb{R}_+\times[M,+\infty),\qquad \frac{\partial}{\partial q}(X(t,q))=\frac{F[X(t,q)]}{F(q)}.$$

The ratio on the right-hand side is well defined since $F(q)>0$ for all $q\geqslant M$. By the chain rule, and assuming that $\mathfrak{C}$ is smooth (a property which will be proved later on in Section 2.3.3), it holds, for all $(t,q)\in\mathbb{R}_+\times[M,+\infty)$,

$$\frac{\partial\mathfrak{C}}{\partial q}(t,q)=\frac{F[X(t,q)]}{F(q)}\frac{\partial\mathscr{C}}{\partial x}(t,X(t,q)),$$

and

$$\frac{\partial^2\mathfrak{C}}{\partial q^2}(t,q)=\frac{F^2[X(t,q)]}{F^2(q)}\frac{\partial^2\mathscr{C}}{\partial x^2}(t,X(t,q))+\frac{F[X(t,q)](F'[X(t,q)]-F'(q))}{F^2(q)}\frac{\partial\mathscr{C}}{\partial x}(t,X(t,q)).$$

Given (2.16) and using Assumption 3, we obtain that, in the limit $q \to +\infty$,

$$\frac{F[X(t,q)](F'[X(t,q)] - F'(q))}{F^2(q)} = \mathrm{O}\left(\frac{1}{q}\right), \qquad \frac{F^2(X(t,q))}{F^2(q)} \sim 1.$$

We then make the assumption, which will be proved to hold as a consequence of Theorem 10, that, as $q \to +\infty$,

$$\frac{1}{q}\frac{\partial \mathscr{C}}{\partial x}(t, X(t,q)) \ll \frac{\partial^2 \mathscr{C}}{\partial x^2}(t, X(t,q)). \tag{2.19}$$

Then, for $q$ large,

$$\frac{\partial^2 \mathfrak{C}}{\partial q^2}(t,q) \simeq \frac{\partial^2 \mathscr{C}}{\partial x^2}(t, X(t,q)).$$

Assuming further that

$$\frac{1}{q}\mathscr{C}(t, X(t,q)) \ll \frac{\partial \mathscr{C}}{\partial x}(t, X(t,q)), \tag{2.20}$$

as $q \to +\infty$, which will also be proved to hold later on, we obtain with Assumption 3 that

$$\frac{\partial (F\mathscr{C})}{\partial x}(t, X(t,q)) \simeq F[X(t,q)]\frac{\partial \mathscr{C}}{\partial x}(t, X(t,q)), \tag{2.21}$$

and

$$\frac{\partial^2 (D\mathscr{C})}{\partial x^2}(t, X(t,q)) \simeq D[X(t,q)]\frac{\partial^2 \mathscr{C}}{\partial x^2}(t, X(t,q)) \simeq D(q)\frac{\partial^2 \mathscr{C}}{\partial x^2}(t, X(t,q)).$$

We finally consider the time derivative of $\mathfrak{C}$ and combine the previous results in order to write the diffusion equation satisfied by $\mathfrak{C}$. Since

$$\frac{\partial \mathfrak{C}}{\partial t}(t,q) = \frac{\partial \mathscr{C}}{\partial t}(t, X(t,q)) + F[X(t,q)]\frac{\partial \mathscr{C}}{\partial x}(t, X(t,q)),$$

we formally obtain that $\mathfrak{C}$ is the solution of the following diffusion equation for large $q$:

$$\frac{\partial \mathfrak{C}}{\partial t}(t,q) = \frac{1}{2}\sigma^2(q)\frac{\partial^2 \mathfrak{C}}{\partial q^2}(t,q), \tag{2.22}$$

with diffusion coefficient

$$\sigma^2(q) = D(q). \tag{2.23}$$

### 2.3.3 Decay estimates of the solution of the diffusion equation

The diffusion equation (2.22) allows to relate the ODEs of CD (2.1) with fixed $C_1$ and the Fokker–Planck equation (2.3). However, before we state more precisely this result, we first need to present some results on the decay of the spatial derivatives of the solution of this diffusion equation, which will allow us to make rigorous the heuristic derivation of Section 2.3.2. Consider the following Cauchy problem:

$$\begin{cases} \dfrac{\partial \mathfrak{C}}{\partial t} = \dfrac{1}{2}\sigma^2(q)\dfrac{\partial^2 \mathfrak{C}}{\partial q^2}, \\ \mathfrak{C}(0,q) = \mathfrak{C}_0(q). \end{cases} \tag{P-Diff}$$

Assumptions on the initial condition $\mathfrak{C}_0$ will be made precise hereafter. To our knowledge, the decay of the spatial derivatives of the solutions to (P-Diff) has never been studied in

the case of an unbounded diffusion coefficient. We propose here a stochastic approach to this end, as long as $\sigma$ and $\mathfrak{C}_0$ satisfy sufficient conditions of growth and regularity. The main difficulty in giving decay estimates of the solution of such a problem comes from the fact that the diffusion coefficient $\sigma$ is not bounded, so that it does not satisfy some parabolic condition as in [Fri12, Chapter 1.1]. While Hörmander's theorem (see [Hai11, Theorem 1.3]) ensures the existence and uniqueness of a smooth solution, for a whole class of diffusion coefficients (positive with bounded derivatives on the whole space, see Assumption 4), it does not provide decay estimates on the solution. In this section, we first discuss the form of $\sigma$ as defined in (2.23), before stating decay estimates on the solution of (P-Diff).

### Characterization of the coefficient $\sigma$

Since we want to prove the correctness of the Fokker–Planck approximation in the asymptotic limit $M \to +\infty$, where $M$ is the size of a cluster, we only need to control the spatial derivatives of the solution when the space variable goes to infinity. While the expression (2.23) holds true only on $[M, +\infty)$, the use of stochastic tools requires $\sigma$ to be defined on the whole space. In order to guarantee the existence and uniqueness of the solution to the problem (P-Diff), we require that $\sigma \in \mathcal{C}^\infty(\mathbb{R}_+ \times \mathbb{R})$ is positive with bounded derivatives (which is guaranteed by Assumption 4 below). Let us now give an expression of $\sigma$ in a simple case, which will give us a useful guideline for the following.

**Example 4.** *Fix $C_1 > 0$ and consider the coefficients $\alpha$ and $\beta$ defined in Example 2. Then, for all $q \geqslant 1$,*

$$D(q) = \beta_0 C_1 q^\gamma + \alpha_0 (q-1)^\gamma \exp\left(\frac{\omega}{q^\gamma}\right) = q^\gamma \left(\beta_0 C_1 + \alpha_0 \left(1 - \frac{1}{q}\right)^\gamma \exp\left(\frac{\omega}{q^\gamma}\right)\right).$$

*Therefore, $\sigma$ writes as*

$$\sigma(q) = \sigma_0(q)\sigma_{\mathrm{b}}(q)$$

*where $\sigma_0(q) = q^{\gamma/2}$ and $\sigma_{\mathrm{b}}$ is bounded with bounded derivatives. In fact, the derivative of order $k$ of $\sigma_{\mathrm{b}}$ asymptotically decays as $q^{-k}$. The function $\sigma_0$ represents the main difficulty of our problem since it is not bounded.*

As suggested in Example 4, and in view of Assumption 3, we assume that $\sigma$ can be written as

$$\sigma(q) = \sigma_0(q)\sigma_{\mathrm{b}}(q),$$

with

$$\forall\, q \geqslant M, \qquad \sigma_0^2(q) = q^\gamma,$$

and $\sigma_{\mathrm{b}}$ a smooth positive bounded function on $[M, +\infty)$. In fact, in view of Assumption 3, $\sigma_{\mathrm{b}}$ is automatically bounded since there exist $K_{D,-}, K_{D,+} \in \mathbb{R}_+^*$ such that, for $q \geqslant M$,

$$K_{D,-} q^\gamma \leqslant D(q) \leqslant K_{D,+} q^\gamma,$$

so that, for $q \geqslant M$,

$$K_{D,-} \leqslant \sigma_{\mathrm{b}}^2(q) := \frac{D(q)}{q^\gamma} \leqslant K_{D,+}.$$

Moreover, in view of Assumption 3, we also obtain estimates on the derivatives of $\sigma$, up to the order 2. In the next section, in order to obtain general results, we assume bounds on derivatives of all order for $\sigma_{\mathrm{b}}$.

### Decay estimates of the solution of (P-Diff)

We present in this section two results concerning the decay of the solution of the Cauchy problem (P-Diff). Let us notice that the results are stated on the whole space $\mathbb{R}_+ \times \mathbb{R}$ for the Cauchy problem (P-Diff). Our assumptions on $\sigma$ are the following.

**Assumption 4.** *The diffusion coefficient $\sigma$ is a smooth positive function with bounded derivatives. Moreover,*

$$\forall\, q \in \mathbb{R}, \qquad \sigma(q) = \sigma_0(q)\sigma_{\mathrm{b}}(q),$$

*where there is $0 \leqslant \gamma \leqslant 1/2$ such that*

$$\forall\, |q| \geqslant 1, \qquad \sigma_0(q) = |q|^{\frac{\gamma}{2}},$$

*and $\sigma_{\mathrm{b}}$ is a bounded smooth function with bounded derivatives. In particular, there exist $\delta_-, \delta_+ > 0$, such that*

$$\forall\, q \in \mathbb{R}, \qquad \delta_- \leqslant \sigma_{\mathrm{b}}(q) \leqslant \delta_+.$$

*Finally, for any $n \geqslant 1$, there exists $S_n \in \mathbb{R}_+$ for which*

$$\forall\, |q| \geqslant 1, \qquad \left|\sigma^{(n)}(q)\right| \leqslant S_n |q|^{\gamma/2-n}.$$

Note that the function $\sigma$ of Example 4 satisfies Assumption 4. We also need assumptions on the initial condition $\mathfrak{C}_0$.

**Assumption 5.** *The initial condition $\mathfrak{C}_0$ is a smooth bounded function. Moreover, for all $n \geqslant 1$, there is a constant $R_n \in \mathbb{R}_+$ such that $\left\|\mathfrak{C}_0^{(n)}\sigma_0^n\right\|_{\mathcal{C}^0} \leqslant R_n$.*

Note that functions in $\mathcal{S}(\mathbb{R})$, the Schwartz space of rapidly decreasing functions, satisfy Assumption 5. We are then in position to prove the following result.

**Theorem 10.** *Fix an initial condition $\mathfrak{C}_0$ satisfying Assumption 5, and suppose that Assumption 4 holds. Then, there exists a unique classical solution of* (P-Diff)*, which is smooth and bounded. Moreover, for all $n \geqslant 1$, there exists a non-negative function $K_n \in \mathcal{C}^0(\mathbb{R}_+)$ such that*

$$\forall\, (t,q) \in \mathbb{R}_+ \times \mathbb{R}, \qquad \left|\frac{\partial^n \mathfrak{C}}{\partial q^n}(t,q)\right| \leqslant K_n(t)|q|^{-n\gamma/2}. \tag{2.24}$$

Notice that the bound depends on time through the functions $K_n$. Typically $K_n$ grows exponentially in time. Therefore, this result is useful for estimates at finite times. The proof of this result is given in Appendix 2.C.1. It relies on stochastic techniques, where the fundamental solution of (P-Diff) is interpreted as the law of a stochastic process. Let us emphasize that the assumptions stated in (2.19) and (2.20) hold true in view of Theorem 10 and Lemma 8.

**Remark 11.** *In practice, Theorem 10 holds true for $0 \leqslant \gamma < 1$. Moreover, a use of Malliavin's calculus [Nua06] allows one to conclude for $\gamma = 1$ if $\sigma = \sigma_0$. Nonetheless, in order to be consistent with the remainder of our work, we limit ourselves to $0 \leqslant \gamma \leqslant 1/2$.*

**Remark 12.** *Cerrai [Cer01, Chap. 1.5] proves the existence of a unique smooth classical solution of* (P-Diff) *assuming only $\sigma \in \mathcal{C}^3(\mathbb{R})$ with polynomial growth and an initial condition $\mathfrak{C}_0 \in \mathcal{C}_b^2(\mathbb{R})$. Therefore, since we are only interested in the third spatial derivative of the solution of* (P-Diff)*, and as a careful inspection of the proof in Appendix 2.C.1 shows, we*

*could relax some assumptions on $\sigma$ and $\mathfrak{C}_0$ and limit the assumptions on their derivatives up to order 3.*

### 2.3.4 Relating Cluster Dynamics with its Fokker–Planck approximation

We are now in position to rigorously relate the Fokker–Planck approximation (FP) and the equations of Cluster Dynamics (2.1). This section is divided into two parts. We first present a result relating the diffusion problem (P-Diff) and the Fokker–Planck approximation (FP) (see Section 2.3.4); and then a result relating the diffusion problem and the equations of Cluster Dynamics (see Section 2.3.4). We discuss the domain on which such results hold true and quantify the error arising from the approximations.

#### From the diffusion equation to the Fokker–Planck equation

Let us first relate the diffusion equation and the Fokker–Planck equation by proving that the solution of the diffusion equation satisfies up to a change of variable the Fokker–Planck equation, up to an error term whose magnitude we quantify.

**Theorem 13.** *Suppose that Assumptions 3, 4 and 5 hold true and denote by $\mathfrak{C} \in \mathcal{C}^\infty(\mathbb{R}_+ \times \mathbb{R})$ the solution of the diffusion problem* (P-Diff) *with initial condition $\mathfrak{C}_0$. Define $\mathscr{C} \in \mathcal{C}^\infty(Z_M)$ as*

$$\forall\, (t,x) \in Z_M, \qquad \mathscr{C}(t,x) = \mathfrak{C}(t, Q(t,x)),$$

*where $Q$ is introduced in* (2.14)*. Then,*

$$\forall\, (t,x) \in Z_M, \qquad \frac{\partial \mathscr{C}}{\partial t} = -\frac{\partial (F\mathscr{C})}{\partial x} + \frac{1}{2}\frac{\partial^2 (D\mathscr{C})}{\partial x^2} + R_{\mathfrak{C}}(t,x),$$

*where there exists a non-negative function $K \in \mathcal{C}^0(\mathbb{R}_+)$ such that*

$$\forall\, (t,x) \in Z_M, \qquad |R_{\mathfrak{C}}(t,x)| \leqslant K(t) x^{\gamma - 1}.$$

The estimates we obtain for $R_{\mathfrak{C}}$ on $Z_M$ show that the error arising from the reformulation of (FP) as (P-Diff) becomes smaller as the size of the clusters increases. In practice, since we do not obtain lower bounds on $|R_{\mathfrak{C}}|$, we cannot ensure whether $R_{\mathfrak{C}}$ is negligible in front of $\partial_x(F\mathscr{C})$ and $\partial_{xx}(D\mathscr{C})$.

#### From the diffusion equation to Cluster Dynamics

We now give a result relating the diffusion equation and the Cluster Dynamics (CD), up to a small error term which can be quantified. Since we work with discrete variables, we consider the discrete version $\mathcal{Z}_M$ of the space $Z_M$ defined in (2.13):

$$\mathcal{Z}_M = \{(t,n) \in \mathbb{R}_+ \times \mathbb{N} \mid n \geqslant M \text{ and } t \leqslant G(n)\},$$

on which the following approximation holds.

**Theorem 14.** *Suppose that Assumptions 3, 4 and 5 hold true and denote by $\mathfrak{C} \in \mathcal{C}^\infty(\mathbb{R}_+ \times \mathbb{R})$ the solution of the diffusion problem* (P-Diff) *with initial condition $\mathfrak{C}_0$. Fix an integer $n_0 \geqslant M$*

*and a time* $t^*_{n_0} = G(n_0)$. *Consider the sequence of smooth functions* $\widehat{C} = \left(\widehat{C}_{n_0}, \cdots, \widehat{C}_n, \cdots\right)$ *defined, for all* $n \geqslant n_0$, *by*

$$\forall\, t \in [0, t^*_{n_0}], \qquad \widehat{C}_n(t) = \mathfrak{C}(t, Q(t, n)),$$

*where* $Q$ *is defined in* (2.14). *Then, there exists a non-negative function* $K \in \mathcal{C}^0(\mathbb{R}_+)$ *for which, for all* $n \geqslant n_0$, *there is a continuous function* $R_n \in \mathcal{C}^0([0, t^*_{n_0}], \mathbb{R})$ *such that* $\widehat{C}_n$ *satisfies the following equation for all* $t \in [0, t^*_{n_0}]$:

$$\frac{d\widehat{C}_n}{dt} = \beta_{n-1}\widehat{C}_{n-1}C_1 - (\beta_n C_1 + \alpha_n)\widehat{C}_n + \alpha_{n+1}\widehat{C}_{n+1} + R_n,$$

*with*

$$\forall\, (t, n) \in \mathcal{Z}_{n_0}, \qquad |R_n(t)| \leqslant \frac{K(t)}{n^{\gamma/2}}.$$

This result gives us an estimate of the error due to the approximation based on the diffusion equation. It shows that the approximation improves when the sizes of the cluster increase. Note that the approximation is only valid for a limited time $t^*_{n_0}$, depending on the minimal size $n_0$ of the clusters for which the Fokker–Planck approximation is considered. As the minimal size $n_0$ grows, the approximation stays valid for longer times. Nevertheless, in view of the splitting introduced in Section 2.2.2, the approximation only needs to hold true on a limited time step $\Delta t$. This gives us the minimal size one can chose, which is characterized by

$$n_0 = \max\left(\lceil M \rceil, G^{-1}(\Delta t)\right).$$

In practice $\lim_{\Delta t \to 0} G^{-1}(\Delta t) = M$, therefore the limitation on $n_0$ is characterized by the real number $M$ which ensures that $F$ is positive on $[M, +\infty)$.

## 2.A Proofs for the well-posedness of CD

The proof of Theorem 3 is conducted in two main steps:

1. We first introduce a regularized version of (CD) where the unbounded operator $A$ is replaced by a bounded operator $A^\varepsilon$ for a parameter $\varepsilon > 0$, and we prove that the regularized problem has a unique classical solution $u^\varepsilon$.
2. We then show that (CD) has a unique classical solution by taking the limit as $\varepsilon \to 0$ of the solutions $u^\varepsilon$.

Section 2.A.1 is devoted to the study of the regularized version of (CD). Section 2.A.2 introduces the solution of (CD) as the limit of the aforementioned problem. The uniqueness of the solution is finally proved in Section 2.A.3.

### 2.A.1 Regularized Cluster Dynamics

In this section, we consider a bounded approximation of the operator $A$. Fix $\varepsilon > 0$ and $p \geqslant 2$ and consider the coefficients

$$\forall\, n \geqslant 1, \qquad \alpha^\varepsilon_n = \frac{\alpha_n}{1 + \varepsilon(n-1)^p} \qquad \text{and} \qquad \beta^\varepsilon_n = \frac{\beta_n}{1 + \varepsilon(n-1)^p}. \tag{2.25}$$

Similarly to (2.7) and (2.8), we define

$$A_{\mathrm{L}}^{\alpha,\varepsilon} e_1 = 0, \qquad A_{\mathrm{L}}^{\alpha,\varepsilon} e_n = \alpha_n^\varepsilon (e_1 + e_{n-1} - e_n),$$

and

$$A_{\mathrm{L}}^{\beta,\varepsilon} e_1 = -\beta_1^\varepsilon (2e_1 - e_2), \qquad A_{\mathrm{L}}^{\beta,\varepsilon} e_n = -\beta_n^\varepsilon (e_1 + e_n - e_{n+1}),$$

so that $A^\varepsilon(b) = A_{\mathrm{L}}^{\alpha,\varepsilon} + bA_{\mathrm{L}}^{\beta,\varepsilon}$ for $b \in \mathbb{R}$.

**Lemma 15.** *The operators $A_{\mathrm{L}}^{\alpha,\varepsilon}$, $A_{\mathrm{L}}^{\beta,\varepsilon}$ are bounded on $\mathcal{H}$. Therefore $A^\varepsilon(b)$ is bounded on $\mathcal{H}$ for any $b \in \mathbb{R}$.*

*Proof.* Consider $u \in \mathcal{H}$ with $u \neq 0$. In view of Assumption 1, since $p \geqslant 2$, there exists $K_\alpha^\varepsilon \geqslant 0$ such that $\alpha_n^\varepsilon \leqslant K_\alpha^\varepsilon / n$ for all $n \geqslant 2$. Then,

$$\begin{aligned}
\|A_{\mathrm{L}}^{\alpha,\varepsilon} u\|^2 &= \left( \sum_{n \geqslant 2} \alpha_n^\varepsilon u_n + \alpha_2^\varepsilon u_2 \right)^2 + \sum_{n \geqslant 2} \left( \alpha_{n+1}^\varepsilon u_{n+1} - \alpha_n^\varepsilon u_n \right)^2 \\
&\leqslant 4 \left( \sum_{n \geqslant 2} \alpha_n^\varepsilon u_n \right)^2 + 4 \sum_{n \geqslant 2} (\alpha_n^\varepsilon u_n)^2.
\end{aligned}$$

Using the Cauchy-Schwarz inequality on the first term of the right-hand side, it holds

$$\begin{aligned}
\|A_{\mathrm{L}}^{\alpha,\varepsilon} u\|^2 &\leqslant 4 \sum_{n \geqslant 2} (\alpha_n^\varepsilon)^2 \sum_{n \geqslant 2} u_n^2 + 4 \left( \sup_{n \geqslant 2} \alpha_n^\varepsilon \right)^2 \sum_{n \geqslant 2} u_n^2 \\
&\leqslant \frac{2 (K_\alpha^\varepsilon)^2 \pi^2}{3} \|u\|^2 + 4 \left( \sup_{n \geqslant 2} \alpha_n^\varepsilon \right)^2 \|u\|^2,
\end{aligned}$$

which leads to the boundedness of $A_{\mathrm{L}}^{\alpha,\varepsilon}$. A similar reasoning can be used to prove the boundedness of $A_{\mathrm{L}}^{\beta,\varepsilon}$, which concludes the proof. □

We consider the following Cauchy problem:

$$\begin{cases} \dfrac{du^\varepsilon}{dt} = A^\varepsilon(u_1^\varepsilon) u^\varepsilon, \\ u^\varepsilon(0) = u^0. \end{cases} \tag{CD$_\varepsilon$}$$

In order to prove that (CD$_\varepsilon$) has a unique classical solution, we first consider a linear version of this problem and then we use a fixed point argument to prove existence and uniqueness of a local solution. In order to obtain the existence and uniqueness of a global-in-time solution we prove that the solution remains non-negative and then we use the total quantity of matter as a Lyapunov function for the dynamics, proving this way that the solution is bounded and has values in $\mathcal{Q}$.

### Existence and uniqueness of a local solution

Fix $b \in \mathcal{C}^0(\mathbb{R}_+, \mathbb{R})$. For a given $T > 0$, we now consider the following linear problem for $t \in [0, T]$:

$$\begin{cases} \dfrac{du^\varepsilon}{dt} = A^\varepsilon(b(t))u^\varepsilon, \\ u^\varepsilon(0) = u^0. \end{cases} \tag{$\mathrm{CD}_{\varepsilon,L}$}$$

Since $t \mapsto A^\varepsilon(b(t))$ is continuous in the uniform operator topology, and, by construction, since $A^\varepsilon(b(t))$ is bounded for all $t \in [0, T]$, the following result holds

**Proposition 16.** *[Paz12, Chapter 5, Theorems 5.1 and 5.2] Consider an initial condition $u^0 \in \mathcal{H}$. The problem ($\mathrm{CD}_{\varepsilon,L}$) has a unique classical solution $u^\varepsilon \in \mathcal{C}^1([0, T], \mathcal{H})$. Moreover, there exists a family of bounded operators $(U_b^\varepsilon(t, s))_{0\leqslant s\leqslant t\leq T}$ such that $u^\varepsilon(t) = U_b^\varepsilon(t, 0)u^0$ and*

*(i)* $\|U_b^\varepsilon(t, s)\| \leqslant \exp(\omega_b^\varepsilon(t - s))$ *for* $0 \leqslant s \leqslant t \leqslant T$
*with* $\omega_b^\varepsilon = \|A_{\mathrm{L}}^{\alpha,\varepsilon}\| + |b|_{\mathcal{C}^0([0,T])}\left\|A_{\mathrm{L}}^{\beta,\varepsilon}\right\|$.

*(ii)* $U_b^\varepsilon(t, t) = I$, $U_b^\varepsilon(t, r)U_b^\varepsilon(r, s) = U_b^\varepsilon(t, s)$ *for* $0 \leqslant s \leqslant r \leqslant t \leqslant T$.

*(iii)* $(t, s) \mapsto U_b^\varepsilon(t, s)$ *is continuous in the uniform operator topology for* $0 \leqslant s \leqslant t \leqslant T$.

*(iv)* $\dfrac{\partial}{\partial t}(U_b^\varepsilon(t, s)) = A^\varepsilon(b(t))U_b^\varepsilon(t, s)$ *for* $0 \leqslant s \leqslant t \leqslant T$,

*(v)* $\dfrac{\partial}{\partial s}(U_b^\varepsilon(t, s)) = -U_b^\varepsilon(t, s)A^\varepsilon(b(s))$ *for* $0 \leqslant s \leqslant t \leqslant T$.

*In the above equations, the equalities hold in $\mathcal{L}(\mathcal{H})$.*

Before proving the existence and uniqueness of a local solution of ($\mathrm{CD}_\varepsilon$), we show the following useful technical result.

**Lemma 17.** *Fix a time $T > 0$ and $Q_* \geqslant 0$. Consider $u, v \in \mathcal{C}^0([0, T], \mathcal{H})$ with $0 \leqslant u_1(t), v_1(t) \leqslant Q_*$ for all $0 \leqslant t \leqslant T$. Then, there exists $K_\varepsilon(T, Q_*) \geqslant 0$ (depending on $\varepsilon$, $T$ and $Q_*$ but not on $u$ and $v$) such that, for all $0 \leqslant s \leqslant t \leqslant T$,*

$$\|(U_{u_1}(t, s) - U_{v_1}(t, s))\| \leqslant K_\varepsilon(T, Q_*) \int_s^t |u_1(r) - v_1(r)|\, dr \leqslant K_\varepsilon(T, Q_*) \int_s^t \|u(r) - v(r)\|\, dr. \tag{2.26}$$

*Proof.* Consider $r \mapsto U_{u_1}^\varepsilon(t, r)U_{v_1}^\varepsilon(r, s)$, which belongs to $\mathcal{C}^1([0, T], \mathcal{L}(\mathcal{H}))$ in view of Proposition 16. Then,

$$\begin{aligned} U_{v_1}^\varepsilon(t, s) - U_{u_1}^\varepsilon(t, s) &= -\int_s^t \frac{\partial}{\partial r}\left[U_{u_1}^\varepsilon(t, r)U_{v_1}^\varepsilon(r, s)\right] dr \\ &= -\int_s^t \left(\left[\frac{\partial}{\partial r}U_{u_1}^\varepsilon(t, r)\right]U_{v_1}^\varepsilon(r, s) + U_{u_1}^\varepsilon(t, r)\left[\frac{\partial}{\partial r}U_{v_1}^\varepsilon(r, s)\right]\right) dr \\ &= \int_s^t U_{u_1}^\varepsilon(t, r)[A^\varepsilon(u_1(r)) - A^\varepsilon(v_1(r))]U_{v_1}^\varepsilon(r, s)\, dr. \end{aligned}$$

Since $0 \leqslant u_1(t) \leqslant Q_*$ for all $0 \leqslant t \leqslant T$, we have $\|U_{u_1}^\varepsilon(t, s)\| \leqslant \mathrm{e}^{\omega_{Q_*}^\varepsilon(t-s)}$ (see Proposition 16). Moreover,

$$\|[A^\varepsilon(u_1(r)) - A^\varepsilon(v_1(r))]U_{v_1}^\varepsilon(r, s)\| = |u_1(r) - v_1(r)|\left\|A_{\mathrm{L}}^{\beta,\varepsilon}U_{v_1}^\varepsilon(r, s)\right\| \leqslant L_0^\varepsilon|u_1(r) - v_1(r)|,$$

where $L_0^\varepsilon = \left\| A_{\mathrm{L}}^{\beta,\varepsilon} \right\| \exp(\omega_{Q_*}^\varepsilon T)$ since $0 \leqslant v_1(t) \leqslant Q_*$ for all $0 \leqslant t \leqslant T$. This gives the estimate (2.26). □

We are now in position to prove the existence and uniqueness of a local solution.

**Proposition 18.** *Consider an initial condition $u^0 \in \mathcal{H}$ with $u_1^0 > 0$. Then, for any $\varepsilon > 0$, there exists a time $T > 0$ (which depends on $\varepsilon$) such that the Cauchy problem* (CD$_\varepsilon$) *admits a unique classical solution $u^\varepsilon \in \mathcal{C}^1([0,T],\mathcal{H})$.*

*Proof.* The proof is organized in three steps. The first one consists in reformulating the Cauchy problem (CD$_\varepsilon$) as a fixed point problem for some function $F_\varepsilon$ on $\mathcal{C}^0([0,T],\mathcal{H})$. We then define a subset $\mathcal{E}$ of this space and show that $F_\varepsilon$ maps $\mathcal{E}$ to itself. We prove that $F_\varepsilon$ is a contraction on $\mathcal{E}$ and conclude to the existence and uniqueness of a classical solution.

*Step 1: Reformulation as a fixed point equation.* Fix $u^\varepsilon \in \mathcal{C}^0(\mathbb{R}_+,\mathcal{H})$, a time $T_0 > 0$ and an initial condition $u^0 \in \mathcal{Q}$, with $u_1^0 > 0$. Proposition 16 shows that the problem

$$\begin{cases} \dfrac{dv^\varepsilon}{dt} = A^\varepsilon(u_1^\varepsilon)v^\varepsilon, \\ v^\varepsilon(0) = u^0, \end{cases}$$

has a unique solution $v^\varepsilon \in \mathcal{C}^1([0,T_0],\mathcal{H})$ such that $v^\varepsilon(t) = U^\varepsilon_{u_1^\varepsilon}(t,0)u^0$ for all $0 \leqslant t \leqslant T_0$. We can therefore define the following map $F_\varepsilon$ from $\mathcal{C}^0([0,T_0],\mathcal{H})$ into itself:

$$\forall\, t \in \mathbb{R}_+, \qquad F_\varepsilon(u)(t) = U^\varepsilon_{u_1^\varepsilon}(t,0)u^0.$$

Note that a classical solution $u^\varepsilon$ of (CD$_\varepsilon$) on the time interval $[0,T_0]$ is such that $u^\varepsilon = F_\varepsilon(u^\varepsilon)$.

*Step 2: Definition of the subset $\mathcal{E}$.* Consider a time $t_1 > 0$ such that

$$\max_{0\leqslant t\leqslant t_1} \left\| U^\varepsilon_{u_1^0}(t,0)u^0 - u^0 \right\| \leqslant \frac{u_1^0}{2},$$

where, with some abuse of notation, we denote by $U^\varepsilon_{u_1^0}$ the propagator associated with the constant function $t \mapsto u_1^0$. The time $t_1$ can indeed be chosen positive since $t \mapsto U^\varepsilon_{u_1^0}(t,0)u^0$ belongs to $\mathcal{C}^0(\mathbb{R}_+,\mathcal{H})$. Note also that $\left(U^\varepsilon_{u_1^0}(t,0)u^0\right)_1$ remains positive for all $t \in [0,t_1]$. Let us then define the time

$$T = \min\left(t_1, T_0, \frac{1}{2K_\varepsilon(T_0, 2u_1^0)\|u^0\|_D}\right) > 0,$$

where the constant $K_\varepsilon(T_0, 2u_1^0)$ is introduced in Lemma 17. We define the closed convex subset $\mathcal{E}$ of the Banach space $\mathcal{C}^0([0,T],\mathcal{H})$ as

$$\mathcal{E} = \left\{ u \in \mathcal{C}^0([0,T],\mathcal{H}) \;\middle|\; u(0) = u^0,\ \sup_{0\leqslant t\leqslant T} \left\| u(t) - u^0 \right\| \leqslant u_1^0 \right\}.$$

The function $F_\varepsilon$ maps $\mathcal{E}$ into itself. Indeed, it clearly holds that $F_\varepsilon(u)(0) = u^0$. Moreover,

for $u \in \mathcal{E}$ and $t \in [0,T]$,

$$\left\|F_\varepsilon(u)(t) - u^0\right\| \leqslant \left\|U^\varepsilon_{u_1}(t,0)u^0 - U^\varepsilon_{u_1^0}(t,0)u^0\right\| + \left\|U^\varepsilon_{u_1^0}(t,0)u^0 - u^0\right\|$$
$$\leqslant K_\varepsilon\left(T_0, 2u_1^0\right)u_1^0\left\|u^0\right\|_D T + \frac{u_1^0}{2} \leqslant u_1^0,$$

where we used Lemma 17 and the definition of $T$.

*Step 3: $F_\varepsilon$ is a contraction on $\mathcal{E}$.* For any $v, w \in \mathcal{E}$ and $t \in [0,T]$,

$$\|F_\varepsilon(v)(t) - F_\varepsilon(w)(t)\| = \left\|U^\varepsilon_{v_1}(t,0)u^0 - U^\varepsilon_{w_1}(t,0)u^0\right\| \leqslant K_\varepsilon\left(T_0, 2u_1^0\right)\left\|u^0\right\| \int_0^t \|v(s) - w(s)\|\, ds$$
$$\leqslant K_\varepsilon\left(T_0, 2u_1^0\right)T\|u^0\|\|v - w\|_{\mathcal{C}^0([0,T],\mathcal{H})} \leqslant \frac{1}{2}\|v - w\|_{\mathcal{C}^0([0,T],\mathcal{H})},$$

where we used once again Lemma 17 and the definition of $T$. Then,

$$\|F_\varepsilon(v) - F_\varepsilon(w)\|_{\mathcal{C}^0([0,T],\mathcal{H})} \leqslant \frac{1}{2}\|v - w\|_{\mathcal{C}^0([0,T],\mathcal{H})}.$$

Therefore, $F_\varepsilon$ is a contraction from $\mathcal{E}$ into itself. By the Banach fixed point theorem [Ban22; Zei95], there exists therefore a unique fixed point $u^\varepsilon \in \mathcal{E}$ of $F_\varepsilon$. □

### Existence and uniqueness of a global-in-time solution

In this section, we prove the following result.

**Theorem 19.** *Fix an initial condition $u^0 \in \mathcal{Q}$ such that $u_1^0 > 0$. Then, there exists a unique global-in-time classical solution $u^\varepsilon \in \mathcal{C}^0(\mathbb{R}_+, \mathcal{Q}) \cap \mathcal{C}^1(\mathbb{R}_+, \mathcal{H})$ of the problem* (CD$_\varepsilon$)*. Moreover,*

$$\forall\, t \geqslant 0, \qquad Q(u^\varepsilon(t)) = Q\left(u^0\right) \qquad \textit{and} \qquad \|u^\varepsilon(t)\| \leqslant \frac{\pi Q(u^0)}{\sqrt{6}}.$$

In order to prove Theorem 19, we first show that the non-negativity of the solution is conserved. The precise statement is the following.

**Proposition 20.** *Consider an initial solution $u^0 \in \mathcal{H}_+$ with $u_1^0 > 0$. If there exists a solution $u^\varepsilon \in \mathcal{C}^1([0,T], \mathcal{H})$ to the problem* (CD$_\varepsilon$) *for some time $T > 0$, then $u^\varepsilon(t) \in \mathcal{H}_+$ for all $t \in [0,T]$.*

*Proof.* We first note that since $u_1^0 > 0$, by continuity of the solution, there is a time $\tau > 0$ such that $u_1^\varepsilon(t) > 0$ for all $0 \leqslant t \leqslant \tau$. Let us now prove that $u_1^\varepsilon$ remains positive on $[0,T]$. We consider

$$\tau_* = \sup\{t \in [0,T] \,:\, u_1^\varepsilon(t) > 0\},$$

and prove that $\tau_* = T$. We suppose that $\tau_* < T$ and obtain a contradiction. Since the dynamics on $u_n^\varepsilon$, for $n \geqslant 2$, is a Kolmogorov forward equation of a (time-inhomogeneous) Markov jump process [Ter+17], it holds $u_n^\varepsilon(t) \geqslant 0$ for all $n \geqslant 2$ and $0 \leqslant t \leqslant \tau_*$. Introduce $\mathcal{A}^\varepsilon(t) = \sum_{n\geqslant 2} \alpha_n^\varepsilon u_n^\varepsilon(t) + \alpha_2^\varepsilon u_2^\varepsilon(t)$ and $\mathcal{B}^\varepsilon(t) = \sum_{n\geqslant 2} \beta_n^\varepsilon u_n^\varepsilon(t)$ for $t \in [0, \tau_*]$. Then, $\mathcal{A}^\varepsilon(t) \geqslant 0$ and, since $u^\varepsilon \in \mathcal{C}^0([0,\tau_*], \mathcal{H})$, there is $\mathcal{B}_*^\varepsilon > 0$ such that $\mathcal{B}^\varepsilon(t) \leqslant \mathcal{B}_*^\varepsilon$ for all $0 \leqslant t \leqslant \tau_*$. The

function $u_1^\varepsilon$ therefore satisfies the following ordinary differential inequality

$$\frac{du_1^\varepsilon}{dt} \geqslant -2\beta_1^\varepsilon u_1^\varepsilon(t)^2 - \mathcal{B}_*^\varepsilon u_1^\varepsilon(t).$$

Consider the solution $\widetilde{u}_1$ of the associated ODE $\widetilde{u}_1' = -2\beta_1^\varepsilon - \mathcal{B}_*^\varepsilon \widetilde{u}_1$ with initial condition $\widetilde{u}_1(0) = u_1^0$. In fact, for $0 \leqslant t \leqslant \tau_*$,

$$\widetilde{u}_1(t) = \frac{u_1^0 r_*^\varepsilon \exp(-\mathcal{B}_*^\varepsilon t)}{u_1^0(1 - \exp(-\mathcal{B}_*^\varepsilon t) + r_*^\varepsilon},$$

where $r_*^\varepsilon = \mathcal{B}_*^\varepsilon/(2\beta_1^\varepsilon)$. Therefore $\widetilde{u}_1(t) > 0$ for all $0 \leqslant t \leqslant \tau_*$. By comparison [Har02, Chapter III], it holds $u_1^\varepsilon(t) \geqslant \widetilde{u}_1(t) > 0$ for all $t \in [0, \tau_*]$. Then, by continuity of the solution, there is $\eta > 0$ such that $u_1^\varepsilon$ is positive on $[0, \tau_* + \eta]$ which is in contradiction with the definition of $\tau_*$ since we assumed that $\tau_* < T$. We conclude the proof by noting that $u_1^\varepsilon > 0$ on $[0, T]$ implies $u_n^\varepsilon \geqslant 0$ on $[0, T]$ for all $n \geqslant 2$ by the interpretation of the dynamics at fixed $u_1$ as a Markov jump process. □

We are now in position to prove Theorem 19.

*Proof.* We prove that the total quantity of matter $Q(u^\varepsilon)$ is conserved on the time interval on $[0, T]$. Indeed,

$$\begin{aligned}
\frac{dQ(u^\varepsilon(t))}{dt} =& \frac{du_1^\varepsilon}{dt}(t) + \sum_{n\geqslant 2} n \frac{du_n^\varepsilon}{dt}(t) \\
=& -2\beta_1^\varepsilon u_1^\varepsilon(t)^2 - \sum_{n\geqslant 2} \beta_n^\varepsilon u_n^\varepsilon(t) u_1^\varepsilon(t) + \sum_{n\geqslant 3} \alpha_n^\varepsilon u_n^\varepsilon(t) + 2\alpha_2^\varepsilon u_2^\varepsilon(t) \\
&+ \sum_{n\geqslant 2} n\big[\beta_{n-1}^\varepsilon u_{n-1}^\varepsilon(t) u_1^\varepsilon(t) - (\beta_n^\varepsilon u_1^\varepsilon(t) + \alpha_n^\varepsilon) u_n^\varepsilon(t) + \alpha_{n+1}^\varepsilon u_{n+1}^\varepsilon(t)\big] \\
=& \, 0.
\end{aligned} \tag{2.27}$$

Note that all the sums above are well defined since the sequences $(n\alpha_n^\varepsilon)_{n\geqslant 2}$ and $(n\beta_n^\varepsilon)_{n\geqslant 2}$ are in $\ell^2(\mathbb{N}^*, \mathbb{R})$ (by the choice $p \geqslant 2$ in (2.25)) so that the sequences $(n\alpha_n^\varepsilon u_n^\varepsilon)_{n\geqslant 2}$ and $(n\beta_n^\varepsilon u_n^\varepsilon)_{n\geqslant 2}$ are in $\ell^1(\mathbb{N}^*, \mathbb{R})$. Moreover, for all $n \geqslant 1$ and $t \in [0, T]$, it holds $0 \leqslant n u_n^\varepsilon(t) \leqslant Q(u^\varepsilon(0)) = Q(u^0)$. This leads to the following estimate:

$$\forall\, t \in [0, T], \qquad \|u^\varepsilon(t)\|^2 = \sum_{n\geqslant 1} (u_n^\varepsilon(t))^2 \leqslant \sum_{n\geqslant 1} \frac{\big(Q(u^0)\big)^2}{n^2} \leqslant \frac{\big(Q(u^0)\big)^2 \pi^2}{6}.$$

This shows finally that the norm of the solution is uniformly bounded. The solution can therefore be extended to all times $t \in \mathbb{R}_+$. □

## 2.A.2 Proof of Theorem 3 – Existence

In this section, we show that we can build a solution of (CD) from a solution of ($\text{CD}_\varepsilon$) when $\varepsilon \to 0$. We first prove that the solution $u^\varepsilon$ of ($\text{CD}_\varepsilon$) is uniformly bounded in $\mathcal{C}^0(\mathbb{R}_+, \mathcal{Q}) \cap \mathcal{C}^1(\mathbb{R}_+, \mathcal{H})$. Then, we extract a candidate solution $u \in \mathcal{C}^0(\mathbb{R}_+, \mathcal{H})$ and show that it is a classical solution of (CD).

**Proposition 21.** *Consider an initial condition $u^0 \in \mathcal{Q}$ with $u_1^0 > 0$ and $\varepsilon > 0$. Fix a time $T > 0$ and define $Q_0 = Q(u^0)$. Then, there exists $K(Q_0) \in \mathbb{R}_+$, which is independent of $\varepsilon$,*

*such that the solution* $u^\varepsilon \in \mathcal{C}^0([0,T],\mathcal{Q}) \cap \mathcal{C}^1([0,T],\mathcal{H})$ *of* (CD$_\varepsilon$) *satisfies*

$$\|u^\varepsilon\|_{\mathcal{C}^1([0,T],\mathcal{H})} \leqslant K(Q_0)$$

*Proof.* In view of Theorem 19, the solution $u^\varepsilon$ satisfies $\|u^\varepsilon(t)\| \leqslant \pi Q_0/\sqrt{6}$ for all $t \geqslant 0$, so that $\|u^\varepsilon(t)\|_{\mathcal{C}^0([0,T],\mathcal{H})} \leqslant \pi Q_0/\sqrt{6}$, which is independent of $\varepsilon$. Moreover, for $0 \leqslant t \leqslant T$, it holds

$$\left\|\frac{du^\varepsilon}{dt}(t)\right\| = \|A^\varepsilon(u_1^\varepsilon(t))u^\varepsilon(t)\| \leqslant \|A_{\mathrm{L}}^{\alpha,\varepsilon}u^\varepsilon(t)\| + |u_1^\varepsilon|_{\mathcal{C}^0([0,T])}\left\|A_{\mathrm{L}}^{\beta,\varepsilon}u^\varepsilon(t)\right\|.$$

Then, in view of Assumption 1, since $u^\varepsilon \in \mathcal{Q}$ with $Q(u^\varepsilon(t)) \leqslant Q_0$ for all $t \geqslant 0$, there exists $K(Q_0)$ such that $\|A_{\mathrm{L}}^{\alpha,\varepsilon}u^\varepsilon(t)\|, \left\|A_{\mathrm{L}}^{\beta,\varepsilon}u^\varepsilon(t)\right\| \leqslant K(Q_0)$ (see Lemma 32 in Appendix 2.B.2). This concludes the proof. □

Using the Arzelà-Ascoli theorem (see for instance [Sch91]), we obtain the following result.

**Proposition 22.** *Consider an initial condition* $u^0 \in \mathcal{Q}$ *with* $u_1^0 > 0$. *Fix a time* $T > 0$ *and define* $Q_0 = Q(u^0)$. *There exists a sequence* $(\varepsilon_n)_{n\geqslant 1}$ *with* $\varepsilon_n \to 0$ *as* $n \to +\infty$ *such that* $u^{\varepsilon_n} \to u$ *in* $\mathcal{C}^0([0,T],\mathcal{H})$. *Moreover,* $u(t) \in \mathcal{Q}$ *for all* $t \in [0,T]$ *and* $Q(u(t)) \leqslant Q_0$.

*Proof.* We introduce the following subset of $\mathcal{C}^0([0,T],\mathcal{H})$:

$$\mathcal{M} = \Big\{u \in \mathcal{C}^0([0,T],\mathcal{Q}) \cap \mathcal{C}^1([0,T],\mathcal{H}) \,\Big|\, \forall\, t \in [0,T],\, Q(u(t)) = Q_0\Big\}.$$

The subset $\mathcal{M}$ is clearly equicontinuous in view of Proposition 21. Consider next, for all $t \in [0,T]$, the set $\mathcal{M}(t) = \{u(t), u \in \mathcal{M}\} \subset \mathcal{H}$. We prove that $\mathcal{M}(t)$ is relatively compact in $\mathcal{H}$. Introducing the Banach space $\widetilde{\mathcal{Q}} = \{u \in \mathcal{H} \,|\, \sum_{n\geqslant 1} n|u_n| < +\infty\}$ equipped with the norm $\|u\|_{\mathcal{Q}} = \sum_{n\geqslant 1} n|u_n|$, we define the canonical injection $i : (\widetilde{\mathcal{Q}}, \|\cdot\|_{\mathcal{Q}}) \to (\mathcal{H}, \|\cdot\|)$ and the sequence of finite-rank operators $(i_N)_{N\geqslant 1}$ by

$$i_N : u \in \widetilde{\mathcal{Q}} \mapsto (u_1, \cdots, u_N, 0, \cdots) \in \mathcal{H}.$$

Then, for all $u \in \widetilde{\mathcal{Q}}$, it holds

$$\|(i - i_N)u\|^2 = \sum_{n\geqslant N+1} |u_n|^2 \leqslant \|u\|_{\mathcal{Q}} \sum_{n\geqslant N+1} \frac{|u_n|}{n} \leqslant \frac{\|u\|_{\mathcal{Q}}}{(N+1)^2} \sum_{n\geqslant N+1} n|u_n| \leqslant \frac{\|u\|_{\mathcal{Q}}^2}{(N+1)^2},$$

so that $\|i - i_N\|_{\mathcal{L}(\widetilde{\mathcal{Q}},\mathcal{H})} \leqslant 1/(N+1) \to 0$. Since the canonical injection $i$ is the limit of compact operators (since $i_N$ is finite-rank for all $N \geqslant 1$), the operator $i$ is compact, which proves that $\mathcal{M}(t)$ is relatively compact in $\mathcal{H}$ for all $t \in [0,T]$. In view of the Arzelá-Ascoli theorem, we conclude that there exists a sequence $(\varepsilon_n)_{n\geqslant 1}$ with $\varepsilon_n \to 0$ as $n \to +\infty$ such that $u^{\varepsilon_n} \to u$ in $\mathcal{C}^0([0,T],\mathcal{H})$. Moreover, $u_n \geqslant 0$ for all $n \geqslant 0$, and,

$$\forall\, N \geqslant 1, \qquad \sum_{n=1}^{N} n u_n^\varepsilon \to \sum_{n=1}^{N} n u_n,$$

with $0 \leqslant \sum_{n=1}^{N} n u_n \leqslant Q_0$, so that $0 \leqslant Q(u) \leqslant Q_0$. This shows that $u$ has values in $\mathcal{Q}$. □

We finally conclude this section with the proof of the existence of a global-in-time solution.

*Proof of Theorem 3 – Existence.* Consider the sequence $(\varepsilon_n)_{n\geqslant 1}$ of Proposition 22 and fix

$T > 0$. Then, for all $k \geqslant 2$ and all $0 \leqslant t \leqslant T$, it holds

$$u_k^{\varepsilon_n}(t) = u_k^0 + \int_0^t \left[\beta_{k-1}^{\varepsilon_n} u_1^{\varepsilon_n}(s) u_{k-1}^{\varepsilon_n}(s) - (\beta_k^{\varepsilon_n} u_1^{\varepsilon_n}(s) + \alpha_k^{\varepsilon_n}) u_k^{\varepsilon_n}(s) + \alpha_{k+1}^{\varepsilon_n} u_{k+1}^{\varepsilon_n}(s)\right] ds.$$

Noting that all the terms under the integral sign are uniformly bounded in view the estimate $Q(u^{\varepsilon_n}(t)) = Q_0$ and the corresponding limit also makes sens for $u$, it holds

$$u_k(t) = u_k^0 + \int_0^t \left[\beta_{k-1} u_1(s) u_{k-1}(s) - (\beta_k u_1(s) + \alpha_k) u_n(s) + \alpha_{k+1} u_{k+1}(s)\right] ds.$$

Similarly, for $k = 1$,

$$u_1^{\varepsilon_n}(t) = u_1^0 - \int_0^t 2\beta_1^{\varepsilon_n} (u_1^{\varepsilon_n}(s))^2 ds - \int_0^t \sum_{k \geqslant 2} \beta_k^{\varepsilon_n} u_1^{\varepsilon_n}(s) u_k^{\varepsilon_n}(s)\, ds$$

$$+ \int_0^t \sum_{k \geqslant 2} \alpha_k^{\varepsilon_n} u_k^{\varepsilon_n}(s)\, ds + \int_0^t \alpha_2^{\varepsilon_n} u_2^{\varepsilon_n}(s)\, ds,$$

so that, passing to the limit $\varepsilon_n \to 0$,

$$u_1(t) = u_1^0 - \int_0^t 2\beta_1 (u_1(s))^2 ds - \int_0^t \sum_{k \geqslant 2} \beta_k u_1(s) u_k(s)\, ds$$

$$+ \int_0^t \sum_{k \geqslant 2} \alpha_k u_k(s)\, ds + \int_0^t \alpha_2 u_2(s)\, ds.$$

This proves that $u \in \mathcal{C}^0([0,T], \mathcal{Q})$ is the solution of the integral formulation of (CD). Moreover, since $Q(u) \leqslant Q_0$, it holds $u \in \mathcal{C}^1([0,T], \mathcal{H})$ and $u$ solves (CD) by differentiating in time the integral formulation of (CD). Finally, similarly to (2.27), a simple computation shows that $Q(u) = Q_0$. Since $T$ was arbitrary, this shows that $u \in \mathcal{C}^0([0,T], \mathcal{Q}) \cap \mathcal{C}^1([0,T], \mathcal{H})$ is a global in time solution to (CD). □

## 2.A.3 Proof of Theorem 3 – Uniqueness

We finally conclude the proof of Theorem 3 by showing the uniqueness of the solution with an argument based on the dissipativity of the operator $A$. We start by studying the operator $A(b)$ when the parameter $b$ is fixed to a constant value in $\mathbb{R}_+$. Let us emphasize that the non-negativity of $b$ is crucial for proving the dissipativity of $A$. We consider the domain

$$D(A(b)) = \left\{ u \in \mathcal{H} \;\middle|\; \sum_{n \geqslant 1} (A(b)u)_n^2 < +\infty \right\}$$

for the operator $A(b) = A_{\mathrm{L}}^{\alpha} + bA_{\mathrm{L}}^{\beta}$ on $\mathcal{H}$, which is dense since it contains $c_{00}$, the space of sequences which have only finitely many nonzero elements. Since $D(A(b))$ is dense, we can define the adjoint of $A(b)$. The operator $A(b)^*$ has action

$$A(b)^* e_1 = -2\beta_1 b e_1 + (2\alpha_2 - \beta_2 b) e_2 + \sum_{n \geqslant 3} (\alpha_n - \beta_n b) e_n,$$

$$A(b)^* e_n = \beta_{n-1} b e_{n-1} - (\beta_n b + \alpha_n) e_n + \alpha_{n+1} e_{n+1}, \qquad n \geqslant 2,$$

and dense domain $D(A(b)^*) = \{u \in \mathcal{H} \mid \sum_{n \geqslant 1} (A(b)^* u)_n^2 < +\infty\}$.

**Remark 23.** *For the sake of simplicity and uniformity of notation, we use the convention that* $\alpha_1 = 0$ *in the sequel.*

**Remark 24.** *The operators* $A(b)$ *and* $A(b)^*$ *are clearly unbounded. Indeed, for* $n \geqslant 2$*, consider the sequence* $(e_n)_{n\geqslant 1}$ *of elements of* $\mathcal{H}$*. One gets for* $n \geqslant 2$

$$\frac{\|A(b)e_n\|^2}{\|e_n\|^2} = (A(b)e_n)^2 = (\alpha_n - \beta_n b)^2 + (\beta_n b)^2 + (\beta_n b + \alpha_n)^2 + \alpha_n^2 = 3(\alpha_n^2 + \beta_n^2 b^2)$$

*and*

$$\frac{\|A(b)^*e_n\|^2}{\|e_n\|^2} = (A(b)^*e_n)^2 = \beta_{n-1}^2 b^2 + (\beta_n b + \alpha_n)^2 + \alpha_{n+1}^2.$$

*Since* $\alpha$ *and* $\beta$ *are unbounded, this shows that the operators* $A(b)$ *and* $A(b)^*$ *are also unbounded.*

Let us now introduce a sequence which naturally arises in the following analysis. Fix $\lambda > 0$ and define the sequence $(\delta_n^\lambda)_{n\geqslant 1}$ as

$$\begin{aligned} \delta_1^\lambda &= \lambda + \beta_1 b + \alpha_1, \\ \delta_n^\lambda &= \lambda + \beta_n b + \alpha_n - \frac{1}{4}\frac{(\beta_{n-1} b + \alpha_n)^2}{\delta_{n-1}^\lambda}, \end{aligned} \tag{2.28}$$

which is well defined as long as $\delta_n^\lambda > 0$ for $n \geqslant 1$.

**Lemma 25.** *Suppose that Assumption 1 holds. Then, there exists* $\lambda_b > 0$ *such that the sequence* $\delta^{\lambda_b}$ *is well defined and satisfies the following lower bound:*

$$\forall\, n \in \mathbb{N}^*, \qquad \delta_n^{\lambda_b} \geqslant \frac{1}{2}(\alpha_{n+1} + \beta_n b) > 0.$$

*Proof.* We proceed by induction. Define

$$\lambda_b = \frac{1}{2}\max\left(B(1+b), \alpha_2\right) > 0, \tag{2.29}$$

where $B$ is introduced in Assumption 1. For $n = 1$, it holds

$$\delta_1^{\lambda_b} = \lambda_b + \alpha_1 + \beta_1 b \geqslant \frac{1}{2}(\alpha_2 + \beta_1 b) > 0.$$

Assume now that $\delta_n^{\lambda_b} \geqslant \frac{1}{2}(\alpha_{n+1} + \beta_n b) > 0$ for some integer $n \geqslant 1$. Since $(\delta_n^{\lambda_b})^{-1} \leqslant \frac{2}{\alpha_{n+1}+\beta_n b}$, one obtains

$$\begin{aligned} \delta_{n+1}^{\lambda_b} &\geqslant \lambda_b + (\alpha_{n+1} + \beta_{n+1} b) - \frac{1}{2}(\alpha_{n+1} + \beta_n b) \\ &= \lambda_b + \frac{1}{2}(\alpha_{n+2} + \beta_{n+1} b) + \frac{1}{2}(\alpha_{n+1} - \alpha_{n+2}) + \frac{b}{2}(\beta_{n+1} - \beta_n) \\ &= \frac{1}{2}(\alpha_{n+2} + \beta_{n+1} b) + \left(\lambda_b - \frac{1}{2}(R_{n+2}^\alpha - b R_{n+1}^\beta)\right). \end{aligned}$$

The inequality $\lambda_b \geqslant B(1+b)/2 \geqslant (R_{n+2}^\alpha - b R_{n+1}^\beta)/2$ then implies that $\delta_{n+1}^{\lambda_b} \geqslant \frac{1}{2}(\alpha_{n+2} + \beta_{n+1} b)$, which concludes the proof. □

**Lemma 26.** *The operators $A(b) - \lambda_b I$ and $(A(b) - \lambda_b I)^*$ are closed.*

*Proof.* Since $A(b) - \lambda_b I$ is a densely defined operator, $(A(b) - \lambda_b I)^*$ is closed. Let us prove that $A(b)$ is closed. Consider a sequence $(u^n)_{n\geqslant 0}$ in $D$ converging to $u$ in $\mathcal{H}$ such that $A(b)u^n \to v$ in $\mathcal{H}$. For every $k \in \mathbb{N}^*$, $u_k^n \to u_k$ and $(A(b)u^n)_k \to v_k$. Therefore, $(A(b)u^n)_k \to (A(b)u)_k$ and $(A(b)u)_k = v_k$. Hence, $u \in D(A(b))$ and $A(b)u = v$. This proves that $A(b)$ is closed and $A(b) - \lambda_b I$ too. □

**Proposition 27.** *Suppose that Assumption 1 holds. Then, the operators $A(b) - \lambda_b I$ and $(A(b) - \lambda_b I)^*$ are dissipative.*

*Proof.* We first prove by induction that, for all $u \in c_{00}$,

$$\langle (A(b) - \lambda_b I)u, u\rangle \leqslant 0, \tag{2.30}$$

and then conclude by a density argument. We define to this end the sequence space $c_{00}^n = \left\{u \in \mathbb{R}^{\mathbb{N}^*} \;\middle|\; \forall\, k > n, u_k = 0\right\}$, composed of sequences whose non-vanishing coefficients are the first $n$ components. We prove by induction the following statement (for the sequence $\delta_n^{\lambda_b}$ defined in (2.28)):

$$P(n): \qquad \forall u \in c_{00}^n, \qquad \langle (A(b) - \lambda_b I)u, u\rangle \leqslant -\delta_n^{\lambda_b} u_n^2 \leqslant 0.$$

This amounts to proving that the operator $P_n(A(b) - \lambda_b I)P_n$ is dissipative, where $P_n$ is the projection onto $c_{00}^n$ defined as $P_n u = (u_1, \cdots, u_n, 0, \cdots)$.

*Induction basis*: for $n = 1$, one simply gets, with $u = u_1 e_1 \in c_{00}^1$,

$$\langle (A(b) - \lambda_b I)u, u\rangle = -(\lambda_b + 2\beta_1 b)u_1^2 = -(\lambda_b + \beta_1 b + \alpha_1)u_1^2 - \beta_1 b u_1^2 \leqslant -\delta_1^{\lambda_b} u_1^2 \leqslant 0.$$

*Inductive step*: assume that $P(n)$ holds for some integer $n \geqslant 1$. Consider $u \in c_{00}^{n+1}$, with $u = \sum_{k=1}^{n+1} u_k e_k = \widehat{u_n} + u_{n+1}e_{n+1}$. Then,

$$\begin{aligned}
\langle (A(b) - \lambda_b I)u, u\rangle &= \langle (A(b) - \lambda_b I)\widehat{u_n}, \widehat{u_n}\rangle + \langle (A(b) - \lambda_b I)e_n, e_{n+1}\rangle u_{n+1}u_n \\
&\qquad + \langle (A(b) - \lambda_b I)e_{n+1}, e_n\rangle u_{n+1}u_n + \langle (A(b) - \lambda_b I)e_{n+1}, e_{n+1}\rangle u_{n+1}^2 \\
&= \langle (A(b) - \lambda_b I)\widehat{u_n}, \widehat{u_n}\rangle + (\beta_n b + \alpha_{n+1})u_{n+1}u_n - (\lambda_b + \beta_{n+1}b + \alpha_{n+1})u_{n+1}^2 \\
&\leqslant -\delta_n^{\lambda_b} u_n^2 + (\beta_n b + \alpha_{n+1})u_{n+1}u_n - (\lambda_b + \beta_{n+1}b + \alpha_{n+1})u_{n+1}^2,
\end{aligned}$$

using $P(n)$ with $\widehat{u_n} \in c_{00}^n$. Let $\mathcal{R}(u_n)$ be the second-order polynomial function defined for $u_{n+1}$ fixed as

$$\mathcal{R}(u_n) = -\delta_n^{\lambda_b} u_n^2 + (\beta_n b + \alpha_{n+1})u_{n+1}u_n - (\lambda_b + \beta_{n+1}b + \alpha_{n+1})u_{n+1}^2.$$

Since $-\delta_n^{\lambda_b} \leqslant 0$ (in view of Lemma 25), the maximum of $\mathcal{R}$ is attained for $u_n^{\max} = \frac{(\beta_n b + \alpha_{n+1})u_{n+1}}{2\delta_n^{\lambda_b}}$, so that

$$\mathcal{R}(u_n) \leqslant \mathcal{R}(u_n^{\max}) = -(\lambda_b + \beta_{n+1}b + \alpha_{n+1})u_{n+1}^2 + \frac{1}{4}\frac{(\beta_n b + \alpha_{n+1})^2}{\delta_n^{\lambda_b}}u_{n+1}^2 = -\delta_{n+1}^{\lambda_b} u_{n+1}^2.$$

This shows that $P(n+1)$ holds. At this stage, we have therefore proved that (2.30) holds. Since $c_{00} \subset D \subset \mathcal{H}$ is dense in $\mathcal{H}$ and $A(b) - \lambda_b I$ is a closed operator, we can conclude that $A(b) - \lambda_b I$ is a dissipative operator.

Since the coefficients of $A(b)$ are real-valued, it holds $\langle (A(b) - \lambda_b I)u, u\rangle = \langle (A(b) - \lambda_b I)^* u, u\rangle$ for any $u \in c_{00}$, which allows to conclude that $(A(b) - \lambda_b I)^*$ is dissipative too. □

We are now in position to prove the uniqueness of the solution to the Cauchy problem (CD).

*Proof of Theorem 3 – Uniqueness.* In order to prove the uniqueness of the solution, we consider two solutions and prove that they are equal using the dissipativity of $A$ and a Gronwall argument. Let $u$ and $v$ be two solutions of (CD) with initial condition $u^0$. Then, $u - v$ is solution of

$$\frac{d(u-v)}{dt} = A(u_1)(u-v) - (A(v_1) - A(u_1))v = A(u_1)(u-v) - (v_1 - u_1)A_{\mathrm{L}}^{\beta} v,$$

so that

$$\frac{d\|u-v\|^2}{dt} = 2\langle A(u_1)(u-v), u-v\rangle + 2(v_1-u_1)\langle A_{\mathrm{L}}^{\beta} v, u-v\rangle.$$

Then, in view of Proposition 27, and since $u_1(t) \leqslant Q_0$ for all $t \geqslant 0$, it holds

$$\langle A(u_1)(u-v), u-v\rangle \leqslant \lambda_{Q_0}\|u-v\|^2.$$

Moreover, using a Cauchy-Schwarz inequality, the equalities $Q(u(t)) = Q(v(t)) = Q_0$ and Lemma 32,

$$(v_1-u_1)\langle A_{\mathrm{L}}^{\beta} v, u-v\rangle \leqslant |u_1 - v_1| K(Q_0)\|u-v\| \leqslant K(Q_0)\|u-v\|^2.$$

Therefore,

$$\frac{d\|u-v\|^2}{dt} \leqslant (\lambda_{Q_0} + K(Q_0))\|u-v\|^2. \tag{2.31}$$

Since $u(0) = v(0)$, we conclude that $u(t) - v(t) = 0$ for all $t \geqslant 0$ by a Gronwall inequality. □

## 2.B Proofs of the results of Section 2.2.2

This section is organized as follows. In Section 2.B.1, we use the fact that the linear operator $A(b)$ is dissipative for every $b \geqslant 0$, to obtain estimates on the sub-dynamics (2.11). In Section 2.B.2 we give estimates on elements of $\mathcal{Q}$ and prove Proposition 5. In Section 2.B.3 we prove the convergence of the splitting.

### 2.B.1 Estimates on the subdynamics (2.11)

In this section we consider the operator $A(b)$ when the parameter $b$ is fixed to a constant value in $\mathbb{R}_+$. Note that linear CD can be rewritten as the following Cauchy problem:

$$\begin{cases} \dfrac{du}{dt} = A(b)u, \\ u(0) = u^0. \end{cases} \tag{LCD}$$

Using the fact that $A$ is dissipative and using standard results of the theory of semi-groups (LCD) has a unique classical solution in $\mathcal{C}^0(\mathbb{R}_+, D(A(b))) \cap \mathcal{C}^1(\mathbb{R}_+, \mathcal{H})$ when $u^0 \in D(A(b))$ (see [Paz12, Chapter 4, Theorem 1.3]). This is summarized in the following result.

**Proposition 28.** *The operator $A(b)$ is the infinitesimal generator of a strongly continuous semigroup $(T_b(t))_{t\in\mathbb{R}_+}$. For all $u^0 \in D(A(b))$, the problem* (LCD) *therefore has a unique solution $u \in \mathcal{C}^0(\mathbb{R}_+, D(A(b))) \cap \mathcal{C}^1(\mathbb{R}_+, \mathcal{H})$ defined as $u(t) = T_b(t)u^0$ for all $t \geqslant 0$. Moreover the following a priori estimates hold true:*

$$\forall\, t \geqslant 0, \qquad \|u(t)\| \leqslant \mathrm{e}^{\lambda_b t}\left\|u^0\right\|, \qquad \left\|\frac{du}{dt}(t)\right\| = \|A(b)u(t)\| \leqslant \mathrm{e}^{\lambda_b t}\left\|A(b)u^0\right\|.$$

We next give *a priori* estimates on the sub-dynamics (2.11) which are useful for the proof of Proposition 4. Introducing the projection $\Pi$ such that

$$\forall\, u = (u_i)_{i\geqslant 1} \in \mathcal{H}, \qquad \Pi u = (0, u_2, u_3, \cdots),$$

we can define the operator $A^\Pi(u_1)$ as $A^\Pi(u_1) = \Pi A(u_1)$. The sub-dynamics (2.11) can then be written compactly as the following linear evolution problem:

$$\begin{cases} \dfrac{du}{dt} = A^\Pi\left(u_1^0\right)u, \\ u(0) = u^0. \end{cases} \tag{2.32}$$

The following results are direct consequences of Proposition 28 and [Paz12, Chapter 4, Corollary 2.5].

**Proposition 29.** *Fix $u_1 \geqslant 0$ and suppose that Assumption 1 holds. Then, the operator $A^\Pi(u_1)$ is the infinitesimal generator of a strongly continuous semigroup $(T_{u_1}(t))_{t\in\mathbb{R}_+}$. The problem* (2.11) *therefore has a unique solution $u \in \mathcal{C}^0(\mathbb{R}_+, D(A(u_1))) \cap \mathcal{C}^1(\mathbb{R}_+, \mathcal{H})$ for all $u^0 \in D(A(u_1))$, and $u(t) = T_{u_1}(t)u^0$ for all $t \geqslant 0$. Moreover, there exists $\lambda_{u_1} \geqslant 0$ such that the following a priori estimates hold true:*

$$\forall\, t \geqslant 0, \qquad \|u(t)\| \leqslant \mathrm{e}^{\lambda_{u_1} t}\left\|u^0\right\| \qquad \textit{and} \qquad \left\|\frac{du}{dt}(t)\right\| = \left\|A^\Pi(u_1)u(t)\right\| \leqslant \mathrm{e}^{\lambda_{u_1} t}\left\|A^\Pi(u_1)u^0\right\|.$$

*Finally, fix $T > 0$ and consider $f \in \mathcal{C}^1([0,T], \mathcal{H})$. Then, the problem*

$$\begin{cases} \dfrac{du}{dt} = A^\Pi(u_1)u + f, \\ u(0) = u^0, \end{cases}$$

*has a unique classical solution $u \in \mathcal{C}^0([0,T], D(A(u_1))) \cap \mathcal{C}^1([0,T], \mathcal{H})$ defined as*

$$\forall\, t \in [0,T], \qquad u(t) = T_{u_1}(t)u^0 + \int_0^t T_{u_1}(t-s)f(s)\,ds.$$

**Remark 30.** *In fact, the dependence of $\lambda_{u_1}$ on $u_1$ can be made precise, see* (2.29).

Finally, in order to prove the convergence of the splitting, we need estimates in $\mathcal{Q}$ of the solutions of (2.11).

**Lemma 31.** *Fix an initial condition $u^0 \in \mathcal{Q}$ and $b \geqslant 0$. Suppose that Assumptions 1 and 2 hold true. Then, the unique classical solution of the second sub-dynamics $u : t \mapsto \chi_t^b(u^0) \in \mathcal{C}^0([0,T], D(A(b))) \cap \mathcal{C}^1([0,T], \mathcal{H})$ remains in $\mathcal{Q}$. Moreover,*

$$Q(u(t)) \leqslant Q(u(0))\exp(2bKt).$$

*Proof.* We first note that all components of $u$ remain non-negative since the dynamics $t \mapsto \chi_t^b(u)$ is in fact the Kolmogorov forward equation of a Markov jump process [Ter+17]. Then, proceeding as in the proof of Theorem 3, we see that $Q(u(t)) = b + \sum_{n\geqslant 2} n u_n(t)$ is well defined, continuously differentiable, and

$$\frac{d}{dt}[Q(u(t))] = 2\beta_1 b^2 + \sum_{n\geqslant 2} b\beta_n u_n(t) - \sum_{n\geqslant 2} \alpha_n u_n(t) - \alpha_2 u_2(t)$$

$$\leqslant 2bK\left(b + \sum_{n\geqslant 2} n^\gamma u_n(t)\right) \leqslant 2bKQ(t).$$

The claimed estimate then follows from a Gronwall inequality. □

## 2.B.2 Some estimates on elements of $\mathcal{Q}$

We first state estimates for elements of the set $\mathcal{Q}$ introduced in (2.9).

**Lemma 32.** *Fix $Q_* \in \mathbb{R}_+$ and suppose that Assumptions 1 and 2 hold true. Then, there exists $R(Q_*) \in \mathbb{R}_+$ such that, for any $w \in \mathcal{Q}$ with $Q(w) \leqslant Q_*$,*

$$\|A_{\mathrm{L}}^\alpha w\|, \left\|A_{\mathrm{L}}^\beta w\right\| \leqslant R(Q_*),$$

*and*

$$\left\|(A_{\mathrm{L}}^\alpha)^2 w\right\|, \left\|A_{\mathrm{L}}^\alpha A_{\mathrm{L}}^\beta w\right\|, \left\|A_{\mathrm{L}}^\beta A_{\mathrm{L}}^\alpha w\right\|, \left\|\left(A_{\mathrm{L}}^\beta\right)^2 w\right\| \leqslant R(Q_*).$$

*Proof.* Fix $w \in \mathcal{Q}$ such that $Q(w) \leqslant Q_*$. Note first that the following bound holds:

$$\forall\, n \geqslant 1, \qquad 0 \leqslant w_n \leqslant \frac{Q_*}{n}.$$

Then, using Assumption 1 to bound $\alpha_n$ as $0 \leqslant \alpha_n \leqslant Kn$ for $n \geqslant 1$,

$$(A_{\mathrm{L}}^\alpha w)_1^2 = \left(\sum_{n\geqslant 2} \alpha_n w_n + \alpha_2 w_2\right)^2 \leqslant \left(2K \sum_{n\geqslant 2} n w_n\right)^2 \leqslant (2KQ_*)^2.$$

Moreover,

$$\sum_{n\geqslant 2} (A_{\mathrm{L}}^\alpha w)_n^2 = \sum_{n\geqslant 2} (\alpha_{n+1} w_{n+1} - \alpha_n w_n)^2 \leqslant 4 \sum_{n\geqslant 2} (\alpha_n w_n)^2$$

$$\leqslant 4 \sum_{n\geqslant 2} \alpha_n^2 \frac{Q_*}{n^2}(n w_n) \leqslant 4K^2 Q_* \sum_{n\geqslant 2} n w_n.$$

Therefore,

$$\sum_{n\geqslant 2} (A_{\mathrm{L}}^\alpha w)_n^2 \leqslant 4K^2 Q_* \sum_{n\geqslant 2} n w_n \leqslant 4K^2 Q_*^2,$$

which gives us $\|A_{\mathrm{L}}^\alpha w\| \leqslant 2\sqrt{2}KQ_*$. A similar reasoning can be used to bound $\left\|A_{\mathrm{L}}^\beta w\right\|$. Let us next consider $\left\|(A_{\mathrm{L}}^\alpha)^2 w\right\|$. Since $\alpha$ is non-decreasing, it holds

$$|(A_{\mathrm{L}}^\alpha A_{\mathrm{L}}^\alpha w)_1| = \left|\sum_{n\geqslant 2} \alpha_n (A_{\mathrm{L}}^\alpha w)_n + \alpha_2 (A_{\mathrm{L}}^\alpha w)_2\right| \leqslant 2 \sum_{n\geqslant 2} \alpha_n |\alpha_{n+1} w_{n+1} - \alpha_n w_n| \leqslant 4 \sum_{n\geqslant 2} \alpha_n^2 w_n.$$

In view of Assumption 2, it holds, with $0 \leqslant \gamma \leqslant 1/2$,

$$\sum_{n\geqslant 2} \alpha_n^2 w_n \leqslant K^2 \sum_{n\geqslant 2} n^{2\gamma-1} n w_n \leqslant K^2 \sum_{n\geqslant 2} n w_n \leqslant K^2 Q_*.$$

Moreover, for $n \geqslant 2$,

$$\begin{aligned}
|(A_{\mathrm{L}}^\alpha A_{\mathrm{L}}^\alpha w)_n| &= \left|\alpha_{n+1}(A_{\mathrm{L}}^\alpha w)_{n+1} - \alpha_n (A_{\mathrm{L}}^\alpha w)_n\right| \\
&= \left|\alpha_{n+2}\alpha_{n+1} w_{n+2} - \left(\alpha_{n+1}^2 + \alpha_{n+1}\alpha_n\right) w_{n+1} + \alpha_n^2 w_n\right| \\
&\leqslant \alpha_{n+2}^2 w_{n+2} + 2\alpha_{n+1}^2 w_{n+1} + \alpha_n^2 w_n.
\end{aligned}$$

Therefore, since $0 \leqslant \alpha_n^4/n^2 \leqslant K^4$ and $n w_n \leqslant Q_*$, it holds

$$\sum_{n\geqslant 2} (A_{\mathrm{L}}^\alpha A_{\mathrm{L}}^\alpha w)_n^2 \leqslant 16 \sum_{n\geqslant 2} \alpha_n^4 w_n^2 \leqslant 16 \sum_{n\geqslant 2} \frac{\alpha_n^4}{n^2} (n w_n)^2 \leqslant 16 K^4 Q_* \sum_{n\geqslant 2} n w_n \leqslant 16 K^4 Q_*^2.$$

In conclusion, $\|A_{\mathrm{L}}^\alpha A_{\mathrm{L}}^\alpha w\| \leqslant 4\sqrt{2} K^2 Q_*$. Similar computations can be performed for $A_{\mathrm{L}}^\alpha A_{\mathrm{L}}^\beta w$, $A_{\mathrm{L}}^\beta A_{\mathrm{L}}^\alpha w$ and $A_{\mathrm{L}}^\beta A_{\mathrm{L}}^\beta w$, which leads to the claimed estimates. □

Let us next prove the technical estimates provided by Proposition 5.

*Proof of Proposition 5.* Fix a time $T > 0$, a constant $Q_* \in \mathbb{R}_+$ and a non-negative initial condition $u^0 \in \mathcal{Q}$. Suppose that the total quantity of matter of the initial condition satisfies $Q(u^0) \leqslant Q_*$ and denote by $|\cdot|_{\mathcal{C}^0}$ the uniform norm for functions in $\mathcal{C}^0([0,T],\mathbb{R})$, *i.e.*

$$\forall\, f \in \mathcal{C}^0([0,T],\mathbb{R}), \qquad |f|_{\mathcal{C}^0} = \sup_{0\leqslant t\leqslant T} |f(t)|.$$

Recall that the total quantity of matter is conserved, so that $Q(u(t)) \leqslant Q_*$. Since $u$ stays non-negative, it holds $0 \leqslant u_n(t) \leqslant Q_*/n$ for all $t \geqslant 0$ and $n \geqslant 1$. In particular, $|u_1|_{\mathcal{C}^0} \leqslant Q_*$. Therefore, for all $0 \leqslant t \leqslant T$,

$$\left\|\frac{du}{dt}(t)\right\| = \left\|A_{\mathrm{L}}^\alpha u(t) + u_1(t) A_{\mathrm{L}}^\beta u(t)\right\| \leqslant \|A_{\mathrm{L}}^\alpha u(t)\| + Q_* \left\|A_{\mathrm{L}}^\beta u(t)\right\| \leqslant (1+Q_*) R(Q_*),$$

which concludes the proof of the bound for $du/dt$ in view of Lemma 32. In particular,

$$\left|\frac{du_1}{dt}\right|_{\mathcal{C}^0} \leqslant \sup_{0\leqslant t\leqslant T} \left\|\frac{du}{dt}(t)\right\| \leqslant (1+Q_*) R(Q_*). \tag{2.33}$$

Then, for all $0 \leqslant t \leqslant T$,

$$\begin{aligned}
\left\|\frac{d^2u}{dt^2}(t)\right\| &\leqslant \left\|A_{\mathrm{L}}^\alpha \left(A_{\mathrm{L}}^\alpha + u_1(t) A_{\mathrm{L}}^\beta\right) u(t)\right\| + |u_1|_{\mathcal{C}^0} \left\|A_{\mathrm{L}}^\beta \left(A_{\mathrm{L}}^\alpha + u_1(t) A_{\mathrm{L}}^\beta\right) u(t)\right\| \\
&\quad + \left|\frac{du_1}{dt}\right|_{\mathcal{C}^0} \left\|A_{\mathrm{L}}^\beta u(t)\right\| \\
&\leqslant \left\|(A_{\mathrm{L}}^\alpha)^2 u(t)\right\| + |u_1|_{\mathcal{C}^0} \left(\left\|A_{\mathrm{L}}^\alpha A_{\mathrm{L}}^\beta u(t)\right\| + \left\|A_{\mathrm{L}}^\beta A_{\mathrm{L}}^\alpha u(t)\right\|\right) \\
&\quad + |u_1|_{\mathcal{C}^0}^2 \left\|\left(A_{\mathrm{L}}^\beta\right)^2 u(t)\right\| + \left|\frac{du_1}{dt}\right|_{\mathcal{C}^0} \left\|A_{\mathrm{L}}^\beta u(t)\right\|,
\end{aligned}$$

from which we obtain the estimate for $d^2u/dt^2$ in view of Lemma 32 and (2.33). □

## 2.B.3 Proof of Proposition 4

We can now write the proof of the convergence of the splitting of the dynamics. The proof can be decomposed in three steps. We first prove the consistency of the splitting for elements of $\mathcal{Q}$ which are bounded in an appropriate norm. We next prove its stability, under the same conditions on elements of $\mathcal{Q}$. We finally conclude to the convergence for arbitrary times using the fact that solutions of (CD) are uniformly bounded.

**Step 0: Technical results on $\varphi$.** Let us recall that the flow $\varphi_t$ defined in (2.10) acts only upon the first component of an element $v \in \mathcal{H}$. For $v \in \mathcal{Q}$, denote by

$$a = 2\beta_1, \qquad b(v) = \sum_{n\geqslant 2} \beta_n v_n, \qquad c(v) = \sum_{n\geqslant 2} \alpha_n v_n + \alpha_2 v_2, \tag{2.34}$$

where $a > 0$ and $b(v), c(v) \geqslant 0$ are fixed. The dynamics on $t \mapsto \varphi_t^{(v_2,\cdots)}(v_1)$ therefore writes

$$\frac{d\varphi_t^{(v_2,\cdots)}(v_1)}{dt} = -a\Big(\varphi_t^{(v_2,\cdots)}(v_1)\Big)^2 - b(v)\varphi_t^{(v_2,\cdots)}(v_1) + c(v). \tag{2.35}$$

In order to prove stability and consistency results on the flow $\varphi_t$, we need the following technical results, whose proofs are given at the end of this section.

**Lemma 33.** *Fix a time $t \geqslant 0$ and $Q_* > 0$. Then, there is $\mathcal{B}(Q_*) \in \mathbb{R}_+$ such that, for any $v \in \mathcal{Q}$ with $Q(v) \leqslant Q_*$ and $(v_2, v_3 \cdots) \neq (0, 0, \cdots)$, it holds $\varphi_t^{(v_2,\cdots)}(v_1) \geqslant 0$ for all $t \geqslant 0$, and*

$$\left|\varphi_t^{(v_2,\cdots)}(v_1)\right|, \left|\frac{d\varphi_t^{(v_2,\cdots)}(v_1)}{dt}\right|, \left|\frac{d^2\varphi_t^{(v_2,\cdots)}(v_1)}{dt^2}\right| \leqslant \mathcal{B}(Q_*). \tag{2.36}$$

**Lemma 34.** *Consider $v \in \mathcal{Q}$ and suppose that there exists $k \geqslant 2$ such that $v_k > 0$. Then, for all $t \geqslant 0$, there exists $\ell \geqslant 2$ such that $(\chi_t(v))_\ell > 0$.*

**Step 1: Consistency**

We prove that, for any $Q_* > 0$, there exists a constant $L_1(Q_*) \in \mathbb{R}_+$ such that, for all $u^0 \in \mathcal{Q}$ with $Q(u^0) \leqslant Q_*$, it holds

$$\forall\, 0 \leqslant \Delta t \leqslant 1, \qquad \Big\|S_{\Delta t}\Big(u^0\Big) - u(\Delta t)\Big\| \leqslant L_1(Q_*)\Delta t^2, \tag{2.37}$$

where $u(t)$ is the solution of (CD) at time $t$ with initial condition $u^0$. We first estimate the error between $u_1(\Delta t)$ and $\varphi_{\Delta t}(u_1^0)$, before quantifying the error between $u(\Delta t)$ and $S_{\Delta t}(u^0)$. In the remainder of this part, we fix $Q_* > 0$. Moreover, recall that the flows $\varphi$ and $\chi$ preserve the non-negativity (see the proofs of Lemmas 33 and 31 respectively).

**Step 1.1: Error estimate on $u_1(\Delta t) - \varphi_{\Delta t}(u_1^0)$.** We first show that there is $P_1(Q_*) \in \mathbb{R}_+$ such that, for all $u^0 \in \mathcal{Q}$ with $Q(u^0) \leqslant Q_*$, it holds

$$\forall\, 0 \leqslant \Delta t \leqslant 1, \qquad \Big|u_1(\Delta t) - \varphi_{\Delta t}\Big(u_1^0\Big)\Big| \leqslant P_1(Q_*)\Delta t^2. \tag{2.38}$$

The dynamics on $t \mapsto \varphi_t(u_1^0)$ reads

$$\frac{d\varphi_t(u_1^0)}{dt} = -a\Big(\varphi_t\Big(u_1^0\Big)\Big)^2 - b(u(0))\varphi_t\Big(u_1^0\Big) + c(u(0)), \qquad \varphi_0\Big(u_1^0\Big) = u_1^0,$$

while the one on $t \mapsto u_1(t)$ reads

$$\frac{du_1}{dt} = -au_1(t)^2 - b(u(t))u_1(t) + c(u(t)), \qquad u_1(0) = u_1^0,$$

where $b$ and $c$ are defined in (2.34). Since $t \mapsto u_1(t)$ and $t \mapsto \varphi_t(u_1^0)$ are twice continuously differentiable (see Proposition 5 and Lemma 33), and

$$\frac{du_1}{dt}(0) = \left.\frac{d\varphi_t(u_1^0)}{dt}\right|_{t=0},$$

it follows that

$$\left|u_1(\Delta t) - \varphi_{\Delta t}\left(u_1^0\right)\right| \leqslant \frac{1}{2}\Delta t^2\left[\sup_{0\leqslant\theta\leqslant\Delta t}\left|\frac{d^2u_1}{dt^2}(\theta)\right| + \sup_{0\leqslant\theta\leqslant\Delta t}\left|\frac{d^2\varphi_t(u_1^0)}{dt^2}\right|\right].$$

The second order derivative $d^2\varphi_t(u_1^0)/dt^2$ is uniformly bounded in time by $\mathcal{B}(Q_*)$ (see (2.36)). Moreover, in view of Proposition 5, $d^2u_1/dt^2$ is also uniformly bounded in time, by a constant which depends on $Q_*$. This leads to (2.38).

**Step 1.2: Error estimates on $\Pi(S_{\Delta t}(u^0) - u(\Delta t))$.** We prove that there is $P_2(Q_*) \in \mathbb{R}_+$ such that, for all $u^0 \in \mathcal{Q}$ with $Q(u^0) \leqslant Q_*$, it holds

$$\forall\, 0 \leqslant \Delta t \leqslant 1, \qquad \left\|\Pi\left(u(\Delta t) - S_{\Delta t}\left(u^0\right)\right)\right\| \leqslant P_2(Q_*)\Delta t^2. \tag{2.39}$$

Let us first reinterpret $\Pi S_{\Delta t}$ as the flow of some time continuous dynamics. We rewrite to this end (2.32) as

$$\frac{d\chi_t(u^0)}{dt} = \left(\Pi A_{\mathrm{L}}^{\alpha} + \varphi_{\Delta t}\left(u_1^0\right)\Pi A_{\mathrm{L}}^{\beta}\right)\chi_t\left(u^0\right), \qquad \chi_0\left(u^0\right) = \left(\varphi_{\Delta t}\left(u_1^0\right), u_2^0, \cdots\right).$$

Consider $\widetilde{S}_t = \chi_t \circ \varphi_{\Delta t}$ and note that $S_{\Delta t} = \widetilde{S}_{\Delta t}$. Moreover, $w = \Pi\left(u - \widetilde{S}_t(u^0)\right)$ is solution of

$$\frac{dw}{dt}(t) = A^{\Pi}\left(\varphi_{\Delta t}\left(u_1^0\right)\right)w(t) + \left(u_1(t) - \varphi_{\Delta t}\left(u_1^0\right)\right)\Pi A_{\mathrm{L}}^{\beta}u(t), \qquad w(0) = 0.$$

Using Proposition 29, since $t \mapsto u_1(t) - \varphi_{\Delta t}(u_1^0)$ and $t \mapsto A_{\mathrm{L}}^{\beta}u(t)$ are continuously differentiable, we can write

$$w(\Delta t) = \int_0^{\Delta t} T_{\varphi_{\Delta t}(u_1^0)}(\Delta t - s)\left[u_1(s) - \varphi_{\Delta t}\left(u_1^0\right)\right]\Pi A_{\mathrm{L}}^{\beta}u(s)\, ds,$$

so that

$$\|w(\Delta t)\| \leqslant \Delta t \sup_{0\leqslant s\leqslant\Delta t}\left\|T_{\varphi_{\Delta t}(u_1^0)}(s)\right\| \sup_{0\leqslant s\leqslant\Delta t}\left\{\left|u_1(s) - \varphi_{\Delta t}\left(u_1^0\right)\right|\left\|\Pi A_{\mathrm{L}}^{\beta}u(s)\right\|\right\}.$$

Since $\varphi_t(u_1^0)$ is bounded by $\mathcal{B}(Q_*)$ (see (2.36)), in view of Proposition 28, it holds, for any $0 \leqslant s \leqslant \Delta t$,

$$\left\|T_{\varphi_{\Delta t}(u_1^0)}(s)\right\| \leqslant \exp\left(\lambda_{\mathcal{B}(Q_*)}\Delta t\right).$$

Moreover, in view of Proposition 5 and Lemma 33, and since $u_1(0) = u_1^0$,

$$\left|u_1(s) - \varphi_{\Delta t}\left(u_1^0\right)\right| \leqslant \Delta t \left( \left| \frac{du_1}{dt} \right|_{\mathcal{C}^0([0,\Delta t])} + \left| \frac{d\varphi_t(u_1^0)}{dt} \right|_{\mathcal{C}^0([0,\Delta t])} \right) \leqslant \Delta t (R(Q_*) + \mathcal{B}(Q_*)).$$

Finally, in view of Lemma 32, since $Q(u(s)) = Q(u^0) \leqslant Q_*$ for any $0 \leqslant s \leqslant \Delta t$ (see Theorem 3), we obtain $\left\|\Pi A_{\mathrm{L}}^{\beta} u(s)\right\| \leqslant R(Q_*)$. Then,

$$\left\|\Pi\left(u(\Delta t) - S_{\Delta t}\left(u^0\right)\right)\right\| \leqslant \Delta t^2 \exp\left(\lambda_{\mathcal{B}(Q_*)} \Delta t\right) [R(Q_*) + \mathcal{B}(Q_*)] R(Q_*),$$

which leads to (2.39).

**Step 1.3: The splitting is consistent.** Consider now $u^0 \in \mathcal{Q}$ with $Q(u^0) \leqslant Q_*$. We first note that $\|S_{\Delta t}(u^0) - u(\Delta t)\|^2 = \|(u_1(\Delta t) - \varphi_{\Delta t}(u^0), 0, \cdots)\|^2 + \|\Pi(u(\Delta t) - S_{\Delta t}(u^0))\|^2$. The estimate (2.37) then follows from (2.38) and (2.39).

**Step 2: Stability**

We prove that, for any $Q_* > 0$, there exists $L_2(Q_*) \in \mathbb{R}_+$, such that for all $u, v \in \mathcal{Q}$ with $Q(u), Q(v) \leqslant Q_*$, it holds

$$\forall\, 0 \leqslant \Delta t \leqslant 1, \qquad \|S_{\Delta t}(u) - S_{\Delta t}(v)\| \leqslant \exp(L_2(Q_*)\Delta t)\|u - v\| + 2L_1(Q_*)\Delta t^2. \quad (2.40)$$

Fix $Q_* > 0$ and consider $u, v \in \mathcal{Q}$ with $Q(u), Q(v) \leqslant Q_*$. Denote by $\widetilde{u}, \widetilde{v}$ the solutions of (CD) with initial conditions $u, v$ respectively. Then,

$$\|S_{\Delta t}(u) - S_{\Delta t}(v)\| \leqslant \|S_{\Delta t}(u) - \widetilde{u}(\Delta t)\| + \|\widetilde{u}(\Delta t) - \widetilde{v}(\Delta t)\| + \|S_{\Delta t}(u) - \widetilde{v}(\Delta t)\|,$$

where $\|S_{\Delta t}(u) - \widetilde{u}(\Delta t)\| + \|S_{\Delta t}(u) - \widetilde{v}(\Delta t)\| \leqslant 2L_1(Q_*)\Delta t^2$ in view of (2.37). Moreover, in view of (2.31), it holds

$$\frac{d\|\widetilde{u} - \widetilde{v}\|^2}{dt} \leqslant (\lambda_{Q_*} + K(Q_*))\|\widetilde{u} - \widetilde{v}\|^2.$$

Therefore, using a Gronwall inequality, there exists $L_2(Q_*) \in \mathbb{R}_+$ such that

$$\|\widetilde{u}(\Delta t) - \widetilde{v}(\Delta t)\| \leqslant \exp(L_2(Q_*)\Delta t)\|u - v\|. \quad (2.41)$$

The estimate (2.40) follows by combining (2.41) together with the consistency estimates given by (2.37).

**Step 3: Convergence**

The convergence of the splitting as $\Delta t \to 0$ classically follows from the stability and consistency estimates obtained in Steps 1 and 2. We however first need to make sure that $Q(u^n) \leqslant Q_*$ in order to apply (2.37) and (2.40) for a well-chosen $Q_*$. The proof proceeds by induction.

Fix a time $\tau > 0$, an initial condition $u^0 \in \mathcal{Q}$ and let $Q_0 = Q(u^0)$. Our aim is to prove that

there exist $\Delta t_* > 0$ and $K(\tau, u^0) \in \mathbb{R}_+$ such that

$$\forall\, 0 < \Delta t \leqslant \Delta t_*, \qquad \forall\, 0 \leqslant n \leqslant \frac{\tau}{\Delta t}, \qquad \|u(n\Delta t) - u^n\| \leqslant K(\tau, u^0)\Delta t.$$

Let $M_* = 2\sup_{0\leqslant t\leqslant \tau} \|u(t)\|$ and consider $Q_* \geqslant 2Q_0 \exp(2KM_*\tau)$. We also introduce the constant $K(\tau, u^0) = 3L_1(Q_*)\tau \exp(L_2(Q_*)\tau)$ where the prefactors $L_i(Q_*)$ are the ones appearing in (2.37) and (2.40), and the time step

$$\Delta t_* = \min\left(1, \frac{M_*}{2K(\tau, u^0)}, \frac{\ln(2)}{2K\mathcal{B}(Q_*)\tau}\right), \tag{2.42}$$

where $\mathcal{B}(Q_*)$ is the constant appearing in Lemma 33. Fix $0 < \Delta t \leqslant \Delta t_*$. We prove by induction that, for $0 \leqslant n \leqslant \tau/\Delta t$, it holds

$$\begin{aligned} \forall\, 0 \leqslant k \leqslant n, \quad Q\left(u^k\right) &\leqslant Q_0 \exp\left(2K(M_* + \Delta t\mathcal{B}(Q_*))k\Delta t\right), \\ \left\|u^k\right\| \leqslant M_*, \quad \left\|u(k\Delta t) - u^k\right\| &\leqslant K(\tau, u^0)\Delta t. \end{aligned} \tag{2.43}$$

The induction basis is clear since $u^0 = u(0)$. Assume now that (2.43) holds for some integer $0 \leqslant n \leqslant \tau/\Delta t$ such that $n + 1 \leqslant \tau/\Delta t$. First, in view of Lemma 31,

$$Q(S_{\Delta t}(u^n)) \leqslant Q(u^n) \exp\left(2K\varphi_{\Delta t}(u_1^n)\Delta t\right).$$

Moreover, in view of Lemma 33, we also have $\varphi_{\Delta t}(u_1^n) \leqslant |u_1^n| + \Delta t\mathcal{B}(Q_*)$ and $|u_1^n| \leqslant \|u^n\| \leqslant M_*$. Then, by the induction hypothesis, it holds

$$Q\left(u^{n+1}\right) \leqslant Q_0 \exp\left(2K(M_* + \Delta t\mathcal{B}(Q_*))(n+1)\Delta t\right).$$

Therefore,

$$Q\left(u^{n+1}\right) \leqslant Q_0 \exp\left(2KM_*\tau\right) \exp(2K\mathcal{B}(Q_*)\tau\Delta t) \leqslant Q_*,$$

where we used the fact that $\Delta t \leqslant \ln(2)(2K\mathcal{B}(Q_*)\tau)^{-1}$ so that $\exp(2K\mathcal{B}(Q_*)\tau\Delta t) \leqslant 2$, as well as the fact that $Q_0 \exp(2KM_*\tau) \leqslant Q_*/2$.

We are then in position to prove the two other inequalities in (2.43). Note that $Q(u(t)) = Q_0 \leqslant Q_*$ for all $t \geqslant 0$. Therefore, in view of the estimates (2.37) and (2.40), for all $0 < \Delta t \leqslant \Delta t_*$ and all $0 \leqslant n \leqslant \tau/\Delta t$, it holds

$$\begin{aligned} \left\|u((n+1)\Delta t) - u^{n+1}\right\| &\leqslant \|u((n+1)\Delta t) - S_{\Delta t}(u(n\Delta t))\| + \left\|S_{\Delta t}(u(n\Delta t)) - u^{n+1}\right\| \\ &\leqslant 3L_1(Q_*)\Delta t^2 + \exp(L_2(Q_*)\Delta t)\|u(n\Delta t) - u^n\|. \end{aligned}$$

Since $u(0) = u^0$, we obtain by recursion

$$\begin{aligned} \left\|u((n+1)\Delta t) - u^{n+1}\right\| &\leqslant 3L_1(Q_*)\Delta t^2 \sum_{k=0}^{n} \exp(L_2(Q_*)k\Delta t) \\ &\leqslant 3L_1(Q_*)\tau \exp(L_2(Q_*)\tau)\Delta t = K(\tau, Q_*)\Delta t, \end{aligned}$$

from which the last inequality in (2.43) follows for $n + 1$. Finally, using a reverse triangle

inequality and (2.42), it holds

$$\left\|u^{n+1}\right\| \leqslant K(\tau, Q_*)\Delta t + \frac{1}{2}M_* \leqslant M_*,$$

which concludes the proof. □

Let us conclude this section by providing the proof of Lemma 33.

*Proof of Lemma 33.* We first give an analytic expression of the solution of (2.35) and then prove the estimates (2.36). Since $v \in \mathcal{Q}$, the quantities $b(v)$ and $c(v)$ are finite. Moreover, since there exists $k \geqslant 2$ such that $v_k > 0$, $\delta(v)^2 = b(v)^2 + 4ac(v) > 0$. Then,

$$\frac{d\varphi_t^{(v_2,\cdots)}(v_1)}{dt} = -a\Big(\varphi_t^{(v_2,\cdots)}(v_1) - r_+(v)\Big)\Big(\varphi_t^{(v_2,\cdots)}(v_1) - r_-(v)\Big),$$

where

$$r_+(v) = -\frac{b(v) - \sqrt{b(v)^2 + 4ac(v)}}{2a} > 0 \quad \text{and} \quad r_-(v) = -\frac{b(v) + \sqrt{b(v)^2 + 4ac(v)}}{2a} < 0. \tag{2.44}$$

Formally,

$$\frac{d\varphi_t^{(v_2,\cdots)}(v_1)}{(\varphi_t^{(v_2,\cdots)}(v_1) - r_+(v))(\varphi_t^{(v_2,\cdots)}(v_1) - r_-(v))} = -adt,$$

and using a partial fraction decomposition we have

$$\frac{a}{\delta(v)}\frac{d\varphi_t^{(v_2,\cdots)}(v_1)}{\varphi_t^{(v_2,\cdots)}(v_1) - r_+(v)} - \frac{a}{\delta(v)}\frac{d\varphi_t^{(v_2,\cdots)}(v_1)}{\varphi_t^{(v_2,\cdots)}(v_1) - r_-(v)} = -adt.$$

Therefore,

$$\log\left(\frac{\varphi_t^{(v_2,\cdots)}(v_1) - r_+(v)}{v_1 - r_+(v)}\right) - \log\left(\frac{\varphi_t^{(v_2,\cdots)}(v_1) - r_-(v)}{v_1 - r_-(v)}\right) = -\delta(v)t,$$

which gives us, with $\omega(v) = \frac{v_1 - r_+(v)}{v_1 - r_-(v)}$,

$$\varphi_t^{(v_2,\cdots)}(v_1) - r_+(v) = \Big(\varphi_t^{(v_2,\cdots)}(v_1) - r_-(v)\Big)\omega(v)\exp\left(-\delta(v)t\right),$$

so that

$$\varphi_t^{(v_2,\cdots)}(v_1) = \frac{r_+(v) - r_-(v)\omega(v)\exp(-\delta(v)t)}{1 - \omega(v)\exp(-\delta(v)t)}.$$

One verifies that $t \mapsto \varphi_t^{(v_2,\cdots)}(v_1)$ is a solution, hence it is unique thanks to Cauchy-Lipschitz's Theorem. We compare in Figure 2.3 the analytical solution (2.45) with a numerically integrated solution of (2.35) to make sure that our expressions are correct. Therefore, the unique solution of (2.35) reads

$$\varphi_t^{(v_2,\cdots)}(v_1) = \frac{v_1[r_+(v) - r_-(v)\exp(-\delta(v)t)] - r_+(v)r_-(v)(1 - \exp(-\delta(v)t))}{v_1[1 - \exp(-\delta(v)t)] + r_+(v)\exp(-\delta(v)t) - r_-(v)}. \tag{2.45}$$

In view of the definition of $r_+(v)$ and $r_-(v)$ (see (2.44)), and since $\exp(-\delta(v)t) \leqslant 1$, the terms $r_+(v) - r_-(v)\exp(-\delta(v)t)$ and $r_+(v)\exp(-\delta(v)t) - r_-(v)$ are positive, while $-r_+(v)r_-(v)(1 - \exp(-\delta(v)t))$ and $1 - \exp(-\delta(v)t)$ are non-negative. Therefore, if $v_1 > 0$,

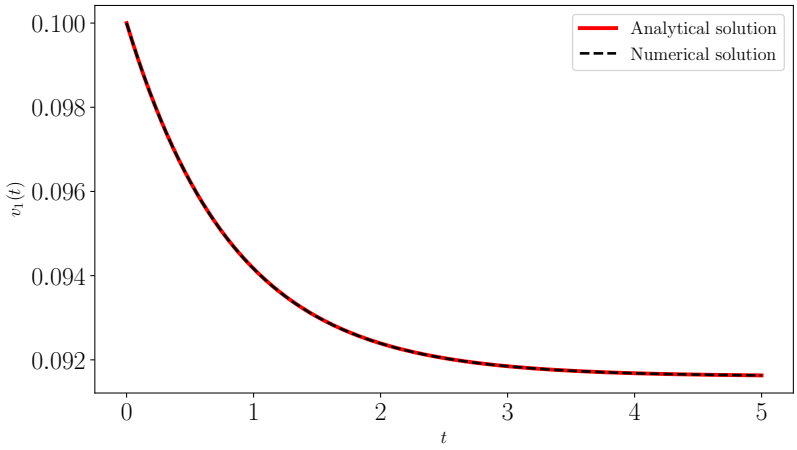



**Fig. 2.3:** Comparison between the analytical solution (2.45) with a numerically integrated solution of (2.35) using a Euler scheme of order 1. The parameters are chosen as $a = 1$, $b = 1$, $c = 0.1$ and $v_1 = 0.1$.

the solution $\varphi_t^{(v_2,\cdots)}(v_1)$ remains positive for all times $t \geqslant 0$.

We next prove the estimates (2.36). Since $Q(v) \leqslant Q_*$, we have in particular $\|A_{\mathrm{L}}^{\alpha} v\| \leqslant R(Q_*)$ in view of Lemma 32. Therefore $c(v) = (A_{\mathrm{L}}^{\alpha} v)_1 \leqslant R(Q_*)$. Similarly, $b(v) = \left(A_{\mathrm{L}}^{\beta} v\right)_1 \leqslant R(Q_*)$. Therefore,

$$r_+(v)^2 \leqslant \frac{1}{2a^2}\Big(b(v)^2 + 2ac(v)\Big) \leqslant \frac{1}{8\beta_1^2}\Big(R(Q_*)^2 + 4\beta_1 R(Q_*)\Big).$$

In view of (2.45), it holds

$$\begin{aligned}\left|\varphi_t^{(v_2,\cdots)}(v_1)\right| &\leqslant \frac{v_1(r_+(v) - r_-(v)) - r_-(v)r_+(v)}{-r_-(v)} \leqslant 2v_1 + r_+(v) \\ &\leqslant 2Q_* + \sqrt{\frac{1}{8\beta_1^2}(R(Q_*)^2 + 4\beta_1 R(Q_*))} := \mathcal{R}(Q_*).\end{aligned}$$

Moreover, in view of (2.35), $d\varphi_t^{(v_2,\cdots)}(v_1)/dt$ is uniformly bounded in time by $2\beta_1\mathcal{R}(Q_*)^2 + R(Q_*)\mathcal{R}(Q_*) + R(Q_*)$. We finally note that

$$\frac{d^2\varphi_t^{(v_2,\cdots)}(v_1)}{dt^2}(t) = -\Big(2a^2\varphi_t^{(v_2,\cdots)}(v_1) + b(v)\Big)\frac{d\varphi_t^{(v_2,\cdots)}(v_1)}{dt},$$

is also uniformly bounded in time since $\varphi_t^{(v_2,\cdots)}(v_1)$ and $d\varphi_t^{(v_2,\cdots)}(v_1)/dt$ are uniformly bounded. This shows that the estimates (2.36) hold true.

□

*Proof of Lemma 34.* Since the subdynamics (2.11) is a Kolmogorov forward equation of a Markov process, the solution remains non-negative. Moreover, since every state is accessible from the state $k$, it holds that, for all $t > 0$ and all $n \geqslant 1$, the solution of the Kolmogorov forward equation satisfies $u_n(t) > 0$ (see [Nor97, Chapter 3.2]). □

## 2.C Proofs for the decay estimates of the solution of (P-Diff)

This section is organized as follows. We first prove Theorem 10 in Appendix 2.C.1, with some technical estimates postponed to Appendix 2.C.2. We finally prove Theorems 13 and 14 in Appendix 2.C.3.

### 2.C.1 Proof of Theorem 10

#### Existence, uniqueness and regularity of the solution

We state some results on the regularity of the solution given that Assumptions 4 and 5 hold true. Let us first recall some important results that allow us to relate partial differential equations (PDEs) with stochastic differential equations (SDEs). Those results are based on the notes by Hairer [Hai11]. Let us consider the stochastic differential equation, written using Stratonovich formulation,

$$dX_t = V_0(X_t)dt + \sum_{j=1}^{m} V_j(X_t) \circ dW_t^j \tag{2.46}$$

where the $V_i$'s are smooth vector fields on $\mathbb{R}^n$ and the $\{W_t^i\}_{t\geqslant 0}$'s are independent standard Brownian motions. We also need to introduce the notion of Lie Bracket $[U, V]$ between to vector fields $U$ and $V$ with values in $\mathbb{R}^n$:

$$[U, V](x) = DV(x)U(x) - DU(x)V(x)$$

where $DU \in \mathbb{R}^{n\times n}$ is the derivative matrix given by $(DU)_{ij} = \partial_j U_i$. The parabolic Hörmander condition then reads as follows.

**Definition 35.** *Let us consider the vector fields*

$$V_1, \cdots, V_m, \qquad [V_i, V_j], 0 \leqslant i, j \leqslant m, \qquad [V_k, [V_i, V_j]], 0 \leqslant i, j, k \leqslant m, \qquad \cdots$$

*Let $\mathcal{V}(x)$ be the vector space spanned by these vector fields for $x$ in $\mathbb{R}^n$. We say that the SDE* (2.46) *satisfies the parabolic Hörmander condition if $\mathcal{V}(x) = \mathbb{R}^n$ for all $x$.*

We then can state Hörmander's theorem:

**Theorem 36.** *[Hairer [Hai11], Theorem 1.3] Consider* (2.46) *and assume that all vector fields have bounded derivatives of all orders. If the SDE* (2.46) *satisfies the parabolic Hörmander condition, then its solutions admit a smooth density with respect to Lebesgue measure and the corresponding Markov semigroup maps bounded functions into smooth functions.*

Let us recall the link between Itô and Stratonovich integrals. The Stratonovich SDE (2.46) has the same solution as the Itô SDE

$$dX_t = \overline{V}_0(X_t)dt + \sum_{j=1}^{m} V_j(X_t)dW_t^j,$$

where, for $i = 1, \cdots, n$

$$\overline{V}_0^i(x) = V_0^i(x) + \frac{1}{2}\sum_{k=1}^{n}\sum_{j=1}^{m} V_j^k(x)\frac{\partial V_j^i}{\partial x_k}(x).$$

We now have the tools for proving the existence and uniqueness of a smooth solution to (P-Diff) under Assumptions 4 and 5. Let us introduce the Stratonovich SDE

$$dX_t = -\frac{1}{2}\sigma(X_t)\sigma'(X_t)\,dt + \sigma(X_t) \circ dW_t,$$

which is equivalent to the Itô SDE

$$dX_t = \sigma(X_t)\,dW_t.$$

Theorem 36 allows us to conclude that its solution admits a smooth density with respect to Lebesgue measure and the corresponding Markov semigroup maps bounded functions into smooth functions. Using the fact that $\mathfrak{C}_0$ is bounded, this ends the proof.

### Proof of the decay estimates (2.24)

We first prove (2.24) for $n = 1$ by introducing the stochastic process associated with (P-Diff) (Steps 1-2). In Step 3 we generalize the previous steps and rely on results proved in Appendix 2.C.2 in order to give a proof for higher order derivatives.

**Step 1: Reformulation as a diffusion with additive noise.** We introduce the stochastic process $(X_t)_{t\geqslant 0}$ defined as

$$dX_t = \sigma(X_t)\,dW_t, \qquad X_0 = x,$$

where $(W_t)_{t\geqslant 0}$ is a standard Brownian motion. Using the Feynman-Kac representation formula [LBL08], the solution of (P-Diff) can be written as

$$u(t,x) = \mathbb{E}[\mathfrak{C}_0(X_t)|X_0 = x] := \mathbb{E}^x[\mathfrak{C}_0(X_t)].$$

Note that the boundedness of $\mathfrak{C}_0$ immediately gives the boundedness of $u$. We next use a Lamperti transform [LP06] on the process $X$. Define

$$\varphi(x) = \int_0^x \frac{1}{\sigma(s)}\,ds. \tag{2.47}$$

The function $\varphi$ is a well defined smooth function since $\sigma$ is a positive smooth function. We then introduce the stochastic process $Y_t = \varphi(X_t)$. Using Itô's formula,

$$dY_t = -\frac{1}{2}\sigma'\Big(\varphi^{-1}(Y_t)\Big)dt + dW_t, \qquad Y_0 = \varphi(x) = y.$$

Defining $v_0 = \mathfrak{C}_0 \circ \varphi^{-1}$, and the function

$$\forall\,(t,y) \in \mathbb{R}_+ \times \mathbb{R}, \qquad v(t,y) := \mathbb{E}^y[v_0(Y_t)], \tag{2.48}$$

the values of $u$ are obtained from $u(t,x) = v(t,\varphi(x))$

**Step 2: Relating the first derivative of $u$ with the flow.** Introduce $\Psi = -\frac{1}{2}\sigma' \circ \varphi^{-1}$, which is a smooth bounded function with bounded derivatives (see Appendix 2.C.2), and the tangent process $\eta$ of $Y$ (*i.e.* the derivative of the flow with respect to the initial

condition [Pro13, Chapter V, Theorem 39]):

$$d\eta_t = \Psi'(Y_t)\eta_t\,dt, \qquad \eta_0 = 1. \tag{2.49}$$

Then (see [Cer01, Chapter 1.3]),

$$\frac{\partial v}{\partial y}(t,y) = \mathbb{E}^y\big[v_0'(Y_t)\eta_t\big].$$

In order to bound the derivative $\partial v/\partial y$, we first notice that $\eta$ is simply the solution of an ODE with a continuous stochastic coefficient:

$$\eta_t = \eta_0 \exp\left[\int_0^t \Psi'(Y_s)\,ds\right].$$

Since $\Psi'$ is bounded, there exists $M_\Psi \in \mathbb{R}_+^*$ such that $0 \leqslant \eta_t \leqslant \eta_0 \exp(M_\Psi t)$ for all $t \geqslant 0$. Moreover, $v_0'$ is bounded in view of Lemma 39 in Appendix 2.C.2. Therefore, there exists $L \geqslant 0$ such that, for all $t \geqslant 0$,

$$\left|\frac{\partial v}{\partial y}(t,y)\right| \leqslant L\exp\left(M_\Psi t\right).$$

The estimate (2.24) for $n = 1$ is finally obtained by noting that

$$\frac{\partial u}{\partial x}(t,x) = \frac{\partial}{\partial x}\big[v(t,\varphi(x))\big] = \frac{\partial v}{\partial y}(t,\varphi(x))\varphi'(x),$$

so that

$$\forall\,(t,x) \in \mathbb{R}_+ \times \mathbb{R}, \qquad \left|\frac{\partial u}{\partial x}(t,x)\right| \leqslant \frac{L\exp\left(M_\Psi t\right)}{\sigma(x)}.$$

The result then follows from Assumption 4.

**Step 3: Generalizing to higher order derivatives.** For the remainder of the proof, let us introduce the tangent process $\eta^{(n)}$ of order $n$ of $Y_t$, recursively defined as the tangent process of $\eta^{(n-1)}$ (see Lemma 41). We also have the following results, proved in Appendix 2.C.2:

1. For all $n \geqslant 1$, there exists a non-negative function $\widetilde{K}_n \in \mathcal{C}^0(\mathbb{R}_+)$ such that,

$$\forall\, t \geqslant 0, \qquad 0 \leqslant \eta_t^{(n)} \leqslant \widetilde{K}_n(t), \tag{2.50}$$

   see Appendix 2.C.2
2. For all $n \geqslant 1$, the derivative of order $n$ of $v_0$ is bounded, see Appendix 2.C.2.
3. For all $n \geqslant 1$, it holds, for $|x| \to +\infty$,

$$\varphi^{(n)}(x) = \mathrm{O}\Big(|x|^{-\frac{\gamma}{2}-n+1}\Big),$$

   see Appendix 2.C.2.

We then use the Faà di Bruno's formula [Joh02] in order to write the higher order derivatives of $v$. Recall that, for $f, g \in \mathcal{C}^\infty(\mathbb{R})$ and $n \geqslant 1$,

$$\frac{d^n}{dx^n}(f(g(x))) = \sum_{k=1}^{n} f^{(k)}(g(x))B_{n,k}\Big(g'(x), g''(x), \ldots, g^{(n-k+1)}(x)\Big), \tag{2.51}$$

where $B_{n,k}$ are the Bell polynomials [Bel27] (see Section 2.C.2). Then, the Faà di Bruno's formula applied to (2.48) together with the results of [Cer01] leads to the following equality: for all $n \geqslant 1$,

$$\forall\, (t,y) \in \mathbb{R}_+ \times \mathbb{R}, \qquad \frac{\partial^n v}{\partial y^n}(t,y) = \mathbb{E}^y\left[\sum_{k=1}^{n} v_0^{(k)}(Y_t) B_{n,k}\Big(\eta_t, \eta_t^{(2)}, \dots, \eta_t^{(n-k+1)}\Big)\right].$$

Using (2.50) and Lemma 39 in Section 2.C.2, there exists, for all $n \geqslant 1$, a non-negative function $M_n \in \mathcal{C}^0(\mathbb{R}_+)$ such that

$$\forall\, n \geqslant 1, \qquad \forall\, (t,y) \in \mathbb{R}_+ \times \mathbb{R}, \qquad \left|\frac{\partial^n v}{\partial y^n}(t,y)\right| \leqslant M_n(t). \tag{2.52}$$

We then use once again the Faa di Bruno's formula to compute the $n$-th partial derivative of $u$:

$$\frac{\partial^n u}{\partial x^n}(t,x) = \sum_{k=1}^{n} \frac{\partial^k v}{\partial y^k}(t,\varphi(x)) B_{n,k}\Big(\varphi'(x), \varphi^{(2)}(x), \dots, \varphi^{(n-k+1)}(x)\Big),$$

so that, with the estimate (2.52), one gets:

$$\left|\frac{\partial^n u}{\partial x^n}(t,x)\right| \leqslant \sum_{k=1}^{n} M_k(t) B_{n,k}\Big(|\varphi'(x)|, \left|\varphi^{(2)}(x)\right|, \dots, \left|\varphi^{(n-k+1)}(x)\right|\Big).$$

Moreover, in view of Lemma 38, it holds

$$\forall\, 1 \leqslant k \leqslant n, \qquad B_{n,k}\Big(|\varphi'(x)|, \left|\varphi^{(2)}(x)\right|, \dots, \left|\varphi^{(n-k+1)}(x)\right|\Big) = \mathrm{O}\Big(|x|^{-n+k-k\gamma/2}\Big).$$

Therefore, we obtain that there exists a non-negative function $K_n \in \mathcal{C}^0(\mathbb{R}_+)$ such that

$$\forall\, (t,x) \in \mathbb{R}_+ \times \mathbb{R}, \qquad \left|\frac{\partial^n u}{\partial x^n}(t,x)\right| \leqslant K_n(t)|x|^{-n\gamma/2}.$$

This concludes the proof of Theorem 10.

## 2.C.2 Some technical results on $\varphi$, $\Psi$, $\eta$ and their derivatives

We gather in this section all the technical results used in the proof of Theorem 10. We will repeatedly use the Faà di Bruno's formula (2.51) and Bell Polynomials.

### Bell polynomials

Bell Polynomials [Bel27] are defined as follows: for any $1 \leqslant k \leqslant n$,

$$B_{n,k}(x_1, x_2, \dots, x_{n-k+1}) = \sum_{(j_1,\cdots,j_{n-k+1}) \in \mathcal{B}_{n,k}} \frac{n!}{j_1! j_2! \cdots j_{n-k+1}!} \left(\frac{x_1}{1!}\right)^{j_1} \cdots \left(\frac{x_{n-k+1}}{(n-k+1)!}\right)^{j_{n-k+1}},$$

where $\mathcal{B}_{n,k}$ is the set of all sequences $(j_1, j_2, \cdots, j_{n-k+1})$ of non-negative integers such that

$$\begin{cases} j_1 + j_2 + \cdots + j_{n-k+1} = k, \\ j_1 + 2j_2 + 3j_3 + \cdots + (n-k+1) j_{n-k+1} = n. \end{cases}$$

### Some estimates on $\varphi$

**Lemma 37.** *The function $\varphi$ defined in* (2.47) *is smooth and its derivatives satisfy*

$$\forall\, n \geqslant 1, \qquad \exists\, K_n \geqslant 0, \qquad \forall\, |x| \geqslant 1, \qquad \left|\varphi^{(n)}(x)\right| \leqslant \frac{K_n}{|x|^{n-1+\gamma/2}}.$$

*Proof.* By definition, since $\sigma$ is a smooth positive function with bounded derivatives, $\varphi$ is also a smooth function and the estimate holds true for $n = 1$. Moreover, in view of Assumption 4, there exists $L_1, L_2 \in \mathbb{R}_+^*$ such that

$$\forall\, |x| \geqslant 1, \qquad L_1|x|^{\gamma/2} \leqslant \sigma(x) \leqslant L_2|x|^{\gamma/2}.$$

Then, using the Faà-di-Bruno's formula, for all $n \geqslant 1$, it holds, since $\varphi'(x) = 1/\sigma(x)$,

$$\varphi^{(n+1)}(x) = \sum_{k=1}^{n} \frac{(-1)^k k!}{\sigma(x)^{k+1}} B_{n,k}\Big(\sigma'(x), \cdots, \sigma^{(n-k+1)}(x)\Big).$$

In view of the definition of Bell polynomials (see Appendix 2.C.2) and Assumption 4, we obtain, for $|x| \geqslant 1$,

$$\begin{aligned}
&\left|B_{n,k}\Big(\sigma'(x), \cdots, \sigma^{(n-k+1)}(x)\Big)\right| \\
&\qquad \leqslant \sum_{(j_1,\cdots,j_{n-k+1})\in\mathcal{B}_{n,k}} \frac{n! S_1^{j_1} \cdots S_{n-k+1}^{j_{n-k+1}}}{j_1! j_2! \cdots j_{n-k+1}!} \left(\frac{|x|^{\gamma/2-1}}{1!}\right)^{j_1} \cdots \left(\frac{|x|^{\gamma/2-n+k-1}}{(n-k+1)!}\right)^{j_{n-k+1}}.
\end{aligned}$$

Noting that $j_1 + \cdots + j_{n-k+1} = k$ and $j_1 + 2j_2 + \cdots + (n-k+1)j_{n-k+1} = n$, we obtain that there is a constant $R_{n,k} \in \mathbb{R}_+$ such that

$$\left|\frac{(-1)^k k!}{\sigma(x)^{k+1}} B_{n,k}\Big(\sigma'(x), \cdots, \sigma^{(n-k+1)}(x)\Big)\right| \leqslant \frac{R_{n,k}}{|x|^{n+k\gamma/2}}.$$

Since $k \geqslant 1$, it finally holds

$$\left|\varphi^{(n+1)}(x)\right| \leqslant \sum_{k=1}^{n} \frac{R_{n,k}}{|x|^{n+k\gamma/2}} = \mathrm{O}\Big(|x|^{-n-\gamma/2}\Big),$$

which concludes the proof. □

**Lemma 38.** *For all $1 \leqslant k \leqslant n$, it holds*

$$B_{n,k}\Big(|\varphi'(x)|, \left|\varphi^{(2)}(x)\right|, \ldots, \left|\varphi^{(n-k+1)}(x)\right|\Big) = \mathrm{O}\Big(|x|^{-n+k-k\gamma/2}\Big).$$

*Proof.* In view of Lemma 37, there exists $\widetilde{S}_{n,k} \in \mathbb{R}_+$ such that

$$B_{n,k}\Big(|\varphi'(x)|, \ldots, \left|\varphi^{(n-k+1)}(x)\right|\Big) \leqslant \widetilde{S}_{n,k} \sum_{(j_1,\cdots,j_{n-k+1})\in\mathcal{B}_{n,k}} \Big(|x|^{-\gamma/2}\Big)^{j_1} \cdots \Big(|x|^{\gamma/2-n+k}\Big)^{j_{n-k+1}}$$

Note that

$$\begin{aligned}
&\Big(|x|^{-\gamma/2}\Big)^{j_1} \Big(|x|^{-\gamma/2-1}\Big)^{j_2} \cdots \Big(|x|^{\gamma/2-n+k}\Big)^{j_{n-k+1}} \\
&\qquad = |x|^{(1-\gamma/2)(j_1+\cdots+j_{n-k+1})} |x|^{-(j_1+2j_2+\cdots+(n-k+1)j_{n-k+1})},
\end{aligned}$$

so that, in view of Appendix 2.C.2,

$$0 \leqslant B_{n,k}\Big(|\varphi'(x)|, \Big|\varphi^{(2)}(x)\Big|, \dots, \Big|\varphi^{(n-k+1)}(x)\Big|\Big) \leqslant \widetilde{S}_{n,k} \sum_{(j_1,\cdots,j_{n-k+1})\in\mathcal{B}_{n,k}} |x|^{-k\gamma/2-n+k},$$

which gives the claimed estimate. □

### On the derivatives of $v_0$

The following technical result holds true in the framework of Theorem 10 in view of Assumption 5.

**Lemma 39.** *The function $v_0$ is smooth and its derivatives are uniformly bounded on $\mathbb{R}$.*

*Proof.* By definition, since $\sigma$ is a positive continuous function, $\varphi$ is an increasing continuous function, and is therefore invertible. Moreover, since $\varphi' = 1/\sigma$ is positive, the inverse function $\varphi^{-1}$ is also differentiable and its first derivative reads $(\varphi^{-1})'(y) = \sigma(\varphi^{-1}(y))$. In fact $\varphi^{-1}$ is smooth and its $n$-th order derivative is $(\varphi^{-1})^{(n)} = R_n\Big(\sigma, \sigma', \cdots, \sigma^{(n-1)}\Big) \circ \varphi^{-1}$, where $R_n$ is a polynomial related to the Bell polynomials. Since $\mathfrak{C}_0$ is smooth, this proves that $v_0 = \mathfrak{C}_0 \circ \varphi^{-1}$ is smooth.

Next, in view of the definition of $v_0$,

$$v_0'(y) = \frac{\mathfrak{C}_0'(\varphi^{-1}(y))}{\varphi'(\varphi^{-1}(y))} = (\mathfrak{C}'\sigma)\Big(\varphi^{-1}(y)\Big).$$

Note that, in view of Assumption 5, $\mathfrak{C}_0'\sigma$ is bounded. Therefore, there exists $R_1 \geqslant 0$ such that $|v_0'(y)| \leqslant R_1$ for all $y \in \mathbb{R}$. A similar argument is used for higher order derivatives. We first prove that the derivative of order $n$ of $v_0$ reads $v_0^{(n)} = P_n \circ \varphi^{-1}$, with

$$P_n = \mathfrak{C}_0^{(n)}\sigma^n + \sum_{k=1}^{n-1} \mathfrak{C}_0^{(k)}\sigma^k \left[\sum_{j=1}^{k} \sum_{(\ell_1^k,\cdots,\ell_j^k)\in\mathcal{L}_j^k} c_{(\ell_1^k,\cdots,\ell_j^k)} \sigma^{\ell_1^k} (\sigma')^{\ell_2^k} \cdots \Big(\sigma^{(j-1)}\Big)^{\ell_j^k}\right], \tag{2.53}$$

where $\mathcal{L}_j^k = \Big\{\Big(\ell_1^k, \cdots, \ell_j^k\Big) \in \mathbb{N}^j \;\Big|\; \ell_1^k + \cdots + \ell_j^k \leqslant \ell_2^k + 2\ell_3^k + \cdots + (j-1)\ell_j^k\Big\}$ and $c_{(\ell_1^k,\cdots,\ell_j^k)}$ are real coefficients. Suppose that (2.53) holds true for some integer $n \geqslant 1$. Since $v_0^{(n+1)} = \sigma P_n' \circ \varphi^{-1}$, it suffices to prove that $P_{n+1} = \sigma P_n'$ is of the form (2.53). It holds

$$\begin{aligned}
\sigma P_n' = {} & \mathfrak{C}_0^{(n+1)}\sigma^{n+1} + n\mathfrak{C}_0^{(n)}\sigma^n\sigma' \\
& + \sum_{k=1}^{n-1} \mathfrak{C}_0^{(k+1)}\sigma^{k+1} \left[\sum_{j=1}^{k} \sum_{(\ell_1^k,\cdots,\ell_j^k)\in\mathcal{L}_j^k} c_{(\ell_1^k,\cdots,\ell_j^k)} \sigma^{\ell_1^k} (\sigma')^{\ell_2^k} \cdots \Big(\sigma^{(j-1)}\Big)^{\ell_j^k}\right] \\
& + \sum_{k=1}^{n-1} k\mathfrak{C}_0^{(k)}\sigma^k \left[\sum_{j=1}^{k} \sum_{(\ell_1^k,\cdots,\ell_j^k)\in\mathcal{L}_j^k} c_{(\ell_1^k,\cdots,\ell_j^k)} \sigma^{\ell_1^k} (\sigma')^{\ell_2^k+1} \cdots \Big(\sigma^{(j-1)}\Big)^{\ell_j^k}\right] \\
& + \sum_{k=1}^{n-1} \mathfrak{C}_0^{(k)}\sigma^k \left[\sum_{j=1}^{k} \sum_{(\ell_1^k,\cdots,\ell_j^k)\in\mathcal{L}_j^k} \ell_1^k c_{(\ell_1^k,\cdots,\ell_j^k)} \sigma^{\ell_1^k} (\sigma')^{\ell_2^k+1} \cdots \Big(\sigma^{(j-1)}\Big)^{\ell_j^k}\right]
\end{aligned}$$

$$+\cdots$$

$$+\sum_{k=1}^{n-1}\mathfrak{C}_0^{(k)}\sigma^k\left[\sum_{j=1}^{k}\sum_{\left(\ell_1^k,\cdots,\ell_j^k\right)\in\mathcal{L}_j^k}\ell_j^k c_{\left(\ell_1^k,\cdots,\ell_j^k\right)}\sigma^{\ell_1^k+1}\left(\sigma'\right)^{\ell_2^k}\cdots\left(\sigma^{(j-1)}\right)^{\ell_j^k-1}\sigma^{(j)}\right],$$

which, by rearranging the terms and noting that $(j-1)\left(\ell_j^k-1\right)+j\left(\ell_{j+1}^k+1\right)=(j-1)\ell_j^k+j\ell_{j+1}^k+1$, reads as (2.53) with $n$ replaced by $n+1$. We next use Assumptions 4 and 5. The terms $\mathfrak{C}_0^{(k)}\sigma^k$ are indeed bounded while the terms $\sigma^{\ell_1^k}\left(\sigma'\right)^{\ell_2^k}\cdots\left(\sigma^{(j-1)}\right)^{\ell_j^k}$ are at most of order $|q|^{\gamma/2\left(\ell_1^k+\cdots+\ell_j^k\right)-\left(\ell_2^k+\cdots+(j-1)\ell_j^k\right)}$ for $|q|\geqslant 1$. This concludes the proof since $0\leqslant\gamma\leqslant 1/2$ and $\ell_1^k+\cdots+\ell_j^k\leqslant\ell_2^k+2\ell_3^k+\cdots+(j-1)\ell_j^k$. □

### The function $\Psi$ is smooth and bounded with bounded derivatives

We prove here the following result.

**Lemma 40.** *The function $\Psi=-\frac{1}{2}\sigma'\circ\varphi^{-1}$ is a smooth and bounded function with bounded derivatives on $\mathbb{R}$.*

*Proof.* By definition, $\Psi$ is smooth, and bounded since $\sigma'$ is bounded. Then, following the proof of Lemma 39 and noting that $\Psi=\mathfrak{S}_0\circ\varphi^{-1}$, with $\mathfrak{S}_0=-\sigma_0'/2$, is similar to $v_0=\mathfrak{C}_0\circ\varphi^{-1}$, the derivative of order $n$ of $\Psi$ reads $\Psi^{(n)}=R_n\circ\varphi^{-1}$, with

$$R_n=\mathfrak{S}_0^{(n)}\sigma^n+\sum_{k=1}^{n-1}\mathfrak{S}_0^{(k)}\sigma^k\left[\sum_{j=1}^{k}\sum_{\left(\ell_1^k,\cdots,\ell_j^k\right)\in\mathcal{L}_j^k}c_{\left(\ell_1^k,\cdots,\ell_j^k\right)}\sigma^{\ell_1^k}\left(\sigma'\right)^{\ell_2^k}\cdots\left(\sigma^{(j-1)}\right)^{\ell_j^k}\right],$$

where $\mathcal{L}_j^k=\left\{\left(\ell_1^k,\cdots,\ell_j^k\right)\in\mathbb{N}^j\,\middle|\,\ell_1^k+\cdots+\ell_j^k\leqslant\ell_2^k+2\ell_3^k+\cdots+(j-1)\ell_j^k\right\}$ and $c_{\left(\ell_1^k,\cdots,\ell_j^k\right)}$ are real coefficients. The conclusion follows by noting that $\mathfrak{S}_0^{(k)}\sigma^k$ is bounded for all $k\geqslant 1$. □

### On the tangent processes of $Y$

The following technical result holds true in the framework of Theorem 10 in view of Lemma 40.

**Lemma 41.** *Let $\eta^{(n)}$ be the tangent process of order $n$ of $Y$, starting from $\eta_0^{(n)}=0$ for $n\geqslant 2$ and $\eta_0^{(1)}=1$. Then, there exists a non-negative function $\widetilde{K}_n\in\mathcal{C}^0(\mathbb{R}_+)$ such that*

$$\forall\, t\geqslant 0,\qquad 0\leqslant\eta_t^{(n)}\leqslant\widetilde{K}_n(t).$$

*Proof.* We show by induction that for all $1\leqslant k\leqslant n$, the process $\eta^{(k)}$ is bounded and solution of

$$d\eta_t^{(k)}=\Psi'(Y_t)\eta_t^{(k)}dt+G_k\left(Y_t,\eta_t^{(1)},\cdots,\eta_t^{(k-1)}\right)dt, \tag{2.54}$$

where $G_1 = 0$ and, for all $2 \leqslant k \leqslant n$,

$$G_k\left(Y_t, \eta_t^{(1)}, \cdots, \eta_t^{(k-1)}\right) = \sum_{p=2}^{k} \Psi^{(p)}(Y_t) \sum_{(\ell_1^p, \cdots, \ell_{k-1}^p) \in \mathcal{J}_{k-1}^p} c_{\ell_1^p, \cdots, \ell_{k-1}^p} \left(\eta_t^{(1)}\right)^{\ell_1^p} \cdots \left(\eta_t^{(k-1)}\right)^{\ell_{k-1}^p}, \tag{2.55}$$

is a bounded function, $\mathcal{J}_k^p = \left\{(\ell_1^p, \cdots, \ell_k^p) \in \mathbb{N}^k \,\middle|\, \ell_1^p + \cdots + \ell_k^p = p\right\}$ and $c_{\ell_1^p, \cdots, \ell_{k-1}^p}$ are real coefficients. The induction basis $k = 1$ holds true with by the definition (2.49) of the tangent process. For the inductive step, let us assume that (2.54) is true for $1 \leqslant k \leqslant n$. Then (see [Pro13]),

$$d\eta_t^{(n+1)} = \Psi'(Y_t)\eta_t^{(n+1)}dt + G_{n+1}\left(Y_t, \eta_t^{(1)}, \cdots, \eta_t^{(n)}\right)dt, \qquad \eta_0^{(n+1)} = 0,$$

with

$$\begin{aligned}
&G_{n+1}\left(Y_t, \eta_t^{(1)}, \cdots, \eta_t^{(n)}\right) \\
&\quad = \Psi^{(2)}(Y_t)\eta_t^{(1)}\eta_t^{(n)} \\
&\quad + \sum_{p=2}^{n} \Psi^{(p+1)}(Y_t)\eta_t^{(1)} \sum_{(\ell_1^p, \cdots, \ell_{n-1}^p) \in \mathcal{J}_{n-1}^p} c_{(\ell_1^p, \cdots, \ell_{n-1}^p)} \left(\eta_t^{(1)}\right)^{\ell_1^p} \cdots \left(\eta_t^{(n-1)}\right)^{\ell_{n-1}^p} \\
&\quad + \sum_{p=2}^{n} \Psi^{(p)}(Y_t) \sum_{(\ell_1^p, \cdots, \ell_{n-1}^p) \in \mathcal{J}_{n-1}^p} \ell_1^p c_{(\ell_1^p, \cdots, \ell_{n-1}^p)} \left(\eta_t^{(1)}\right)^{\ell_1^p - 1} \left(\eta_t^{(2)}\right)^{\ell_2^p + 1} \cdots \left(\eta_t^{(n-1)}\right)^{\ell_{n-1}^p} \\
&\quad + \cdots + \sum_{p=2}^{n} \Psi^{(p)}(Y_t) \sum_{(\ell_1^p, \cdots, \ell_{n-1}^p) \in \mathcal{J}_{n-1}^p} \ell_{n-1}^p c_{(\ell_1^p, \cdots, \ell_{n-1}^p)} \left(\eta_t^{(1)}\right)^{\ell_1^p} \cdots \left(\eta_t^{(n-1)}\right)^{\ell_{n-1}^p - 1} \eta_t^{(n)},
\end{aligned}$$

which, by rearranging the terms, reads as (2.55). Then, in view of (2.55), since $\Psi$ has bounded derivatives and we assumed $\eta^{(k)}$ to be bounded for $1 \leqslant k \leqslant n$, the function $G_{n+1}$ is bounded. Using (2.54) (which is in fact a simple ODE with random coefficients), it holds

$$\eta_t^{(n+1)} = \exp\left(\int_0^t \Psi'(Y_s)\, ds\right) \int_0^t G_{n+1}(Y_s, \eta_s^{(1)}, \cdots, \eta_s^{(n)}) \exp\left(-\int_0^s \Psi'(Y_u)\, du\right) ds.$$

This equality allows to conclude. □

## 2.C.3 Proofs on the relation between Cluster dynamics and its Fokker–Planck approximation

### Proof of Theorem 13

The proof mainly consists in rewriting rigorously what we presented in Section 2.3.2, but with the reverse change of variable. Let us first note that $(G^{-1})' = F \circ G^{-1}$, so that

$$\forall\, (t, x) \in Z_M, \qquad \frac{\partial Q}{\partial x}(t, x) = \frac{F(Q(t, x))}{F(x)}, \qquad \frac{\partial Q}{\partial t}(t, x) = -F(Q(t, x)). \tag{2.56}$$

Then, by the chain rule,

$$\frac{\partial \mathscr{C}}{\partial x}(t, x) = \frac{F(Q(t, x))}{F(x)} \frac{\partial \mathfrak{C}}{\partial q}(t, Q(t, x)),$$

and

$$\frac{\partial^2 \mathscr{C}}{\partial x^2}(t,x) = \frac{F^2(Q(t,x))}{F^2(x)} \frac{\partial^2 \mathfrak{C}}{\partial q^2}(t, Q(t,x)) + \frac{F(Q(t,x))(F'(Q(t,x)) - F'(x))}{F^2(x)} \frac{\partial \mathfrak{C}}{\partial q}(t, Q(t,x)).$$

We also have

$$\begin{aligned}\frac{\partial (F\mathscr{C})}{\partial x}(t,x) &= F(x)\frac{\partial \mathscr{C}}{\partial x}(t,x) + F'(x)\mathscr{C}(t,x) \\ &= F(Q(t,x))\frac{\partial \mathfrak{C}}{\partial q}(t,Q(t,x)) + F'(x)\mathfrak{C}(t,Q(t,x)),\end{aligned}$$

and

$$\begin{aligned}\frac{\partial^2 (D\mathscr{C})}{\partial x^2}(t,x) &= D(x)\frac{\partial^2 \mathscr{C}}{\partial x^2}(t,x) + 2D'(x)\frac{\partial \mathscr{C}}{\partial x}(t,x) + D''(x)\mathscr{C}(t,x) \\ &= D(Q(t,x))\frac{\partial^2 \mathfrak{C}}{\partial q^2}(t,Q(t,x)) \\ &+ \left(D(x)\frac{F^2(Q(t,x))}{F^2(x)} - D(Q(t,x))\right)\frac{\partial^2 \mathfrak{C}}{\partial q^2}(t,Q(t,x)) \\ &+ \left(\frac{D(x)}{F(x)}(F'(Q(t,x)) - F'(x)) + 2D'(x)\right)\frac{F(Q(t,x))}{F(x)}\frac{\partial \mathfrak{C}}{\partial q}(t,Q(t,x)) \\ &+ D''(x)\mathfrak{C}(t,Q(t,x)).\end{aligned}$$

Taking the time derivative of $\mathscr{C}$, we also have

$$\frac{\partial \mathscr{C}}{\partial t}(t,x) = \frac{\partial \mathfrak{C}}{\partial t}(t,Q(t,x)) - F(Q(t,x))\frac{\partial \mathfrak{C}}{\partial q}(t,Q(t,x)).$$

Therefore, it holds

$$\frac{\partial \mathfrak{C}}{\partial t}(t,Q(t,x)) = \frac{\partial \mathscr{C}}{\partial t}(t,x) + \frac{\partial (F\mathscr{C})}{\partial x}(t,x) - F'(x)\mathfrak{C}(t,Q(t,x)),$$

and

$$\begin{aligned}D(Q(t,x))\frac{\partial^2 \mathfrak{C}}{\partial q^2}(t,Q(t,x)) &= \frac{\partial^2 (D\mathscr{C})}{\partial x^2}(t,x) \\ &- \left(D(x)\frac{F^2(Q(t,x))}{F^2(x)} - D(Q(t,x))\right)\frac{\partial^2 \mathfrak{C}}{\partial q^2}(t,Q(t,x)) \\ &- \left(\frac{D(x)}{F(x)}(F'(Q(t,x)) - F'(x)) + 2D'(x)\right)\frac{F(Q(t,x))}{F(x)}\frac{\partial \mathfrak{C}}{\partial q}(t,Q(t,x)) \\ &- D''(x)\mathfrak{C}(t,Q(t,x)).\end{aligned}$$

Combining the last two equations and using (P-Diff) gives us that $\mathscr{C}$ is solution of

$$\frac{\partial \mathscr{C}}{\partial t}(t,x) = -\frac{\partial (F\mathscr{C})}{\partial x}(t,x) + \frac{1}{2}\frac{\partial^2 (D\mathscr{C})}{\partial x^2}(t,x) - R_{\mathfrak{C}}(t,x),$$

where

$$R_{\mathfrak{C}}(t,x) = \frac{1}{2}\left(D(x)\frac{F^2(Q(t,x))}{F^2(x)} - D(Q(t,x))\right)\frac{\partial^2 \mathfrak{C}}{\partial q^2}(t,Q(t,x)) + \left(\frac{1}{2}D''(x) - F'(x)\right)\mathfrak{C}(t,Q(t,x)) + \left(\frac{D(x)}{2F(x)}(F'(Q(t,x)) - F'(x)) + D'(x)\right)\frac{F(Q(t,x))}{F(x)}\frac{\partial \mathfrak{C}}{\partial q}(t,Q(t,x)).$$

Using Lemma 9 and Assumption 3, we have, as $x \to +\infty$ and for $t \leqslant G(x)$,

$$D(x)\frac{F^2(Q(t,x))}{F^2(x)} - D(Q(t,x)) = D(x)\left[\left(\frac{F^2(Q(t,x))}{F^2(x)} - 1\right) - \left(\frac{D(Q(t,x))}{D(x)} - 1\right)\right] = \mathrm{O}\Big(x^{2\gamma-1}\Big),$$

and

$$\left(\frac{D(x)}{F(x)}(F'(Q(t,x)) - F'(x)) + 2D'(x)\right)\frac{F(Q(t,x))}{F(x)} = \mathrm{O}\Big(x^{\gamma-1}\Big),$$

as well as

$$D''(x) - F'(x) = \mathrm{O}\Big(x^{\gamma-1}\Big).$$

Then, in view of Theorem 10, there exists a non-negative function $K \in \mathcal{C}^0(\mathbb{R}_+)$, such that, for all $(t,x) \in Z_M$,

$$|R_{\mathfrak{C}}(t,x)| \leqslant K(t)x^{\gamma-1},$$

which concludes the proof.

### Proof of Theorem 14

Let us first remark that, for all $n \geqslant n_0$, $\widehat{C}_n$ is well defined and smooth by Theorem 10. Then, for all $(t,x) \in Z_M$, in view of Lemma 9 and (2.56), it holds

$$\forall\, (t,n) \in \mathcal{Z}_M, \qquad Q(t,n+1) - Q(t,n) = 1 + R_1(t,n),$$

where there is a non-negative function $K_1 \in \mathcal{C}^0(\mathbb{R}_+)$ such that $|R_1(t,n)| \leqslant K_1(t)n^{\gamma-1}$. Next, using once again a Taylor expansion, for all $(t,n) \in \mathcal{Z}_M$, there exists $\kappa_n \in ]Q(t,n), Q(t,n+1)[$ such that

$$\begin{aligned}\widehat{C}_{n+1}(t) - \widehat{C}_n(t) &= \mathfrak{C}(t,Q(t,n) + 1 + R_1(t,n)) - \mathfrak{C}(t,Q(t,n)) \\ &= (1+R_1(t,n))\frac{\partial \mathfrak{C}}{\partial q}(t,Q(t,n)) + \frac{1}{2}(1+R_1(t,n))^2\frac{\partial^2 \mathfrak{C}}{\partial q^2}(t,Q(t,n)) \\ &\quad + \frac{1}{6}(1+R_1(t,n))^3\frac{\partial^3 \mathfrak{C}}{\partial q^3}(t,\kappa_n).\end{aligned}$$

Using Theorem 10 and Lemma 9, we have

$$\widehat{C}_{n+1}(t) - \widehat{C}_n(t) = \frac{\partial \mathfrak{C}}{\partial q}(t,Q(t,n)) + \frac{1}{2}\frac{\partial^2 \mathfrak{C}}{\partial q^2}(t,Q(t,n)) + R_2(t,n),$$

where there is a non-negative function $K_2 \in \mathcal{C}^0(\mathbb{R}_+)$ such that $|R_2(t,n)| \leqslant K_2(t)n^{-3\gamma/2}$, the dominant terms of the remainder being $R_1 \partial\mathfrak{C}/\partial q$ and $\partial^3\mathfrak{C}/\partial q^3$. Using once again the assumptions on $\alpha$ and $\beta$ (see Assumption 3), in particular that $\alpha' = \mathrm{O}(n^{\gamma-1})$ and the fact that $\mathfrak{C}$ is bounded (see Theorem 10), there is $\xi_n \in [n, n+1]$ such that

$$\begin{aligned}\alpha_{n+1}\widehat{C}_{n+1}(t) - \alpha_n\widehat{C}_n(t) &= (\alpha(n) + \alpha'(\xi_n))\widehat{C}_{n+1}(t) - \alpha_n\widehat{C}_n(t) \\ &= \alpha(n)\frac{\partial\mathfrak{C}}{\partial q}(t, Q(t,n)) + \frac{1}{2}\alpha(n)\frac{\partial^2\mathfrak{C}}{\partial q^2}(t, Q(t,n)) + R_3^\alpha(t,n),\end{aligned}$$

where there is a non-negative function $K_3^\alpha \in \mathcal{C}^0(\mathbb{R}_+)$ such that $|R_3^\alpha(t,n)| \leqslant K_3^\alpha(t)n^{-\gamma/2}$. Similarly,

$$\beta_{n-1}\widehat{C}_{n-1}(t) - \beta_n\widehat{C}_n(t) = -\beta(n)\frac{\partial\mathfrak{C}}{\partial q}(t, Q(t,n)) + \frac{1}{2}\beta(n)\frac{\partial^2\mathfrak{C}}{\partial q^2}(t, Q(t,n)) + R_3^\beta(t,n),$$

where there is a non-negative function $K_3^\beta \in \mathcal{C}^0(\mathbb{R}_+)$ such that $\left|R_3^\beta(t,n)\right| \leqslant K_3^\beta(t)n^{-\gamma/2}$. Combining these results, and noting that $F(n) = F(Q(t,n)) + F(n)R_{F,1}(t,n)$ (where $R_{F,1}$ is defined in Lemma 9) and a similar equality for $D(n)$, we finally obtain

$$\begin{aligned}\beta_{n-1}\widehat{C}_{n-1}C_1 - (\beta_n C_1 + \alpha_n)\widehat{C}_n + \alpha_{n+1}\widehat{C}_{n+1} = &- F(Q(t,n))\frac{\partial\mathfrak{C}}{\partial q}(t, Q(t,n)) \\ &+ \frac{1}{2}D(Q(t,n))\frac{\partial^2\mathfrak{C}}{\partial q^2}(t, Q(t,n)) + R_3(t,n),\end{aligned}$$

where there is a non-negative function $K_3 \in \mathcal{C}^0(\mathbb{R}_+)$ such that $|R_3(t,n)| \leqslant K_3(t)n^{-\gamma/2}$. Since $\mathfrak{C}$ is solution of (P-Diff) and

$$\begin{aligned}\frac{d\widehat{C}_n}{dt}(t) &= \frac{\partial\mathfrak{C}}{\partial t}(t, Q(t,n)) - F(Q(t,n))\frac{\partial\mathfrak{C}}{\partial q}(t, Q(t,n)) \\ &= -F(Q(t,n))\frac{\partial\mathfrak{C}}{\partial q}(t, Q(t,n)) + \frac{1}{2}D(Q(t,n))\frac{\partial^2\mathfrak{C}}{\partial q^2}(t, Q(t,n)),\end{aligned}$$

it follows that $\widehat{C}_n$ satisfies

$$\frac{d\widehat{C}_n}{dt} = \beta_{n-1}\widehat{C}_{n-1}C_1 - (\beta_n C_1 + \alpha_n)\widehat{C}_n + \alpha_{n+1}\widehat{C}_{n+1} - R_3(t,n),$$

from which leads to the desired conclusion.

# A new hybrid deterministic/stochastic coupling approach

# 3

*This chapter is entirely based on the article* "Cluster dynamics modelling of materials: A new hybrid deterministic/stochastic coupling approach" *published in the* Journal of Computational Physics [Ter+17].

## 3.1 Introduction

The microstructural evolution of materials under thermal ageing or irradiation involves complex processes, such as nucleation, growth and coarsening of precipitates or bubbles, that occur on different time scales. The computer simulation of such processes triggered the development of efficient methods able to deal with very different time scales [Vot07; Cat+00; Dom+04; Ort+07; Dup+02; Jou+14; Kir73; Gol+01; Wol+77; Jou+16; Lan74; Gil00; MB11; Dun+16; Sur+04; Ghe+12; AB14]. Long time scale phenomena arise in the evolution of the microstructure of materials under thermal ageing or irradiation. To simulate such events (nucleation, formation of precipitates, growth of bubbles etc.) one needs efficient methods that are able to handle systems with different time scales. Monte Carlo methods, such as kinetic Monte Carlo [Vot07; Cat+00; Dom+04], give physically accurate results but may be limited to short time simulations when frequent events occur.

Mean-field techniques such as rate equation cluster dynamics (RECD) have been used with success to get around this issue [Ort+07; Dup+02; Jou+14]. The modelling of the microstructure is approximated by considering only the defect concentrations, whose evolutions are determined by a system of ordinary differential equations (ODE), called rate equations. Despite its simplicity, two main difficulties occur with RECD. First, since there is one rate equation per cluster type, the number of equations might become very large (clusters might contain up to millions of atoms or defects). In addition, such systems of ODEs are generally stiff, *i.e.* the typical time scale for some reactions is very large while it may be very small for others. Implicit methods are then required and solving such a system of ODE becomes computationally prohibitive as the cluster sizes increase.

Several methods and approximations have been proposed to solve these equations. Deterministic ones include grouping methods where rate equations are gathered into classes [Kir73; Gol+01], and a Fokker-Planck approach where rate equations for large size clusters are approximated by a Fokker-Planck equation [Wol+77]. Recent developments of the Fokker-Planck approach [Jou+14; Jou+16] have proven to be really efficient when only one or two types of defect are considered. They are however strongly limited by the dimensionality of the system. To our knowledge no system with three types of defects/solutes or more have been simulated using a deterministic approach up to large size clusters.

Marian *et al.* recently proposed a stochastic implementation of the rate theory cluster dynamics [MB11]. This method is intended to take into account complex clusters containing different species (point defects, atoms, etc.). The formalism of RECD has been related to purely stochastic approaches such as the well known Stochastic Simulation Algorithm (SSA) introduced by Gillespie [Gil00]. Nevertheless, in stiff systems where

certain reactions occur frequently, the efficiency of the computations is still limited by discrepancies in the time scales.

Attempts have been made to take advantage of both deterministic and stochastic methods. Hybrid deterministic/stochastic algorithms have been proposed [Sur+04; Ghe+12]. Rate equations are used for small size clusters, while large size ones are treated stochastically. In particular, Surh *et al.* [Sur+04] approximate the evolution of large size clusters with a Fokker-Planck equation and use a Langevin dynamics to propagate stochastic particles when clusters reach a certain size. While, to our knowledge, this method is the first one to use such a coupling for cluster dynamics, it has an important limitation. As stochastic particles representing clusters of a certain size are emitted one by one at each time step, the time step should be small to increase the number of particles, and hence reduce the statistical noise. On the contrary, if one wants to rapidly reach large time scales, the time step should be large.

In this work, we propose an alternative way to couple deterministic and stochastic simulations. After a brief presentation of the physical model, we introduce in Section 3.2 a first splitting between the vacancy concentration and the remainder of the distribution. Hence, when the vacancy concentration is fixed, the cluster dynamics becomes linear. Taking advantage of this linearity feature, the dynamics is further decomposed, this time between small and large size clusters. We describe a generic version of the algorithm based on these two splittings. This generic method allows us to design several coupling methods depending on the way the subdynamics are solved. In particular we introduce in Section 3.3 two stochastic methods for computing the evolution of large size clusters, one based on the discrete Markov process associated with the rate equations and the other on a Fokker-Planck approximation. In Section 3.4, we present different ways of computing the vacancy concentration. In Section 3.5 we present a numerical analysis of the algorithm. Numerical results are presented in Section 3.6. We compare in particular some of the methods to compute the vacancy concentration in order to confirm both the validity of our approximation and the overall accuracy of the method.

## 3.2 Model description and main algorithm

We study vacancy clustering during ageing with the model system described in [Ovc+03]. The chosen model is simple but can be enriched with additional sink/source terms, mobile clusters of size two or greater, etc.

### 3.2.1 Rate equations

The RECD approach is used to describe the evolution of cluster size concentration $(C_{\mathrm{vac}}, C_2, C_3, \cdots)$ where $C_{\mathrm{vac}}$ is the vacancy concentration and $C_n$ is the concentration of a cluster of size $n$, *i.e.* containing $n$ vacancies. A set of rate equations governing the time evolution of each concentration is solved. We assume that only mono-vacancies are mobile. Therefore, the rate equation for the concentration $C_n$ of an immobile cluster of size $n \geqslant 2$ is

$$\frac{dC_n}{dt} = \beta_{n-1} C_{n-1} C_{\mathrm{vac}} - \left(\beta_n C_{\mathrm{vac}} + \alpha_n\right) C_n + \alpha_{n+1} C_{n+1}, \tag{3.1}$$

where $\beta_n$ is the absorption rate and $\alpha_n$ the emission rate. These rates take the form [Ovc+03]

$$\begin{cases} \beta_n = \beta_0 n^{1/3}, \quad n \geqslant 1 \\ \alpha_n = \alpha_0 n^{1/3} \exp\left(-\dfrac{E^{\mathrm{b}}_{\mathrm{vac}}(n)}{k_{\mathrm{B}}T}\right), \end{cases}$$

where $\alpha_0 = \beta_0 = (48\pi^2/V_{\mathrm{at}}^2)^{1/3} D_{\mathrm{vac}}$, with $V_{\mathrm{at}}$ the atomic volume and $D_{\mathrm{vac}}$ the diffusion coefficient of vacancies. The term $E^{\mathrm{b}}_{\mathrm{vac}}(n)$ represents the binding energy of a vacancy with a cluster of size $n$:

$$E^{\mathrm{b}}_{\mathrm{vac}}(n) = E^{\mathrm{f}}_{\mathrm{vac}} - \frac{2\gamma V_{\mathrm{at}}}{r(n)},$$

where $\gamma$ is a surface energy, $r$ is the radius of void given by $r(n) = (3nV_{\mathrm{at}}/4\pi)^{1/3}$ and $E^{\mathrm{f}}_{\mathrm{vac}}$ is the vacancy formation energy. The rate equation for $C_{\mathrm{vac}}$ is given by:

$$\frac{dC_{\mathrm{vac}}}{dt} = -2\beta_1 C_{\mathrm{vac}}^2 - \sum_{n\geqslant 2} \beta_n C_n C_{\mathrm{vac}} + \sum_{n\geqslant 2} \alpha_n C_n + \alpha_2 C_2. \tag{3.2}$$

The latter equation is obtained by requiring that cluster dynamics preserve the total quantity of matter $\mathcal{Q}_{\mathrm{tot}}$:

$$\frac{d\mathcal{Q}_{\mathrm{tot}}}{dt} = \frac{d}{dt}\left(C_{\mathrm{vac}} + \sum_{n\geqslant 2} nC_n\right) = 0. \tag{3.3}$$

Initial conditions are given by

$$C_{\mathrm{vac}}(0) = C_{\mathrm{init}} \quad \text{and} \quad C_n(0) = 0, \quad n \geqslant 2, \tag{3.4}$$

where $C_{\mathrm{init}}$ is the quenched-in vacancy concentration.

### 3.2.2 Splitting of the dynamics

Solving the set of rate equations (3.1)–(3.2) when large clusters appear becomes computationally prohibitive. One way to address this problem is to numerically solve the evolution of different classes of clusters with dedicated methods [Jou+14; Sur+04]. We present here a generic algorithm that allows a seamless coupling between such methods.

We propose to first split the dynamics into two elementary dynamics, namely the dynamics of the vacancy concentration $C_{\mathrm{vac}}$ at fixed concentrations $(C_n)_{n\geqslant 2}$ (see Eq. (3.2)) and the dynamics of the cluster concentrations $C = (C_n)_{n\geqslant 2}$ at fixed vacancy concentration $C_{\mathrm{vac}}$ (see Eq. (3.1)). This splitting may be performed after a time $t_0$ corresponding to some initial transient regime where the full set of ODEs (3.1)–(3.2) is integrated by a standard numerical scheme. Let $\delta t$ be a time step and $t_k = t_0 + k\delta t$ for $k \in \{1, \cdots, K\}$ the time at which approximations of the solution are sought, $C^k_{\mathrm{vac}}$ and $C^k$ being respectively approximate solutions of $C_{\mathrm{vac}}(t_k)$ and $C(t_k)$. A good approximation of the cluster dynamics is provided by the following procedure:

$$\begin{cases} C^k = \mathcal{G}_{\delta t}\left(C^{k-1}; C^{k-1}_{\mathrm{vac}}\right), \\ C^k_{\mathrm{vac}} = \mathcal{F}_{\delta t}\left(C^{k-1}_{\mathrm{vac}}; C^k\right), \end{cases} \tag{P1}$$

where $\mathcal{F}_{\delta t}$ and $\mathcal{G}_{\delta t}$ respectively approximate the evolution of (3.2) and (3.1) over a time

step $\delta t$. Figure 3.1 illustrates such a splitting.

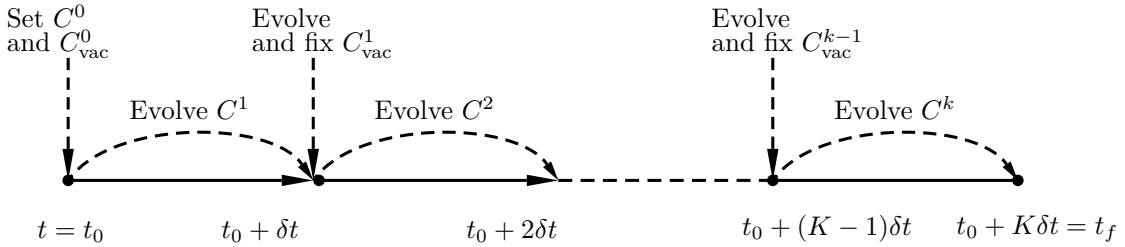



**Fig. 3.1:** Illustration of the splitting between the dynamics of $C_{\text{vac}}$ and $(C_n)_{n\geqslant 2}$.

Notice that in the limit $\delta t \to 0$, the problem (P1) becomes equivalent to the full cluster dynamics (3.1)–(3.2). The error is determined by the time step $\delta t$ and the quality of the approximations $\mathcal{F}_{\delta t}$ and $\mathcal{G}_{\delta t}$. We quantify the error with respect to $\delta t$ in Section 3.6.1, where we also discuss the range of admissible $\delta t$. Since this splitting is a Lie–Trotter splitting, the error on finite time properties scales as $\delta t$ when the sub-ODEs are integrated exactly (by an analysis similar to the one provided in [**HairerLubichWanner06**]), as illustrated in Section 3.5.1. In practice, an additional error arises from the fact that $\mathcal{F}_{\delta t}$ and $\mathcal{G}_{\delta t}$ are not exact, and are typically constructed by substepping strategies; see Section 3.4.1 for $\mathcal{F}_{\delta t}$.

### Integrating the vacancy subdynamics

The numerical scheme $\mathcal{F}_{\delta t}$ in (P1) is obtained by approximating the solution of Eq. (3.2) with $C = (C_n)_{n\geqslant 2}$ fixed. Let us now discuss how to obtain $C^k_{\text{vac}}$ from the knowledge of $C^{k-1}_{\text{vac}}$ and $C^k = (C^k_n)_{n\geqslant 2}$. Introducing

$$\mathcal{A}^k = \sum_{n\geqslant 2} \alpha_n C^k_n + \alpha_2 C^k_2 \tag{3.5}$$

and

$$\mathcal{B}^k = \sum_{n\geqslant 2} \beta_n C^k_n, \tag{3.6}$$

$C^k_{\text{vac}}$ is an approximation of the solution of the following dynamics at time $\delta t$:

$$\frac{dC_{\text{vac}}}{dt} = -\beta_1 C^2_{\text{vac}}(t) - \mathcal{B}^k C_{\text{vac}}(t) + \mathcal{A}^k, \quad C_{\text{vac}}(0) = C^{k-1}_{\text{vac}}. \tag{3.7}$$

The actual numerical method $\mathcal{F}_{\delta t}$ depends on the numerical scheme used to integrate (3.7) (see Section 3.4).

### Integrating the cluster subdynamics

The numerical scheme $\mathcal{G}_{\delta t}$ in (P1) is obtained by approximating the solution of Eq. (3.1) with $C_{\text{vac}}$ fixed. Let us now discuss how to obtain $C^k$ from the knowledge of $C^{k-1}_{\text{vac}}$ and $C^{k-1}$. First notice that, when the vacancy concentration is fixed, the set of rate equations (3.1) forms a linear problem, that can be expressed in matrix form. Denote by

$(e_n)_{n\geqslant 0}$ the basis with components $(e_n)_i = \delta_n(i)$, where $\delta$ is the Kronecker delta. Let $A_0$ be the tridiagonal operator such that:

$$A_0(C_{\text{vac}})e_2 = -(\beta_2 C_{\text{vac}} + \alpha_2)e_2 + \beta_2 C_{\text{vac}} e_3,$$
$$A_0(C_{\text{vac}})e_n = \alpha_n e_{n-1} - (\beta_n C_{\text{vac}} + \alpha_n)e_n + \beta_n C_{\text{vac}} e_{n+1}.$$

The operator $A_0$ can also be represented as the following infinite matrix:

$$A_0(C_{\text{vac}}) = \begin{pmatrix} -(\beta_2 C_{\text{vac}} + \alpha_2) & \alpha_3 & 0 & 0 & \cdots \\ \beta_2 C_{\text{vac}} & -(\beta_3 C_{\text{vac}} + \alpha_3) & \alpha_4 & 0 & \cdots \\ 0 & \beta_3 C_{\text{vac}} & -(\beta_4 C_{\text{vac}} + \alpha_4) & \alpha_5 & \cdots \\ 0 & 0 & \beta_4 C_{\text{vac}} & -(\beta_5 C_{\text{vac}} + \alpha_5) & \ddots \\ \vdots & \vdots & \vdots & \ddots & \ddots \end{pmatrix}$$

The approximation $C^k$ is obtained by numerically integrating the following dynamics:

$$\begin{cases} \dfrac{dC}{dt} = A_0(C_{\text{vac}}^{k-1})C, \\ C(0) = C^{k-1}, \end{cases} \tag{P2}$$

depending on the numerical scheme used to integrate (P2).

As previously noticed, a key feature of Eq. (3.1) is that the dynamics is linear. For any initial condition $C^0$, it is then possible to split the evolution problem (P2) into independent evolutions, corresponding to a decomposition of the initial condition $C^0$. The solution is then obtained by summing the independent sub-solutions. If one writes $C^0 = C^{0,a} + C^{0,b}$, then the solution is $C(t) = C^a(t) + C^b(t)$ with $C^z(t)$ the solution of (P2) with initial condition $C^{0,z}$ for $z = a, b$.

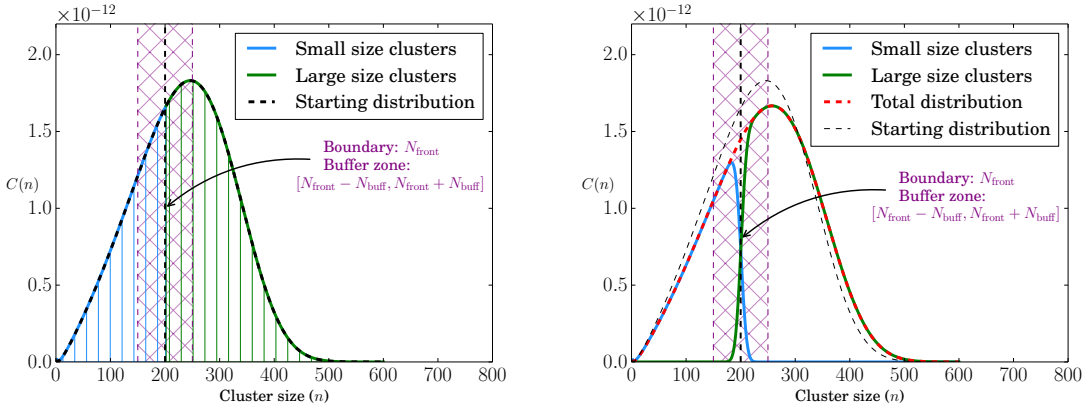
### Splitting and decomposition of the dynamics

Using the linearity of (P2) we choose to separate the evolution of small and large size clusters. With initial conditions $C_{\text{small}}^{k-1}$ and $C_{\text{large}}^{k-1}$ such that $C_{\text{small}}^{k-1} + C_{\text{large}}^{k-1} = C^{k-1}$, the main problem (P1) now writes:

$$\begin{cases} C_{\text{small}}^k = \mathcal{G}_{\delta t}^{\text{small}}\left(C_{\text{small}}^{k-1}; C_{\text{vac}}^{k-1}\right), \\ C_{\text{large}}^k = \mathcal{G}_{\delta t}^{\text{large}}\left(C_{\text{large}}^{k-1}; C_{\text{vac}}^{k-1}\right), \\ C_{\text{vac}}^k = \mathcal{F}_{\delta t}\left(C_{\text{vac}}^{k-1}; C_{\text{small}}^k + C_{\text{large}}^k\right). \end{cases}$$

Note that such a decomposition between small and large size clusters allows us to solve the corresponding dynamics with a different numerical scheme (as emphasized by the notations $\mathcal{G}^{\text{small}}$ and $\mathcal{G}^{\text{large}}$). It is straightforward and computationally effective to numerically solve rate equations for small size clusters (since they consist of a small number of ODEs), so that many options are available for $\mathcal{G}^{\text{small}}$. On the other hand, the treatment of large size clusters requires dedicated techniques. We present in Section 3.3 two stochastic methods

that are highly parallelizable and more appropriate for large size clusters. Figure 3.2 illustrates the decomposition between small and large size clusters on a single time interval. In Figure 3.2.a, the initial distribution is divided into two distributions, one for small size clusters, the other for large size ones. Both distributions are then propagated independently over time and the sum of both propagated distributions (Figure 3.2.b) gives us an approximation of the total distribution.




**(a)**Initial distributions **(b)**Propagated distributions

**Fig. 3.2:** Illustration of the decomposition of the linear dynamics: the initial distribution is divided into two distributions, which are independently propagated. The buffer zone allows the distributions to overlap and limits the calculation cost since both distributions are propagated on a limited space.

### 3.2.3 Main algorithm

From the discussion of Section 3.2.2, the introduction of the coupling algorithm is rather straightforward. Let us introduce a final time $t_{\mathrm{f}}$ for the calculation, a frontier $N_{\mathrm{front}}$ and a buffer zone of size $N_{\mathrm{buff}}$ for the separation and overlapping of small and large size clusters. The size of the buffer zone is chosen sufficiently small to limit the computational cost, as it allows in particular to reduce the number of ODE to solve. Therefore one has to choose $N_{\mathrm{buff}}$ and $\delta t$ such that when one propagates the distribution of small size clusters with $C_n = 0$ for $n \geqslant N_{\mathrm{front}}$ on a time step $\delta t$, it remains negligible for $n \geqslant N_{\mathrm{front}} + N_{\mathrm{buff}}$. Actually, the distribution around $n$ approximately propagates at an average speed of $\beta_n C_{\mathrm{vac}} - \alpha_n$. This property can be observed easily on the Fokker-Planck equation (3.11), presented in Section 3.3.2, where the quantity $F \triangleq \beta C_{\mathrm{vac}} - \alpha$ will act as a drift term. Therefore $N_{\mathrm{buff}}$ can be chosen to be of order $(\beta_{N_{\mathrm{front}}} C_{\mathrm{vac}} - \alpha_{N_{\mathrm{front}}})\delta t$. We also set a maximum cluster size to $N_{\max}$.

As mentioned in Section 3.2.2, we intend to solve the evolution of small and large size clusters with different methods. While the scheme $\mathcal{G}_{\delta t}^{\mathrm{small}}$ can be as simple as a Euler scheme for ODEs, the scheme $\mathcal{G}_{\delta t}^{\mathrm{large}}$ is more elaborated and will be detailed in Section 3.3. The diagram presented in Figure 3.3 summarizes the algorithm presented hereafter.

Let $C^0 = (C_n^0)_{n \geqslant 2}$ be the initial distribution of the cluster concentrations. We denote by $\mathcal{C}_{\mathscr{S}}^0 = (C_2^0, \cdots, C_{\mathrm{N_{front}}-1}^0, 0, \cdots)$ the initial distribution for small size clusters and $\mathscr{C}_{\mathscr{L}}^0 = (0, \cdots, 0, C_{N_{\mathrm{front}}}^0, C_{N_{\mathrm{front}}+1}^0, \cdots)$ the initial distribution for large size clusters and $C_{\mathrm{vac}}^0$ the initial vacancy concentration. To compute the solution from a time $k\delta t$ to a time $(k+1)\delta t$, the general algorithm reads as follows:

0. Decompose the total distribution between small size and large size clusters:

$$\mathcal{C}_{\mathscr{S}}^k = \left(\widetilde{\mathcal{C}}_{\mathscr{S}}^k(2), \cdots, \widetilde{\mathcal{C}}_{\mathscr{S}}^k(N_{\text{front}}-1), 0, \cdots\right),$$
$$\mathscr{C}_{\mathscr{L}}^k = \left(0, \cdots, 0, \widetilde{\mathscr{C}}_{\mathscr{L}}^k(N_{\text{front}}), \widetilde{\mathscr{C}}_{\mathscr{L}}^k(N_{\text{front}}+1), \cdots\right).$$

1. Compute $\widetilde{\mathcal{C}}_{\mathscr{S}}^{k+1}$ on $\{2, \cdots, N_{\text{front}} + N_{\text{buff}}\}$ by integrating the ODE (3.1) with initial condition $\mathcal{C}_{\mathscr{S}}^k$ equal to 0 for $n \geqslant N_{\text{front}}$:

$$\widetilde{\mathcal{C}}_{\mathscr{S}}^{k+1} = \mathcal{G}_{\delta t}^{\text{small}}\left(\mathcal{C}_{\mathscr{S}}^k; C_{\text{vac}}^k\right). \tag{M1}$$

2. Compute $\widetilde{\mathscr{C}}_{\mathscr{L}}^{k+1}$ on $\{N_{\text{front}} - N_{\text{buff}}, \cdots, N_{\max}\}$ by a (possibly approximate) dynamics for large size clusters, with initial condition $\mathscr{C}_{\mathscr{L}}^k$ equal to 0 for $n \leqslant N_{\text{front}} - 1$:

$$\widetilde{\mathscr{C}}_{\mathscr{L}}^{k+1} = \mathcal{G}_{\delta t}^{\text{large}}\left(\mathscr{C}_{\mathscr{L}}^k; C_{\text{vac}}^k\right), \tag{M2}$$

3. Compute the total distribution $C^{k+1}$:

$$C^{k+1} = \widetilde{\mathcal{C}}_{\mathscr{S}}^{k+1} + \widetilde{\mathscr{C}}_{\mathscr{L}}^{k+1},$$

4. Update the vacancy concentration $C_{\text{vac}}^{k+1}$:

$$C_{\text{vac}}^{k+1} = \mathcal{F}_{\delta t}\left(C_{\text{vac}}^k; C^{k+1}\right). \tag{M4}$$

Let us make some general remarks:

- Before the distribution reaches the border $N_{\text{front}}$, equations (3.1) and (3.2) can be solved by an ODE scheme without any splitting between $C_{\text{vac}}$ and $(C_n)_{n\geqslant 2}$. This may be important for the initial transient regime during which $C_{\text{vac}}$ rapidly evolves.
- An alternative way to handle the boundary and the buffer zone is to introduce adaptative time steps $\delta t$ for a fixed value $N_{\text{buff}}$. These steps would be limited by the requirement that $\mathcal{C}_{\mathscr{S}}(N_{\text{front}} + N_{\text{buff}})$ should not grow to a value that is not negligible any more.
- Since Steps (M1) and (M2) are independent, it is possible to switch those steps or to perform them simultaneously.

## 3.3 Discretization of the large size cluster subdynamics

As we want to avoid the computation of a large number of ODEs, we present two methods that allow for a stochastic and parallelizable solving of the evolution of large size clusters (Step (M2) in the algorithm of Section 3.2.3). The first method, called Jump process approach, relies on an original formulation of the set of rate equations when $C_{\text{vac}}$ is fixed, and is better suited for clusters of intermediary sizes. The second method, called Langevin process approach, is based on an approximation by a partial differential equation of the set of rate equations for large size clusters, which is more appropriate for sufficiently large clusters.

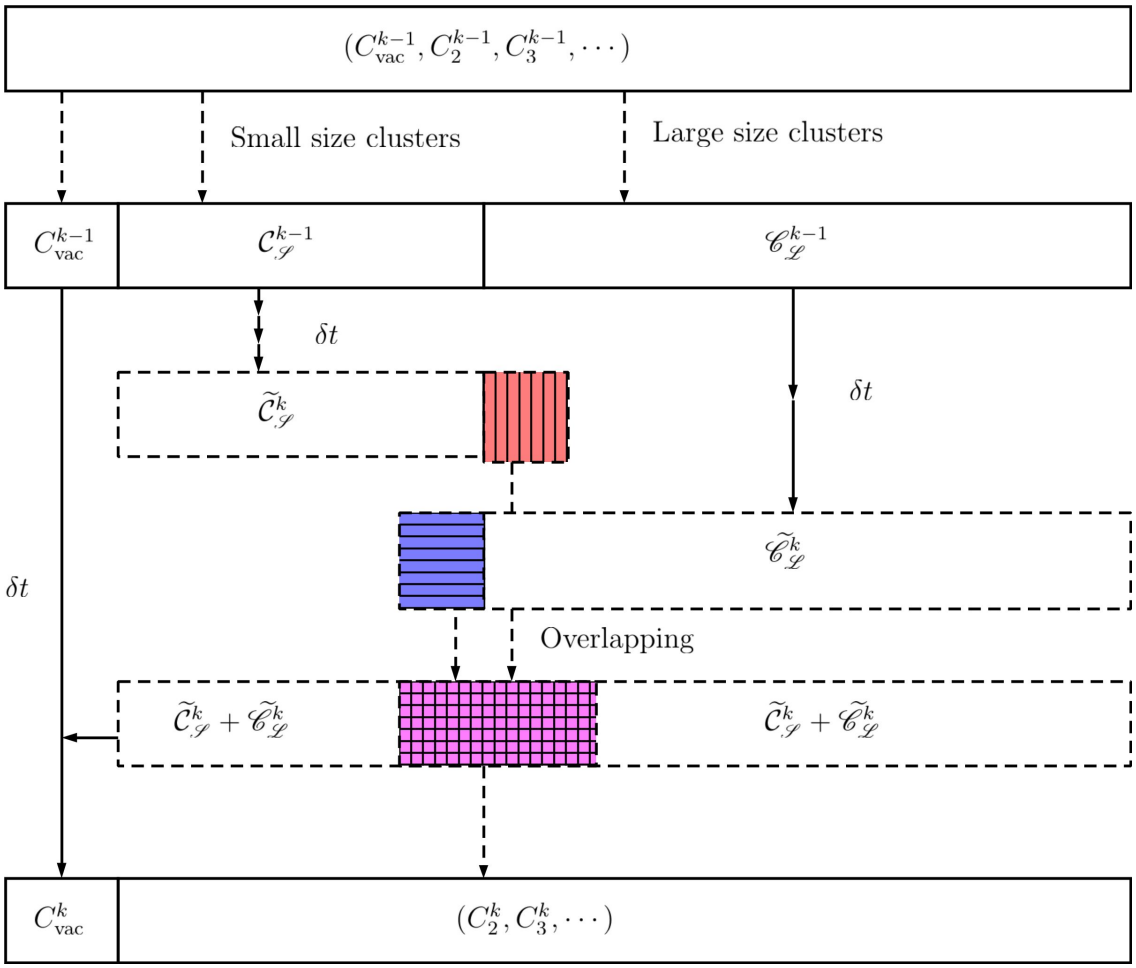



**Fig. 3.3:** Summary of the algorithm presented in Section 3.2.3.

### 3.3.1 The Jump process approach

#### Model presentation

Assuming that the vacancy concentration $C_{\text{vac}}$ is fixed, the rate equations (3.1) are equivalent to a forward-Kolmogorov equation of a Markov process. Such an observation was already made by Goodrich [Goo64]. Indeed a cluster of size $n$ either emits a vacancy and reduces to a cluster of size $n-1$ or absorbs a vacancy and grows to a cluster of size $n+1$. In the size space, such a behaviour can be seen as the jump of a particle in state $n$ to a state $n-1$ or to a state $n+1$. Since the rate of absorption depends on $C_{\text{vac}}$, the Markov process is actually a time-dependent Jump process. It is described as follows: consider a population of size $N(t)$ at a time $t$ and a small time step $\delta\tau$. The transition probabilities for $n \geqslant 2$ are given by

$$\begin{aligned}
&\mathbb{P}\left[N(t+\delta\tau) = n+1 | N(t) = n\right] = \beta_n C_{\text{vac}} \delta\tau + \mathrm{O}(\delta\tau),\\
&\mathbb{P}\left[N(t+\delta\tau) = n-1 | N(t) = n\right] = \alpha_n \delta\tau + \mathrm{O}(\delta\tau),\\
&\mathbb{P}\left[N(t+\delta\tau) = n | N(t) = n\right] = 1 - (\beta_n C_{\text{vac}} + \alpha_n)\delta\tau + \mathrm{O}(\delta\tau).
\end{aligned} \tag{3.8}$$

Let $p(t,n) = \mathbb{P}(N(t) = n)$ be the probability to be in state $n$ at a time $t$. The evolution of $(p(t,n))_{n\geqslant 2}$ is then governed by the following dynamics, a forward-Kolmogorov equation:

$$\frac{dp}{dt}(t,n) = \beta_{n-1}p(t,n-1)C_{\text{vac}} - (\beta_n C_{\text{vac}} + \alpha_n)\, p(t,n) + \alpha_{n+1}p(t,n+1),$$

with initial conditions

$$p(0,n) = 1, \quad \text{and} \quad p(0,i) = 0, \quad i \neq n,$$

for a population initially of size $n$. There exist many algorithms that allow to compute an approximation of $p$ at each time, assuming the concentration $C_{\text{vac}}$ is known. In the following, we implement a simple but rather efficient algorithm in which particles are propagated according to the law of the Jump process (3.8). Each particle represents a cluster of size $n$ and evolves independently on a time step $\delta t$ (during which $C_{\text{vac}}$ is constant). Due to the independence of each particle, the method is highly parallelizable. Moreover, as we only consider large size clusters, we do not suffer from high frequency events due to the small clusters behaviour. This will be later illustrated in Section 3.6.2, where the characteristic jump time appears to be really small for small size clusters.

#### **Jump algorithm: Step** (M2)

The Jump process approach interprets the set of rate equations as a set of forward-Kolmogorov equations and its solution is therefore a probability distribution, denoted by $(p(t,n))_{n\geqslant 2}$ such that $p(t,n) \geqslant 0$ and $\sum_{n=2}^{\infty} p(t,n) = 1$. The total concentration

$$M_{\text{tot}} = \sum_{n=2}^{\infty} \mathscr{C}_{\mathscr{L}}(n) \tag{3.9}$$

should be stored in order to rescale the probability $p$ and get the concentration as $\mathscr{C}_{\mathscr{L}} = M_{\text{tot}}p$.

In order to compute the law $p(t,n)$, starting from a distribution $p_0$, the method we propose generates a large number $N_{\text{part}}$ of particles $(X_n)_{1\leqslant n\leqslant N_{\text{part}}}$ sampled according to $p_0$. There exist various methods to sample from a multinomial distribution ($p_0$ is discrete), see for instance [MI07; Ste94]. These particles are then propagated according to the jump process associated with the transition rates (3.8). For a particle in state $n$, its jump frequency is given by

$$\nu(n) = \beta_n C_{\text{vac}} + \alpha_n \tag{3.10}$$

(*i.e.* the time of the next jump follows an exponential distribution $\mathcal{E}$ of rate $\nu(n)$) and, when it jumps, the particle reaches either the state $n-1$ with probability $\alpha_n/(\beta_n C_{\text{vac}} + \alpha_n)$ or the state $n+1$ with probability $\beta_n C_{\text{vac}}/(\beta_n C_{\text{vac}} + \alpha_n)$. We denote by $\xi(x,\tau,u)$ the function which gives the new state as a function of the previous one $x$ and the two random numbers used in the procedure. Here, $\tau$ is a random time sampled from an exponential distribution of parameter $\nu(x)$ and $u$ is a random number sampled from a uniform distribution $U$ on $[0,1]$ allowing us to choose between the state $n-1$ and $n+1$.

There is in fact no notion of fixed time step in this algorithm apart from the time interval $\delta t$ that corresponds to the final time of the stochastic process. The algorithm summarized as $\mathcal{G}_{\delta t}^{\text{large}}(\mathscr{C}_{\mathscr{L}}^k; C_{\text{vac}}^k)$ in equation (M2) at a step $k$ reads as follows:

1. Sample $N_{\text{part}}$ particles according to the initial distribution $p_0^k(n) = M_{\text{tot}}^{-1}\mathscr{C}_{\mathscr{L}}^k(n)$ as:

$$(x_1^0, \cdots, x_{N\text{part}}^0) \sim p_0^k; \tag{B1}$$

2. Propagate the particles until $(k+1)\delta t$:
   a) Associate the $\ell$-th particle (denoted by $x_\ell$) with a time $\tau_\ell^k$ that is initially set to $k\delta t$;
   b) Propagate independently all particles:
      i. first sample a jump time:

$$\begin{aligned} &\text{Compute the jump frequency: } \nu(x_\ell) = \beta_{x_\ell} C_{\text{vac}}^k + \alpha_{x_\ell}; \\ &\text{Sample a jump time } \delta\tau_\ell \sim \mathcal{E}(\nu(x_\ell)); \\ &\text{Update time } \tau_\ell^k \longleftarrow \tau_\ell^k + \delta\tau_\ell; \end{aligned} \tag{B2.a}$$

      ii. and then propagate until $(k+1)\delta t$:

$$\begin{aligned} &\text{While } \tau_\ell^k \leqslant (k+1)\delta t, \text{ do:} \\ &\text{- Sample: } u \sim U; \\ &\text{- Propagate: } x_\ell \longleftarrow \xi(x_\ell, \delta\tau_\ell, u); \\ &\text{- Compute: } \nu(x_\ell) = \beta_{x_\ell} C_{\text{vac}}^k + \alpha_{x_\ell}; \\ &\text{- Sample the next jump time } \delta\tau_\ell \sim \mathcal{E}(\nu(x_\ell)); \\ &\text{- Update the time as } \tau_\ell^k \longleftarrow \tau_\ell^k + \delta\tau_\ell; \end{aligned} \tag{B2.b}$$

3. Compute the concentration $\widetilde{\mathscr{C}}_{\mathscr{L}}^{k+1}$ at time $(k+1)\delta t$:

$$\widetilde{\mathscr{C}}_{\mathscr{L}}^{k+1}(n) = \frac{M_{\text{tot}}}{N_{\text{part}}} \sum_{\ell=1}^{N_{\text{part}}} \mathbb{1}_n(x_\ell),$$

for $n = N_{\text{front}} - N_{\text{buff}}, \cdots, N_{\max}$, and with $\mathbb{1}_i(j) = 1$ if $i = j$ and 0 otherwise.

Notice that at a step $k > 0$ of the main algorithm, Step (B1) can be performed without actually resampling $N_{\text{part}}$ particles. To this end, from the previous step $k-1$, one just needs to add the particles coming from the part of the distribution of small size clusters that spills out in the large size clusters zone and delete the ones that are added to the small size cluster distribution. In order to keep the number of particles constant, we either randomly suppress some of the particles if there are more than $N_{\text{part}}$ particles, or duplicate some of them if there are less than $N_{\text{part}}$ particles.

Such an algorithm differs from the standard SSA procedure in that all particles are handled independently (which itself comes from the fact that the forward-Kolmogorov equation on $p$ is linear). Parallelizing the scheme is straightforward in the present situation and results in a significant improvement in term of wall-clock time. Moreover this method becomes exact in the limit $N_{\text{part}} \to +\infty$. Nevertheless the frequency $\nu(n)$ increases with $n$ and this might reduce the efficiency of the method for very large clusters. In these situations, Fokker-Planck based methods should be used instead.

### 3.3.2 The Langevin process approach

#### Model presentation

Assuming that the size $n$ of the cluster is large enough and that the concentration $C_{\text{vac}}$ is known at each time, the rate equations can be approximated with a good approximation by a single Fokker-Planck equation [Goo64; Wol+77]:

$$\frac{\partial \mathscr{C}}{\partial t} = -\frac{\partial (F\mathscr{C})}{\partial x} + \frac{1}{2}\frac{\partial^2 (D\mathscr{C})}{\partial x^2}, \tag{3.11}$$

where $F(t,x) = \beta(x)C_{\text{vac}}(t) - \alpha(x)$ and $D(t,x) = \beta(x)C_{\text{vac}}(t) + \alpha(x)$. The size $x$ of the cluster now plays the role of a spatial coordinate. The scalar field $\mathscr{C}$ acts as a concentration which is continuous in space, with:

$$C_n(t) \simeq \mathscr{C}(t,n) \quad \text{for} \quad n \gg 1.$$

When only one type of defect is considered, such a partial differential equation (PDE) is one-dimensional in size space. In this situation, there exist good solvers to efficiently simulate such equations on large scales problems. Jourdan *et al.* [Jou+16] have proposed an efficient method (based on a finite volume formulation) to numerically solve the Fokker-Planck equation when it is coupled to rate equations.

Nevertheless, when two or more types of defects are introduced, *i.e.* when a cluster is identified by a $m$-tuple $(n_1, n_2, \cdots, n_m)$, the Fokker-Planck equation is $m$-dimensional. It then becomes computationally prohibitive to solve it with deterministic mesh-based methods due to the curse of dimensionality (the number of discretization unknowns grows exponentially with the dimension). Stochastic methods are much more appropriate in such situations. The Fokker-Planck equation is related to a stochastic differential equation, called Langevin dynamics. Let $(X_t)_{t\geqslant 0}$ be the stochastic process

$$dX_t = F(t, X_t)dt + \sqrt{D(t, X_t)}dW_t,$$

where $W_t$ is a standard Wiener process. Then the law $p(t,x)$ of $X_t$ satisfies the Fokker-Planck equation (3.11). Therefore, by simulating a large number of trajectories for the process $X_t$, one can obtain a good approximation of the law $p$, *i.e.* the solution of the Fokker-Planck equation (3.11). Since the trajectories are independent of each other, this stochastic method is also highly parallelizable.

#### Langevin process algorithm: Step (M2)

It is important to note that using the stochastic representation of the Fokker-Planck equation (3.11) allows to evolve a probability $p(t,x)$ such that $\int p(t,x)dx = 1$. As in the Jump process approach, the concentration $M_{\mathscr{L}}$ of large size clusters (3.9) should therefore be stored in order to rescale the probability $p$ and obtain the concentration $\mathscr{C}_{\mathscr{L}}$.

Let us introduce an interpolation operator $\mathcal{I}$ that transforms a discrete distribution into a continuous one (such as a linear interpolation of the values at integers). The given initial

condition $\mathscr{C}^0_{\mathscr{L}}$ is associated with an initial density

$$p_0(x) = \frac{\mathcal{I}(\mathscr{C}^0_{\mathscr{L}})(x)}{\int_{N_{\text{front}}}^{\infty} \mathcal{I}(\mathscr{C}^0_{\mathscr{L}})(y)dy},$$

and with a fixed $C_{\text{vac}}$, we formulate the problem as

$$\begin{cases} \text{Find the law } p(t,x) \text{ of } (X_t)_{t\geqslant 0} \text{ solution of} \\ \\ dX_t = F(X_t)dt + \sqrt{D(X_t)}dW_t, \\ \\ p(0,x) = p_0(x), \end{cases}$$

where $F, D$ are defined after (3.11). In order to approximate the law $p(t,x)$ one can generate a large number of trajectories for $X$ and use various methods such as histograms or kernel density estimators to construct an empirical density. Let us therefore introduce the number $N_{\text{part}}$ of Langevin trajectories, $\chi$ a kernel function (*i.e.* a non-negative function that integrates to one and has mean zero) and $h$ a smoothing parameter. We will in particular need to sample the initial distribution $p_0$ with $N_{\text{part}}$ Langevin particles. Here we use a Metropolis algorithm [Met+53; RC13], starting from $N_{\text{front}} + N_{\text{buff}}$, with a uniform proposal distribution of support $[-\alpha, \alpha]$, with $\alpha$ chosen such that the acceptance ratio is around $0.5$. Using the kernel density approach, the law of $X_t$ is approximated by

$$p(t,x) = \frac{1}{N_{\text{part}}h} \sum_{\ell=1}^{N_{\text{part}}} \chi\left(\frac{X_t^\ell - x}{h}\right),$$

where $(X_t^\ell)_{1\leqslant \ell \leqslant N_{\text{part}}}$ are trajectories of the process $X$. Finally to propagate the Langevin dynamics, we use a numerical scheme $\psi(x, \Delta t^{\text{L}}, G)$, here a Euler-Maruyama scheme, with a time step $\Delta t^{\text{L}}$:

$$\psi\left(x, \Delta t^{\text{L}}, G\right) = x + F(x)\Delta t^{\text{L}} + \sqrt{D(x)\Delta t^{\text{L}}}G,$$

where $G$ is a standard Gaussian random variable. The time step $\Delta t^{\text{L}}$ is such that $\delta t = K^{\text{L}}\Delta t^{\text{L}}$ for some $K^{\text{L}} \geqslant 1$.

For the Langevin process approach, the algorithm summarized as $\mathcal{G}^{\text{large}}_{\delta t}(\mathscr{C}^k_{\mathscr{L}}; C^k_{\text{vac}})$ in equation (M2) at a step $k$ writes:

1. Sample $N_{\text{part}}$ particles according to the initial distribution $p_0^k(x) = \frac{\mathcal{I}(\mathscr{C}^k_{\mathscr{L}})(x)}{\int_{N_{\text{front}}}^{\infty} \mathcal{I}(\mathscr{C}^k_{\mathscr{L}})(y)dy}$ as:
$$\left(x_1^0, \cdots, x_{N_{\text{part}}}^0\right) \sim p_0^k; \tag{L1}$$
2. Propagate in time the Langevin particles for $j = 0, \cdots, K^{\text{L}} - 1$:
$$\left(x_1^{j+1}, \cdots, x_{N_{\text{part}}}^{j+1}\right) = \left(\psi\left(x_1^j, \Delta t^{\text{L}}, G_1^j\right), \cdots, \psi\left(x_{N_{\text{part}}}^j, \Delta t^{\text{L}}, G_{N_{\text{part}}}^j\right)\right); \tag{L2}$$
3. Compute the concentration $\widetilde{\mathscr{C}}^{k+1}_{\mathscr{L}}$:
$$\widetilde{\mathscr{C}}^{k+1}_{\mathscr{L}}(n) = \frac{M_{\text{tot}}}{N_{\text{part}}h} \sum_{\ell=1}^{N_{\text{part}}} \chi\left(\frac{x_\ell^{N_L} - n}{h}\right),$$
for $n = N_{\text{front}} - N_{\text{buff}}, \cdots, N_{\max}$.

Once again the stochastic particles are independent and a parallelization of the method is straightforward. Moreover the same remark as in Section 3.3.1 holds for Step (L1), to avoid a full resampling and simply update the population size once the distributions have been separated again into distributions of small and large clusters.

## 3.4 Approximating the dynamics of $C_{\mathrm{vac}}$

The value of $C_{\mathrm{vac}}$ needs to be calculated at multiple values of the time increment $\delta t$ in Step (M4) of the main algorithm. This is encoded in the numerical method $C_{\mathrm{vac}}^k = \mathcal{F}(C_{\mathrm{vac}}^{k-1}; C^k)$. We present here three methods to this end.

### 3.4.1 Decomposition into elementary integrable ODEs [1]

The first method consists in solving the ODE (3.7) with fixed concentrations $(C_n)_{n\geqslant 2}$. A direct integration of the ODE with standard schemes is not appropriate in our case because the values $\mathcal{A}^k$ and $\mathcal{B}^k$ are fluctuating every time step $\delta t$ due to the stochastic evolution of large size clusters. Since the right hand side term of Eq. (3.7) is the difference of two large terms and is observed to be small when integrating the full cluster dynamics (see Figure 3.6), small fluctuations create large instabilities. In order to develop a stable numerical scheme we recommend a decomposition of the evolution into two integrable parts (an affine part and a nonlinear one). For fixed $\mathcal{A}^k$ and $\mathcal{B}^k$ (defined in Eq. (3.5) and (3.6)) we split Eq. (3.7) into the affine part

$$\frac{dC_{\mathrm{vac}}^{\mathrm{L}}}{dt} = -\mathcal{B}^k C_{\mathrm{vac}}^{\mathrm{L}} + \mathcal{A}^k,$$

and the non-linear one

$$\frac{dC_{\mathrm{vac}}^{\mathrm{NL}}}{dt} = -2\beta_1 \left(C_{\mathrm{vac}}^{\mathrm{NL}}\right)^2.$$

Both equations exhibit analytic solutions in closed forms, namely:

$$C_{\mathrm{vac}}^{\mathrm{L}}(t + t_{\mathrm{init}}) = \left(C_{\mathrm{vac}}^{\mathrm{L}}(t_{\mathrm{init}}) - \frac{\mathcal{A}^k}{\mathcal{B}^k}\right) \exp(-\mathcal{B}^k t) + \frac{\mathcal{A}^k}{\mathcal{B}^k},$$

$$C_{\mathrm{vac}}^{\mathrm{NL}}(t + t_{\mathrm{init}}) = \frac{C_{\mathrm{vac}}^{\mathrm{NL}}(t_{\mathrm{init}})}{1 + 2\beta_1 t C_{\mathrm{vac}}^{\mathrm{NL}}(t_{\mathrm{init}})}.$$

To compute the solution one may adopt either a first or second order scheme. To integrate the dynamics with a time step $\Delta t$, such that $\delta t = J\Delta t$, the second order scheme writes

$$\begin{cases} C_{\mathrm{vac}}^{k-1,j+1/2} = \left(C_{\mathrm{vac}}^{k,j} - \dfrac{\mathcal{A}^k}{\mathcal{B}^k}\right) \exp\left(-\dfrac{\mathcal{B}^k \Delta t}{2}\right) + \dfrac{\mathcal{A}^k}{\mathcal{B}^k}, \\ \widetilde{C}_{\mathrm{vac}}^{k-1,j+1} = \dfrac{C_{\mathrm{vac}}^{k,j+1/2}}{1 + \beta_1 \Delta t C_{\mathrm{vac}}^{k,j+1/2}}, \\ C_{\mathrm{vac}}^{k-1,j+1} = \left(\widetilde{C}_{\mathrm{vac}}^{k,j+1} - \dfrac{\mathcal{A}^k}{\mathcal{B}^k}\right) \exp\left(-\dfrac{\mathcal{B}^k \Delta t}{2}\right) + \dfrac{\mathcal{A}^k}{\mathcal{B}^k}. \end{cases} \tag{3.12}$$

1. In view of (2.45) and the expression of an analytical solution to the ODE (3.7) with fixed concentrations $(C_n)_{n\geqslant 2}$, this section might appear superfluous. This is however a prior work and the idea of an analytical solution was not considered at the time.

The value $C_{\text{vac}}^k$ is set to $C_{\text{vac}}^{k-1,J}$ in Step (M4) of the main algorithm.

### 3.4.2 Quasi-stationary limit

The second method consists in making a stronger assumption on the behaviour of $C_{\text{vac}}$, namely that

$$\frac{dC_{\text{vac}}}{dt} \simeq 0.$$

This situation occurs in many physical systems and the quasi-stationnary limit is a good approximation in many cases [Sur+04]. The vacancy concentration is then given by the positive solution of the second order equation

$$-2\beta_1 C_{\text{vac}}^2 - \mathcal{B}^k C_{\text{vac}} + \mathcal{A}^k = 0.$$

The two solutions of this equation are

$$r_\pm = -\frac{\mathcal{B}^k}{4\beta_1} \pm \frac{\sqrt{\left(\mathcal{B}^k\right)^2 + 8\beta_1 \mathcal{A}^k}}{4\beta_1}. \tag{3.13}$$

Since $r_- < 0$, the only physical solution is $r_+ > 0$. The second way of implementing (M4) is therefore given by

$$C_{\text{vac}}^{k+1} = \frac{2\mathcal{A}^k}{\mathcal{B}^k}\left(1 + \sqrt{1 + \frac{8\beta_1 \mathcal{A}^k}{\left(\mathcal{B}^k\right)^2}}\right)^{-1}, \tag{3.14}$$

which is equal to $r_+$ in Eq. (3.13), though numerically more stable when $\mathcal{A}$ and $\mathcal{B}$ are large.

### 3.4.3 Mass conservation

The third method is based on the preservation of the total quantity of matter (see Eq. (3.3)):

$$C_{\text{vac}} + \sum_{n \geqslant 2} n C_n(t) = C_{\text{init}}.$$

The computation of $C_{\text{vac}}$ given the concentrations $(C_n)_{n\geqslant 2}$ is then straightforward:

$$C_{\text{vac}}^k = C_{\text{init}} - \sum_{n \geqslant 2} n C_n^k. \tag{3.15}$$

## 3.5 Numerical analysis

We give in this section error estimates for the schemes we consider. We focus our analysis on two sources of errors: systematic errors due to the splitting of the dynamics (see Section 3.5.1), as well as sampling errors related to the introduction of stochastic algorithms (see Section 3.5.2). We conclude the section by a discussion on the total error which results from these two sources (see Section 3.5.3).

### 3.5.1 Deterministic splitting

Since the splitting introduced in Section 3.2.2 is a Lie-Trotter splitting, we expect to observe an error of order $\delta t$ for properties computed over finite time intervals (which is a standard result in the numerical analysis of ODEs). In order to illustrate such error

estimates, we run simulations with various time steps $\delta t$, in the case when the elementary sub-ODEs in (P1) are both integrated using discretization techniques for ODEs in order not to pollute error estimates with sampling errors. The elementary sub-ODEs are integrated in time using standard second order numerical scheme (Heun scheme) when necessary, and with very small time steps $\Delta t$, in order to limit errors due to the underlying numerical scheme and retain only the error coming from the splitting algorithm. The solutions for $C_{\text{vac}}$ are obtained by using the first method presented in Section 3.4.1, as the error of the integration scheme scales as $\Delta t^2$. A reference solution is obtained by computing a solution $C^{\text{ref}}$ to the full ODE, with the same small timestep $\Delta t$. The errors between the numerical solutions $C^{\text{splitting}}$ of problem (P1) and the reference solution $C^{\text{ref}}$ are measured with the following mean square error norm:

$$\eta_2^{\text{splitting}}(t) = \sqrt{\sum_{n=1}^{N_{\max}} \left( C_n^{\text{splitting}}(t) - C_n^{\text{ref}}(t) \right)^2}. \tag{3.16}$$

The simulations are performed using the setting described more precisely in Section 3.6, for a time $t_{\text{f}} = 1000$ s and with $\Delta t = 10^{-5}$ s. In Figure 3.4, we compare the error at the final time for various values of $\delta t$, ranging from $10^0$ s to $10^2$ s. As expected, the error $\eta_2$ decreases linearly with the splitting time step $\delta t$.

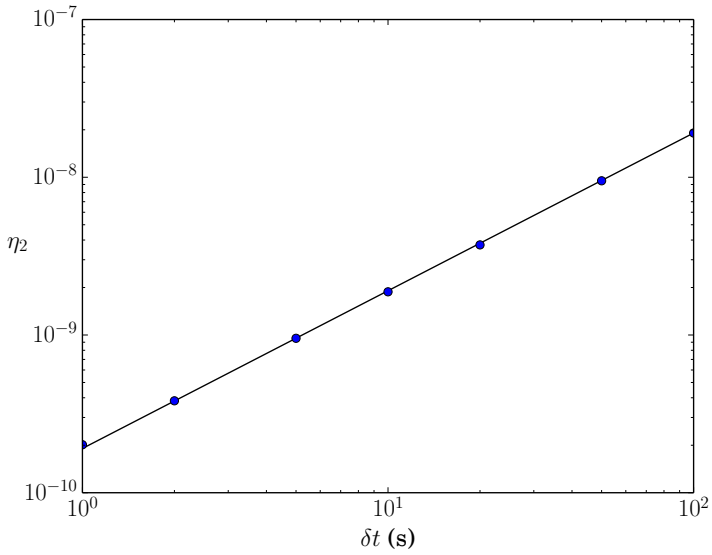



**Fig. 3.4:** Error $\eta_2$ at final time $t_{\text{f}} = 1000$ s as a function of $\delta t$. A linear approximation is superimposed in a black line.

### 3.5.2 Stochastic error

Another source of error in the algorithm is the sampling error arising from the finiteness of the number of particles $N_{\text{part}}$ used to approximate large cluster size dynamics. Since these particles evolve independently, a central limit theorem can be applied, showing that the statistical error scales as $N_{\text{part}}^{-1/2}$. We illustrate such a behaviour by running simulations of a simple system where all clusters are handled stochastically using the jump process (in order to avoid the error coming from the Fokker-Planck approximation), except for the single vacancies, which represent a very small part of the total mass. To measure the error, we consider the average size of clusters, which corresponds to the following observable $\mathcal{M}$

(using the notation of Section 3.3.1)

$$\mathcal{M}(C) = \frac{\sum_{n\geqslant 1} nC_n}{\sum_{n\geqslant 1} C_n} = \frac{\sum_{n\geqslant 1} n\left(\sum_{\ell=1}^{N_{\rm part}} \mathbb{1}_n(x_\ell)\right)}{\sum_{n\geqslant 1}\left(\sum_{\ell=1}^{N_{\rm part}} \mathbb{1}_n(x_\ell)\right)},$$

For each value of $N_{\rm part}$, we perform $N_{\rm sim}$ independent simulations, denoted by $(C^j)_{1\leqslant j\leqslant N_{\rm sim}}$, in order to quantify the variability of the outcome. At a given final time $t_{\rm f}$, we expect that, when $N_{\rm part}$ is sufficiently large for a central limit theorem to hold, the empirical variance of the outcome scales as $N_{\rm part}^{-1/2}$. More precisely, denoting by $C^{t_{\rm f},j}$ the concentrations at time $t_{\rm f}$ for the $j$-th realization, this empirical variance reads

$$\sigma^2_{\rm sim} = \frac{1}{N_{\rm sim}} \sum_{j=1}^{N_{\rm sim}} \left(\mathcal{M}(C^{t_{\rm f},j}) - \overline{\mathcal{M}}_{t_{\rm f},N_{\rm sim}}\right)^2 \simeq \frac{\sigma^2(t_{\rm f})}{N_{\rm part}}, \tag{3.17}$$

with

$$\overline{\mathcal{M}}_{t_{\rm f},N_{\rm sim}} = \frac{1}{N_{\rm sim}} \sum_{n=1}^{N_{\rm sim}} \mathcal{M}(C^{t_{\rm f},j})$$

the average value of the observable $\mathcal{M}$ over the realizations. There is however no prediction for the asymptotic variance $\sigma^2(t_{\rm f})$, which has to be estimated numerically. In fact, the outcomes $\mathcal{M}(C^{t_{\rm f},j})$ should be distributed according to a Gaussian distribution of variance $\sigma^2(t_{\rm f})/N_{\rm part}$.

The scaling (3.17) is illustrated in Figure 3.5.a, for $N_{\rm sim} = 100$ and $N_{\rm part}$ ranging from $10^3$ to $2\times 10^5$, with final simulation time $t_{\rm f} = 10^3$ s. In Figure 3.5.b, we check the fact that the outcomes are indeed distributed according to a Gaussian distribution, in the case when $N_{\rm part} = 2\times 10^3$ and $N_{\rm sim} = 10^4$.

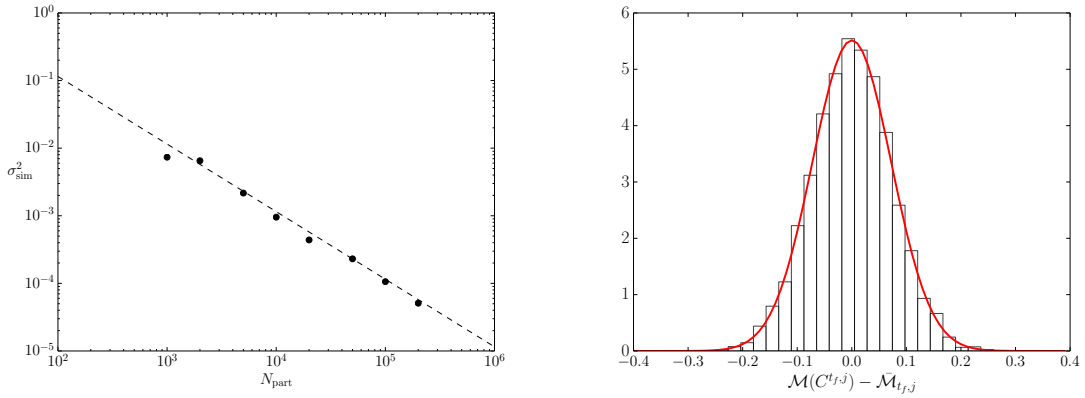



**Fig. 3.5:** Left: Variance (3.17) as a function of $N_{\rm part}$. A linear approximation is superimposed in dashed line. Right: Histogram of the outcomes $\mathcal{M}(C^{t_{\rm f},j}) - \overline{\mathcal{M}}_{t_{\rm f},j}$. The reference Gaussian distribution with variance $\mathcal{N}(0,\sigma^2_{\rm sim})$ is superimposed to the data.

### 3.5.3 Coupling of errors

We identified in the previous sections two important sources of errors: systematic errors arising from the splitting and sampling errors arising from a discretization of the stochastic

representation of the solution. These two sources of errors are the only ones when the integration error in the ODE describing the small size clusters is negligible and jump processes are used for the stochastic parts. The total error at time $t$ on some observable of interest then writes

$$\eta_2^{\mathrm{tot}}(t) \simeq \varepsilon_1(t)\delta t + \frac{\varepsilon_2(t)}{\sqrt{N_{\mathrm{part}}}},$$

where $\varepsilon_1(t)$ and $\varepsilon_2(t)$ depend on the simulation time $t$ (as the notation suggests) and only on the chosen observable. In all our simulations, with splitting time steps no larger than $\delta t = 10$ s and various observables such as the error $\eta_2(t)$ at a given time, see Eq. (3.16), or the total mass, we observe that $\varepsilon_1(t)\delta t \ll \eta_2^{\mathrm{tot}}(t)$ for $N_{\mathrm{part}}$ ranging from $10^3$ to $2 \times 10^7$. The statistical error therefore appears to be the main source of error.

## 3.6 Results

All the simulations reported below are performed using the parameters of [Ovc+03], which are summarized in Table 3.1. The final computation time $t_{\mathrm{f}}$ as well as the time

| | |
|---|---|
| Temperature ($T$) | 823 K |
| Atomic volume ($V_{\mathrm{at}}$) | $1.205 \times 10^{-29}$ m$^3$ |
| Vacancy formation energy ($E^{\mathrm{f}}_{\mathrm{vac}}$) | 1.7 eV |
| Vacancy migration energy ($E^{\mathrm{m}}_{\mathrm{vac}}$) | 1.1 eV |
| Vacancy diffusion coefficient ($D_{\mathrm{vac}}$) | $10^{-6}\exp(-E^{\mathrm{m}}_{\mathrm{vac}}/(k_B T))$ m$^2$s$^{-1}$ |
| Surface energy ($\gamma$) | 1.0 J/m$^2$ |
| Concentration of quenched-in vacancies ($C_{\mathrm{init}}$) | $10^{-7}$ atom$^{-1}$ |
| Kernel function ($\chi$) | $(2\pi)^{-1/2}\exp(-x^2/2)$ |
| Smoothing parameter ($h$) | 0.3 |

**Tab. 3.1:** Parameters for the simulation of a nickel-like metal

interval $\delta t$ and the time steps $\Delta t^{\mathrm{M}}$, $\Delta t^{\mathrm{L}}$ (respectively used for the deterministic integration of Eq. (3.1) and the integration of Langevin dynamics in Step (L2)) are specified for each result. Besides the numerical analysis conducted in Section 3.5, there is a systematic error due to the Fokker-Planck approximation. This error is expected to be of order $N_{\mathrm{front}}^{-2/3}$. A more thorough study of the Fokker–Planck approximation is given in Chapter 2 since the estimation of this error requires to first provide a mathematically rigorous derivation of the Fokker–Planck limit. Nevertheless, the simulations we performed (not reported here) show that, with $N_{\mathrm{front}} = 200$, the error due to the Fokker-Planck approximation is negligible compared to the statistical error as long as the number of particles $N_{\mathrm{part}}$ does not exceed $10^8$. The buffer zone is of size $2N_{\mathrm{buffer}} = 100$ and therefore extends from $n = 150$ to $n = 250$.

### 3.6.1 On the quasi-stationary assumption

We first show that under the assumption that $C_{\mathrm{vac}}$ reaches a quasi-stationary state, the problem (P1) is an excellent approximation of the original RECD problem (3.1)–(3.2) even for large time intervals $\delta t$. Let us introduce a time-dependent function $\mathcal{T}$ defined by

$$\mathcal{T}(t) = \left| \frac{1}{C_{\mathrm{vac}}(t)} \frac{dC_{\mathrm{vac}}}{dt} \right|^{-1}. \tag{3.18}$$

The function $\mathcal{T}$ acts as a characteristic time for the vacancy density evolution. Since

$$C_{\text{vac}}(t+\delta t) = C_{\text{vac}}(t) + \frac{dC_{\text{vac}}}{dt}(t)\delta t + \mathrm{O}(\delta t^2),$$

on a time step $\delta\tau \ll \mathcal{T}$, the variation of $C_{\text{vac}}$ is relatively small. The condition $\delta t \ll \mathcal{T}$ indicates the relevant orders of magnitude of $\delta t$.

We compute a reference solution by solving the full ODE problem (3.1)–(3.2) without any approximation other than the integration scheme. We use a second order Euler-Heun numerical scheme. Starting from the initial condition (3.4), the system first goes through a nucleation stage (for which we use a small time step $\Delta t^{\text{M}} = 10^{-5}$ s), before it enters a growth regime where we start to observe the quasi-stationary state of $C_{\text{vac}}$ (characterized by $\frac{dC_{\text{vac}}}{dt} \simeq 0$). This state is rapidly observed and is reached before clusters grow beyond $N_{\text{front}}$. After the nucleation stage as the system enters the growth regime, we use the time step $\Delta t^{\text{M}} = 10^{-3}$ s. This time step ensures a good conservation of the total quantity of matter $\mathcal{Q}_{\text{tot}} = C_{\text{vac}} + \sum_{n=2}^{N_{\max}} nC(n)$, with a relative error of order $10^{-7}$. The final time of computation $t_{\text{f}}$ is set to $10^4$ s. We denote by $C^{\text{ref}}$ the numerical solution of the dynamics (3.1)–(3.2) obtained by this procedure. The solution $C^{\text{ref}}$ allows to compute the characteristic time $\mathcal{T}$ defined in Eq. (3.18). We observe in Figure 3.6 that $C_{\text{vac}}$ strongly decreases at first from its initial value $C_{\text{init}} = 10^{-7}$ and then enters a quasi-stationary state where the concentration of vacancy slowly decreases. At time $t = 10^3$ s, the characteristic time $\mathcal{T}(t)$ is approximately equal to $4 \times 10^3$ s. This indicates that a time step $\delta t$ of order $10^2$ s is sufficiently small in order to keep the variations of $C_{\text{vac}}$ small.

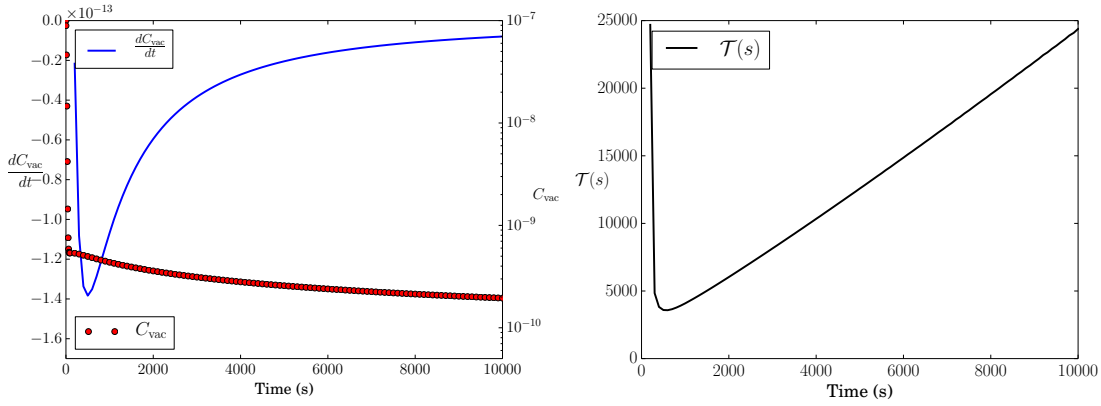



**Fig. 3.6:** Left: $C_{\text{vac}}$ and its derivative as functions of time. Right: characteristic time $\mathcal{T}$ as a function of time.

We next compare the three methods of computing $C_{\text{vac}}$, using a Euler-Heun integration of problem (P1) to update the cluster concentrations (still with $\Delta t^M = 10^{-3}$ s). The initial condition is set to $C^{\text{ref}}(t_0)$ and $t_0 = 10^3$ s ensures that the initial transient regime is over. We then compute the solution obtained when updating the vacancy concentration every time step $\delta t = 10$ s using one of the three methods discussed in Section 3.4, until $t_{\text{f}} = 10^4$ s. For the method presented in Section 3.4.1, we use $\Delta t = 10^{-3}$ s. We then estimate the mean square error between the full ODE solution $C^{\text{ref}}$ and the solutions $C^{\text{splitting}}$ of problem (P1) as defined by (3.16). Figure 3.7.a compares the errors for each of the three ways of integrating the dynamics of $C_{\text{vac}}$ for $C^{\text{splitting}}$. It shows that there is no significant differences between the three methods, as the error remains more than 5

orders of magnitude lower than the total quantity of matter.

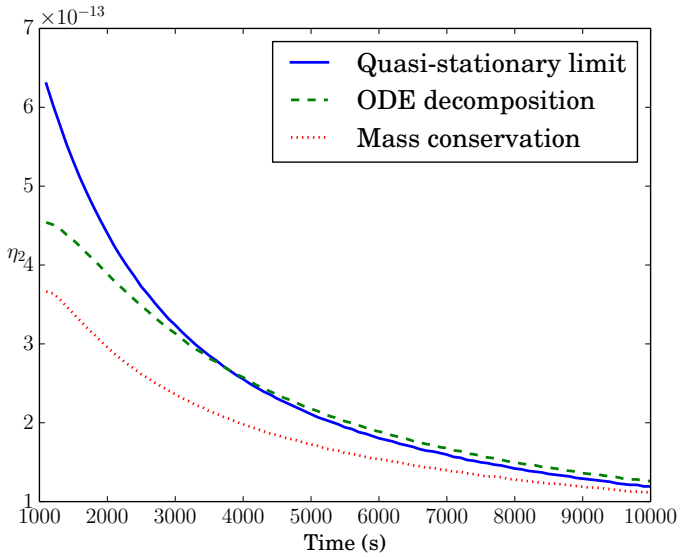



**Fig. 3.7:** Behaviour of the mean square error (3.16) as a function of $\delta t$ (Quasi-stationnary limit: Eq. (3.14); ODE decomposition: Eq. (3.12); Mass conservation: Eq. (3.15)).

In the sequel, we choose the quasi-stationary limit approximation (3.14) as it is stable and straightforward to compute. The conclusion could be different when other types of mobile clusters are taken into account (typically small clusters such as $C_2, \cdots, C_{10}$). In this case the mass conservation method cannot be used. The decomposition into elementary integrable ODEs then becomes the best alternative since the quasi-stationary limit approach requires solving a system of coupled non-linear equations.

### 3.6.2 Accuracy of the splitting algorithm for thermal ageing

We now present simulation results obtained with the main algorithm presented in Section 3.2.3, using the methods presented in Section 3.3.1 and 3.3.2 for the large cluster dynamics. In Steps (B2.a)–(B2.b) and (L2) of the large size cluster dynamics, each particle is propagated independently, which allows dispatching the computations on a parallel architecture. The computations reported here were performed on a cluster of 15 hyper-threaded cores, each thread being used to propagate $N_{\text{proc}} = 2.10^5$ particles, which gives us a total of $N_{\text{part}} = 6.10^6$ particles. The final time of computation $t_{\text{f}}$ is set to $10^5$ s.

The time step used in Langevin process simulations is set to $\Delta t^{\text{L}} = 1$ s while the concentration of vacancies $C_{\text{vac}}$ is updated [2] at times that are multiple of $\delta t = 10$ s. The value of $C_{\text{vac}}$ is calculated using the quasi-stationary limit approach (3.14). For the Jump approach, the time step is not fixed but a characteristic jump time is given by the particles of size $N_{\text{front}}$ and is on the order $(\beta_{N_{\text{front}}} C_{\text{vac}} + \alpha_{N_{\text{front}}})^{-1}$ which is on the order of 10 s when the system is in a growth regime (see Figure 3.8). Moreover it is increasing with time. In contrast, with a fully stochastic approach (without the decomposition between small and large size

2. We could have chosen larger time steps and time intervals in order to speed up the computational time. However we refrained from optimizing the parameters and comparing with state of the art methods since our main objective is to use our method in more complex problems than the ones which can be currently solved with classical methods.

clusters), the time steps of the most frequent events are of order $(\beta_2 C_{\text{vac}} + \alpha_2)^{-1} \simeq 10^{-2}$ s.

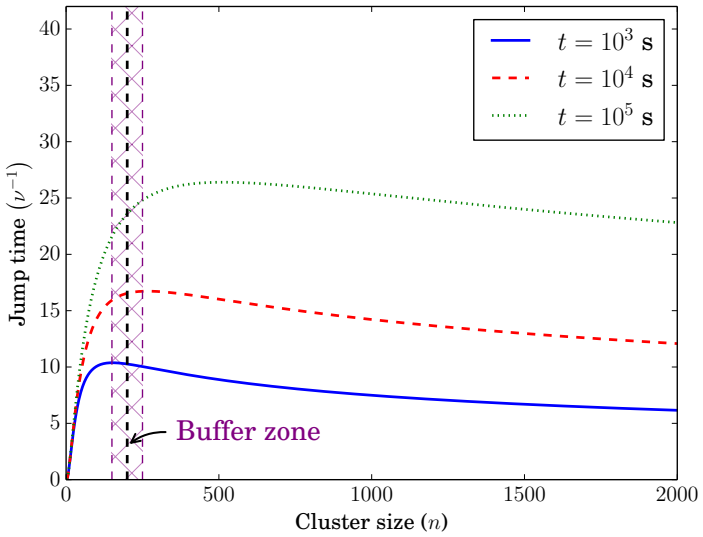



**Fig. 3.8:** Characteristic jump time $\nu^{-1}$ (see Eq. (3.10)) as a function of the cluster size $n$: the most frequent events occur for small $n$.

Aside from the stochastic fluctuations inherent to both methods, the results presented in Figure 3.9 show a perfect agreement with the exact concentration obtained by an integration of the full ODE system. The total concentration is equal to $\mathcal{Q}_{\text{tot}}^{\text{LP}} = 9.989 \times 10^{-8}$ at the final time in the case of the LP approach and is equal to $\mathcal{Q}_{\text{tot}}^{\text{JP}} = 9.968 \times 10^{-8}$ for the JP approach. The relative error on the total concentration is therefore less than $0.4\%$.

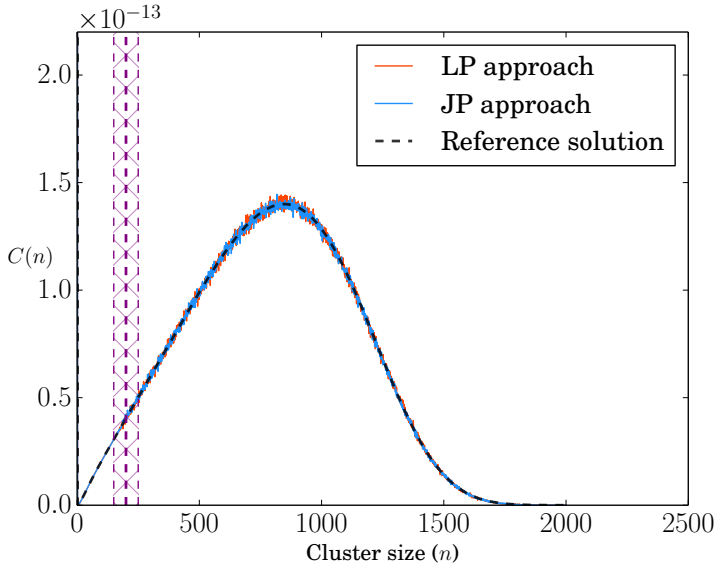



**Fig. 3.9:** Comparison between the exact concentration (black dotted line), the concentration obtained with a Langevin process (LP) approach (red line) and the one obtained with the Jump process (JP) approach (blue line).

### 3.6.3 Parallelization

We finally investigate the gain in efficiency arising from the parallelization of the stochastic part, a crucial advantage of the proposed method. We assess relative efficiencies by computing and comparing the wall-clock time of our simulations depending on the number of processors we used to run the simulations. We report in Figure 3.10 the results of simulations performed with fixed number of particles ($N_{\text{part}} = 256 \times 10^4$ or $N_{\text{part}} = 256 \times 10^5$) and various numbers of processors $N_{\text{proc}}$ ranging from $4$ to $1024$. We observe

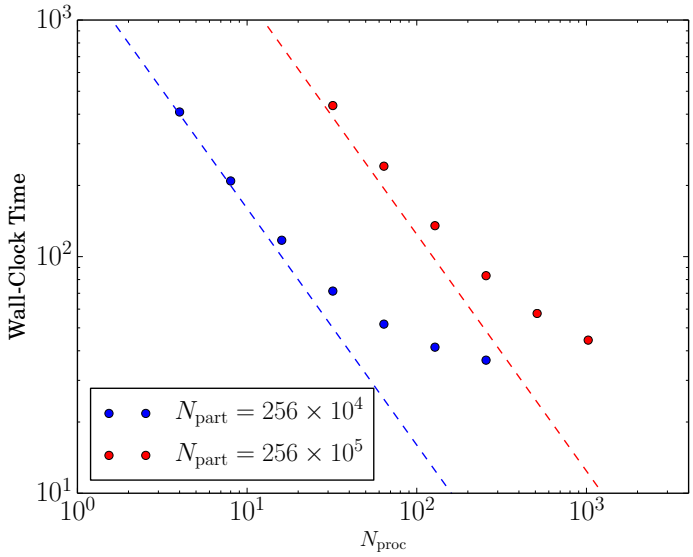



**Fig. 3.10:** Comparison between the wall-clock time of computations for a fixed number of particles $N_{\text{part}}$. The ideal linear scaling behaviour is plotted in dashed lines.

that the increase of the number of processors at first reduces linearly the wall-clock computation time, but then, as the number of particles per processor decreases, we observe that the slope tends to reach an asymptotic limit, determined by computational time associated with the deterministic part. Moreover we notice that, as expected, this asymptotic limit appears for smaller values of $N_{\text{proc}}$ when the total number of stochastic particles decreases.

## 3.7 Conclusion

In this work we have presented a generic coupling algorithm allowing to simulate thermal ageing and ageing under irradiation using cluster dynamics. Our approach consists in coupling the standard rate equations for small size clusters with more efficient stochastic methods for large size clusters. Such a coupling is based on a splitting of the dynamics between the nonlinear dynamics of the vacancy concentration and the linear evolution of the cluster concentrations at fixed vacancy concentration. The dynamics of cluster concentrations is integrated by decomposing the initial condition and independently evolving the dynamics of small and large clusters.

We emphasized two stochastic methods in order to simulate the evolution of the concentration of large size clusters. The Fokker-Planck approximation is well known, but

our stochastic treatment with Langevin dynamics is more recent [Sur+04] in the cluster dynamics community. The Jump process approach is reminiscent of event kinetic Monte Carlo algorithms [Lan74] but can be parallelized much more efficiently and high frequency events associated with small size clusters are avoided. The main interest of these approaches is that they can be extended to higher dimensional situations. Moreover, with both methods, the particles are propagated independently, which allows to dispatch computations on a parallel architecture, henceforth decreasing the wall-clock computation time. With both methods the quantity of matter is accurately conserved and the distribution of concentrations we obtain is very close to the exact solution obtained by a numerical integration of the original full ODE system.

# 4 Some applications of the coupling algorithm

## 4.1 Introduction

In Chapter 2 and 3 we mainly focused on a simple but paradigmatic example of cluster dynamics, namely vacancy clustering under thermal ageing or irradiation. In particular, we showed that such a model was well-posed and that the algorithm we developed to speed up the simulation time was accurate with well controlled errors. Nevertheless, even if our algorithm is versatile and highly parallelizable, current methods of deterministic implementations of CD are more efficient [Jou+14] for such a model. The interest of our method relies in its application for more complex materials with different types of defects and higher dimensionality in the Fokker–Planck approximation.

In this chapter, we first show how to implement the hybrid deterministic/stochastic algorithm in the CD code CRESCENDO [Jou+14]. This code has been developed since 2009 at CEA and EDF R&D and is used for a broad range of applications for various materials. With this code, users can simulate two types of defects, self-defects (vacancies and interstitials) and one type of solute. Despite its efficiency, CRESCENDO presents some of the drawbacks of a purely deterministic approach (combinatorial explosion) and lacks some features in the case of materials with two types of defects (namely the inability to simulate systems with mobile clusters which contains self-defects and solute atoms).

The actual implementation of our algorithm is detailed in Section 4.2. Some modifications to the original algorithm of Chapter 3 have been made to satisfy technical constraints. In Section 4.3 we present two examples of applications. The first one, iron under neutron irradiation, allows us to validate the implementation and opens interesting perspectives concerning further improvements of our algorithm. The second example, the implantation of helium in $\alpha$-iron, allows us to discuss the efficiency of our method in a more challenging context.

We conclude the chapter with a different kind of application of our algorithm (see Section 4.4). In a recent work [Car+], it has been shown that the coefficients of emission and absorption in CD are not simple functions of cluster sizes, but are distributed according to a certain probability law. This dispersion in the coefficients creates a discrepancy in CD simulations with results obtained using an OkMC approach. The versatility of our hybrid algorithm allows us to easily take into account such dispersion and to quantitatively improve the results obtained with cluster dynamics.

## 4.2 Implementation of the coupling algorithm in CRESCENDO

The structure and features of CRESCENDO have been extensively described in [Jou+14]. In this section, we highlight the advantages and drawbacks of the code (Section 4.2.1) and which modifications to the algorithm described in Chapter 3.2.3 were made to overcome technical constraints (Section 4.2.2).

### 4.2.1 The CD code CRESCENDO

We first present a one-dimensional model and explain briefly why the algorithm is efficient for this case. We then present a two-dimensional case and highlight some of the constraints of the original approach of CRESCENDO, *i.e.* a coupling between ODEs of Cluster Dynamics and the Fokker–Planck approximation.

#### General model

Let us first recall the general model for cluster dynamics (1.5)–(1.6). We define $\mathcal{M}$ the set of mobile clusters, whereas $\Omega$ is the set of all clusters. The evolution of the concentrations for mobile clusters of type $\nu \in \mathcal{M}$ reads

$$\frac{dC_\nu}{dt} = G_\nu + \sum_{\mu\in\mathcal{M}} (\beta_{\nu-\mu,\mu}C_{\nu-\mu}C_\mu - (\beta_{\nu,\mu}C_\mu + \alpha_{\nu,\mu})C_\nu + \alpha_{\nu+\mu,\mu}C_{\nu+\mu}) \\ - \sum_{\mu\in\Omega} (\beta_{\nu,\mu}C_\nu C_\mu - \alpha_{\nu+\mu,\mu}C_{\nu+\mu}) - \sum_{i=1}^{N_i} k_{i,\nu}^2 D_\nu (C_\nu - C_\nu^{\mathrm{eq}}),$$

where $G_\nu$ is a source term due to irradiation, $\beta_{\nu,\mu}$ is the absorption rate of a cluster of type $\mu$ by a cluster of type $\nu$ and $\alpha_{\nu,\mu}$ is the emission rate of a cluster of type $\mu$ by a cluster of type $\nu$. Interactions with sinks are characterized by the sink strength $k_{i,\nu}^2$ of type $i$ and thermal equilibrium concentrations $C_\nu^{\mathrm{eq}}$. Two types of sinks are usually considered.

1. *Dislocations.* The sink strength reads $k_{i,\nu}^2 = Z_{i,\nu}\rho_i$, where $\rho_i$ is the dislocation density of type $i \in \mathcal{N}_d$ and $Z_{i,\nu}$ an efficiency factor, characterizing the capacity of a sink to capture a defect.
2. *Grain boundaries.* This type of sink describes the complexity of polycrystalline materials where the sink strength depends on the total sink strength in a single crystal:
$$k_{j,\nu}^2 = \frac{6}{l_{\mathrm{gb}}}\sqrt{\sum_{i\in\mathcal{N}_d} k_{i,\nu}^2 + \frac{1}{D_\nu}\left(\sum_{\mu\in\Omega}\beta_{\mu,\nu}C_\mu + \sum_{\mu\in\mathcal{M}}\beta_{\nu,\mu}C_\mu\right)},$$
where $l_{\mathrm{gb}}$ is a typical grain size.

Equilibrium concentrations are usually given by $C_\nu^{\mathrm{eq}} = V_{\mathrm{at}}^{-1}\exp(-E_\nu^{\mathrm{f}}/k_{\mathrm{b}}T)$ where $V_{\mathrm{at}}$ is the atomic volume, $E_\nu^{\mathrm{f}}$ the formation energy of a cluster of type $\nu$ and $T$ the temperature. For immobile clusters, the dynamics reads

$$\frac{dC_\nu}{dt} = G_\nu + \sum_{\mu\in\mathcal{M}} \beta_{\nu-\mu,\mu}C_{\nu-\mu}C_\mu - (\beta_{\nu,\mu}C_\mu + \alpha_{\nu,\mu})C_\nu + \alpha_{\nu+\mu,\mu}C_{\nu+\mu}. \tag{4.1}$$

#### One-dimensional case

Consider first the case where no solutes are taken into account. In this one-dimensional model, $\nu$ and $\mu$ are integers and called $n$ and $m$ respectively. Moreover, the set of mobile clusters only contains small size clusters. Typically, $\mathcal{M} = \{-m_{\mathrm{v}}, \cdots, -1, 1, \cdots, m_{\mathrm{i}}\}$, where $m_{\mathrm{v}} > 0$ is the maximal size of mobile vacancy clusters and $m_{\mathrm{i}} > 0$ the maximal size of mobile interstitial clusters. The CD code CRESCENDO relies on a deterministic approach using the Fokker–Planck approximation. Introduce the scalar field $\mathscr{C}$ such

that $\mathscr{C}(t,n) \simeq C_n(t)$ for $|n| \gg 1$ and the functions $\alpha_m, \beta_m$ such that $\alpha_m(n) = \alpha_{n,m}$ and $\beta_m(n) = \beta_{n,m}$ for all $n, m \in \mathbb{Z}^*$. Then, the approximation reads, for $|n| \gg |m_\mathrm{i}|, |m_\mathrm{v}|$,

$$\frac{\partial \mathscr{C}}{\partial t} = -\frac{\partial (F\mathscr{C})}{\partial x} + \frac{1}{2}\frac{\partial (D\mathscr{C})}{\partial x^2}, \tag{4.2}$$

with

$$F(x) = \sum_{m=-m_\mathrm{v}}^{m_\mathrm{i}} m\left(\beta_m(x) C_m - \alpha_m(x)\right), \qquad D(x) = \sum_{m=-m_\mathrm{v}}^{m_\mathrm{i}} m^2\left(\beta_m(x) C_m + \alpha_m(x)\right).$$

Formally, this approximation is obtained as in Chapter 2.3.1. Note the presence of factors $m$ and $m^2$ in the functions $F$ and $D$ respectively, which naturally appear in a Taylor expansion. It is interesting to note that when only monomers are mobile (*i.e.* $m_\mathrm{v} = m_\mathrm{i} = 1$), discretizing the PDE (4.2) with a centered scheme and mesh size $\Delta x = 1$ leads to the initial rate equations (4.1). Therefore, a seamless coupling between the rate equations and the Fokker–Planck approximation is possible. Some adaptations are nevertheless required in general to ensure mass and fluxes conservation between the two parts [Jou+14].

### The CVODES solver

The efficiency of CRESCENDO stems from two features. The Fokker–Planck approximation is the first key feature as it allows to reduce the number of ODEs to solve. Indeed, if a transition part with mesh size $\Delta x = 1$ ensures a seamless coupling between the rate equations and the Fokker–Planck approximation, then a mesh with a geometrical progression is chosen in order to reduce the number of calculations. The second key feature in CRESCENDO is the use of an efficient ODE solver. The solver used is CVODES of the library SUNDIALS [Hin+05]. It is particularly well suited to stiff systems since it is based on Backward Differential Formula algorithms [Coh+96]. Moreover, the solver provides an adaptive time stepping with variable order schemes, which allows to reach large time steps.

### Two-dimensional case

The case where we consider solute atoms is more complex and highlights some limitations of the method. Indeed, the Fokker–Planck approximation is only valid for large size clusters, in particular when the absorption and emission coefficients vary slowly with the size of the cluster. For mixed clusters of type $\nu = (n, p)$, where $n \in \mathbb{Z}$ characterizes the number of self-defects (vacancy or interstitial) and $p \in \mathbb{N}$ the number of solute atoms, the Fokker–Planck approximation might for example be valid along the self-defect axis if $n \gg 1$ but not in the solute direction if $p$ is small. In fact, in this configuration, the approximation is more precisely a set of coupled reaction-advection-diffusion equations along the self-defects axis. If mobile clusters containing self-defects and solute atoms are taken into account, discretization schemes are much more challenging to conceive. Nevertheless, in the particular case where clusters containing self-defects and solute atoms are immobile, the same discretization scheme as in the one-dimensional case can be used [Jou+14]. This is because the approximation can be decoupled between reaction rate equations, for the evolution involving pure solutes (or self-defects) clusters, and a Fokker–Planck equation, for the evolution along the self-defects (or solutes) axis. A decomposition of the domain is illustrated in Figure 4.1.

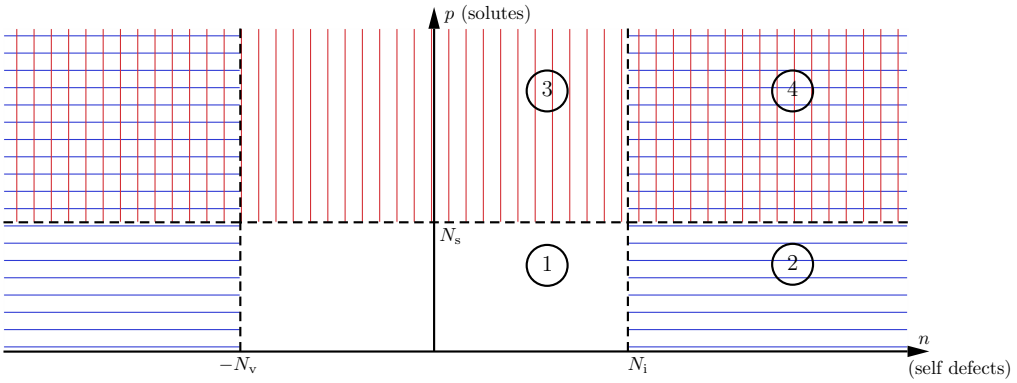



**Fig. 4.1:** Decomposition of the domain in CRESCENDO, describing the evolution of the concentrations in different ways. 1. (Blank) Rate equations. 2. (horizontal blue lines) Rate equations for the transitions involving pure solute clusters and one dimensional Fokker–Planck equation for transitions involving clusters composed of self-defects. 3. (vertical red lines) rate equations for transitions involving clusters of self-defects and one dimensional Fokker–Planck equation for transitions involving pure solute clusters. 4. (horizontal blue lines and vertical red lines) two-dimensional Fokker–Planck equation.

## 4.2.2 Modifications of the hybrid coupling algorithm for CRESCENDO

Two main modifications were brought to the hybrid deterministic/stochastic coupling algorithm when implementing it in CRESCENDO. The first one concerns the splitting of the dynamics which is adapted in order to preserve the main structure of the code for technical and performance reasons. The second one concerns the use of both stochastic approaches (EkMC and Langevin) for large size clusters, which improves the stability and efficiency of the method.

### Modification of the splitting

The use of the CVODES library for solving the deterministic part proves to be very efficient. In view of the structure of CRESCENDO, the implementation of a splitting between the non-linear dynamics on mobile clusters with the remainder of the concentrations would have increased the numerical cost of the method. The solver CVODES indeed needs to be restarted if different dynamics are computed, which automatically reduces the adaptive time step and therefore slows down the computations. Let us recall the original formulation of the splitting before presenting the modifications which we brought in CRESCENDO.

**Original splitting of the algorithm:** Formally, one step of the splitted dynamics is encoded by the mapping $C^{k+1} = S_{\Delta t}\left(C^k\right)$ for a given time step $\Delta t > 0$, defined as

1. update the concentrations of mobile clusters as $\widetilde{C}^{k+1}_{\mathcal{M}} = \varphi_{\Delta t}\left(C^k_{\mathcal{M}}\,;\, C^k_{\Omega\backslash\mathcal{M}}\right)$,
2. update the remaining concentrations as $\widetilde{C}^{k+1}_{\Omega\backslash\mathcal{M}} = \chi_{\Delta t}\left(C^k_{\Omega\backslash\mathcal{M}}\,;\, C^{k+1}_{\mathcal{M}}\right)$.

We recall that $C_X$ is the projection of $C$ on the subset $X \subset \Omega$, while $\widetilde{C}_X$ might contain non-zero elements in $\Omega\backslash X$ (see Chapter 3.2.3 for a more precise description of the algorithm). Moreover, considering $\Omega\backslash\mathcal{M} = \mathcal{S} \cup \mathcal{L}$, the decomposition between immobile clusters, where the disjoint sets $\mathcal{S}$ and $\mathcal{L}$ represent small and large size clusters respectively, it holds $S_{\Delta t}\left(C^k\right) = \chi_{\Delta t}\left(C^k_{\mathcal{S}}\,;\, C^{k+1}_{\mathcal{M}}\right) + \chi_{\Delta t}\left(C^k_{\mathcal{L}}\,;\, C^{k+1}_{\mathcal{M}}\right)$.

**Modified splitting of the algorithm:** In CRESCENDO, we solve the evolution of mobile clusters with the evolution of small size ones. It consists in the following splitting:

1. update the concentrations of mobile clusters and small immobile ones as

$$\widetilde{C}^{k+1}_{\mathcal{M}\cup\mathcal{S}} = \varphi_{\Delta t}\left(C^k_{\mathcal{M}}\,;\, C^k_{\mathcal{S}\cup\mathcal{L}}\right) + \chi_{\Delta t}\left(C^k_{\mathcal{S}}\,;\, C^k_{\mathcal{M}}\right),$$

2. update the remaining concentrations as $\widetilde{C}^{k+1}_{\mathcal{L}} = \chi_{\Delta t}\left(C^k_{\mathcal{L}}\,;\, C^{k+1}_{\mathcal{M}}\right)$.

The splitting is still of order 1. In fact it proves to be more precise and numerically stable in view of our different experimentations in the implementation, which we do not report here.

Another modification we brought in the implementation of the algorithm is the use of adaptive time steps for the splitting. We observed that the transient regime in CD models under irradiation is longer than under thermal ageing. In particular, concentrations of mobile clusters evolve rapidly, which requires to carefully choose the time steps for the splitting. To choose the time step, we fix an upper threshold for the concentration $C(N_{\text{front}}+N_{\text{buff}})$. When the deterministic concentration reaches this threshold, we stop the simulation and then compute the stochastic part of the algorithm. Then, after handling the stochastic part, we restart the solver CVODES and compute the deterministic simulation, which gives us the next splitting time step. In fact, the time step obtained this way indirectly depends on the concentrations of mobile clusters. This dependence can be explained as follows: the advection part of the Fokker–Planck approximation gives a good estimate of the velocity of a stochastic particle at a position $x$, namely $F(x)$, which is proportional to the concentrations of mobile clusters. Consider, for example, the simple case of vacancy clustering. For large sizes, the coefficient $\alpha_n$ is usually negligible in front of $\beta_n C_1$. Physically, the concentration of mobile clusters decreases with time, which means that the velocity decreases, and therefore the time step increases.

### Coupling both stochastic approaches

Another important change in the coupling algorithm involves the way the stochastic part is handled. As noted in the description of CRESCENDO for two-dimensional models, solving the dynamics in the sub-domains 2 and 3 (see Figure 4.1) might be quite challenging, especially when mobile clusters containing different types of defects are taken into account. To overcome this problem, the jump process approach is well suited since it solves exactly the dynamics, in particular for the transitions involving clusters with a small number of single type defects. In addition, we observed that coupling both stochastic approaches gives better results. The conversion of deterministic concentrations into stochastic particles indeed takes place on the discrete mesh $\{N_{\text{front}}, \cdots, N_{\text{front}} + N_{\text{buff}}\}$. Using the jump process approach in this area allows a seamless coupling since this dynamics is exact (see Chapter 3.3.1). Then, when a stochastic particle reaches a size greater than $N_{\text{front}} + N_{\text{buff}}$, it becomes a Langevin particle which evolves along the real line $[N_{\text{front}} + N_{\text{buff}}, +\infty)$. Naturally, if a Langevin particle reaches a size smaller than $N_{\text{front}} + N_{\text{buff}}$, it evolves again according to the jump process along a discrete mesh. The conversion of a Langevin particle into a discrete stochastic particle is as follows. If the size of the Langevin particle is $x \in [n, n+1]$, then it becomes a discrete stochastic particle of size $n$ with probability $p_x = 1 - x + n$ and of size $n+1$ with probability $1 - p_x$ (see Figure 4.2).

A last change is made in handling the stochastic particles in the deterministic zone. In fact, stochastic particles can fully overlap the deterministic zone except for the mobile clusters domain (this limitation comes from the splitting of the dynamics between mobile clusters and the remainder of the concentrations). Moreover, we avoid a transformation of stochastic particles into deterministic concentrations in the deterministic domain since we observed some numerical instabilities which remain unexplained (a bug is not excluded due to the complexity of the code). This new decomposition of the domain is illustrated in Figure 4.2. Finally, let us note that using a stochastic approach allows us to handle models with mobile clusters which contain self-defects and solutes.

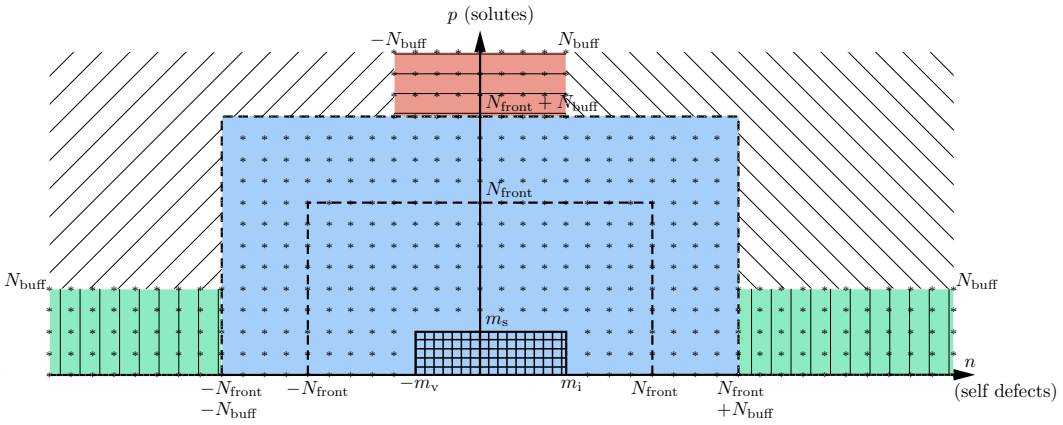



**Fig. 4.2:** Domain decomposition in CRESCENDO implemented with the hybrid dterministic/stochastic algorithm. To simulate the time evolution of cluster concentrations, the deterministic solver CVODES is used inside the blue zone. The sub-domain with stars also describe the evolution of the concentrations using the jump process. The domain with hatched lines describe the evolution of the concentrations using the Langevin process. In the green zones, particles evolves according to the Langevin process in the $n$-direction and according to the jump process in the $p$-direction. It is the opposite in the red zone. Deterministic concentrations are transformed into stochastic particles when $|n| \geqslant N_{\text{front}}$ or $p \geqslant N_{\text{front}}$.

## 4.3 Using the coupling algorithm in CRESCENDO for Fe and FeHe under irradiation

We present results of simulations for two systems. The first one is a one-dimensional problem, which allows us to give a precise description of the observed phenomenons during our simulations. We discuss which aspects of our algorithm still need improvement. The second model is a two-dimensional problem. We discuss the performance of our algorithm in comparison with purely deterministic approaches.

### 4.3.1 Fe under irradiation

In this model, we consider pure $\alpha$-iron under irradiation by neutrons. Vacancy clusters are mobile up to a size $m_{\text{v}} = 4$ and interstitial clusters up to $m_{\text{i}} = 3$. The temperature of the system is $T = 573K$, which is a typical temperature for reactors. Finally, the total damage rate is $G_{\text{tot}} = 5 \times 10^{-9}$ dpa/s (displacement per atoms per second). For vacancy clusters, it is distributed amongst monomers ($G_{-1} = 4.5 \times 10^{-9}$ dpa/s) and vacancy clusters of size 10 ($G_{-10} = 5 \times 10^{-11}$ dpa/s), so that $G_{\text{tot}} = G_{-1} + 10G_{-10}$. For interstitial clusters, $G_{\text{tot}}$

is distributed amongst monomers ($G_1 = 4.5 \times 10^{-9}$ dpa/s) and interstitial clusters of size 10 ($G_{10} = 5 \times 10^{-11}$ dpa/s), so that $G_{\text{tot}} = G_1 + 10G_{10}$. Note that even if damage rates are the same between vacancies and interstitials (see Case 2), it does not presume that the distributions of the concentrations between vacancies and interstitials will be similar since the absorption and emission coefficients are different. We refer to [Jou+14] for a description of the other parameters of the model.

### Results and discussions

Let us now discuss the results we obtained in this one-dimensional example. We first observed that we should carefully choose the frontier $N_{\text{front}}$ and the size of the buffer $N_{\text{buff}}$ between the deterministic and the stochastic domains. We observed that there is a discrepancy between the results of our simulations with the reference solution when $N_{\text{front}}$ is too small. This is illustrated in Figure 4.3, for which 3 different sets of parameters were considered:

**Case 1** $N_{\text{front}} = 10$ and $N_{\text{buff}} = 40$.

**Case 2** $N_{\text{front}} = 30$ and $N_{\text{buff}} = 40$.

**Case 3** $N_{\text{front}} = 60$ and $N_{\text{buff}} = 40$.

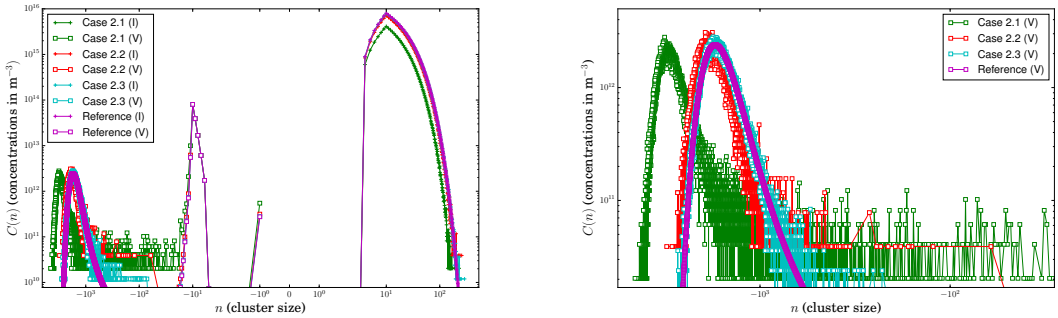



**Fig. 4.3:** Left: Comparison between a reference solution obtained by solving the rate equations of CD with CVODES and our algorithm with various values of $N_{\text{front}}$. The number of particles is $N_{\text{part}} = 1.2 \times 10^6$. Right: focus on the vacancy part of the left figure.

Let us now explain how the number of stochastic particles in the deterministic domain on the interstitial side, along with large time steps, affects the numerical accuracy. Since the stochastic particles do not directly contribute to the deterministic dynamics, due to the splitting and decomposition of the dynamics, the error might increase during the simulation. Comparing Cases 1, 2 and 3 clearly confirms this trend. The concentrations of stochastic particles are two orders of magnitude lower than the deterministic concentrations of the reference in Case 3 when the concentrations of stochastic particles are less than one order of magnitude lower than the deterministic ones in Case 1 (see Figure 4.4). Therefore, since the splitting error only comes from the stochastic part (see the discussion in Section 4.2.2), the error is more important when there are more stochastic particles near the mobile zone.

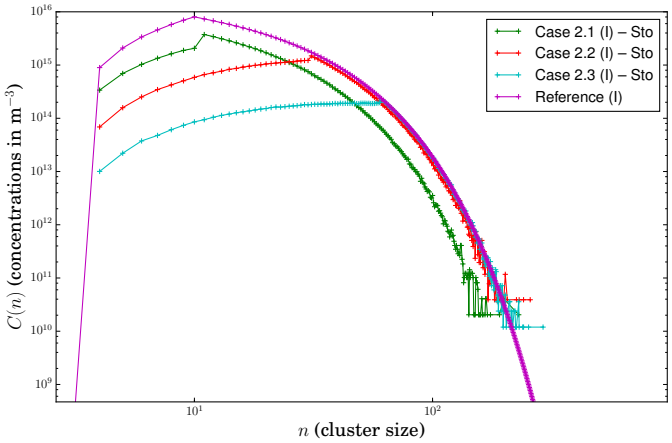



**Fig. 4.4:** Concentrations of the stochastic part of our algorithm with various values of $N_{\text{front}}$.

### Accelerating the dynamics through adaptive time steps

As explained in Section 4.2.2, we use adaptive time steps in order to speed up the computations. In Figure 4.5, we observe that the time step of the splitting $\Delta t_{\text{splitting}}$ increases during the simulation. This phenomenon is explained by the fact that, as time increases, the concentration of mobile clusters decreases, which decreases the intensity of the drift $F$, so that, at the end of the simulation, we reach time steps of the order of several hours.

However, our approach is severely limited in efficiency. This limitation is explained by the evolution of the internal time step of the solver CVODES, displayed in the right part of Figure 4.5. The solver is restarted every macroscopic time step $\Delta t_{\text{Splitting}}$. Since the algorithm of CVODES uses adaptive time steps in order to control the error and since we modify the concentration by turning deterministic concentrations into stochastic particles, we observe a sudden decrease of the value $\Delta t_{\text{CVODES}}$. In our implementation, we tried to avoid the restart of CVODES by internally changing the vector of concentrations, but the result is the same.

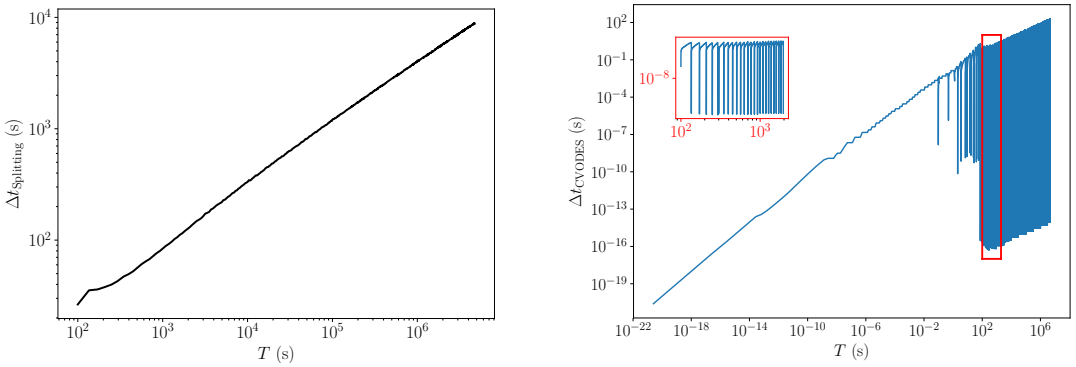



**Fig. 4.5:** [Case 3] Left: Evolution of $\Delta t_{\text{Splitting}}$ during the simulation. Right: Evolution of $\Delta t_{\text{CVODES}}$ during the simulation.

### 4.3.2 FeHe under irradiation

We next explore a two-dimensional case of $\alpha$-iron with Helium. The physical parameters are also described in [Jou+14] and the irradiation conditions are the same as in Section 4.3.1, with an additional irradiation of Helium ions. The insertion rate is such that $G_{\mathrm{He}} = G_{(0,1)} = 10^{-12}$ pa/s (per atom per second). We also fixed the frontier and buffer size to $N_{\mathrm{front}} = 60$ and $N_{\mathrm{buff}} = 40$.

#### Scaling and performance

We first observe the performance of the stochastic part compared to the deterministic part which is handled by CVODES. For a fixed number of stochastic particles ($N_{\mathrm{part}} = 10^6$, we ran simulations on 1 up to 28 processors. The number of particles per processor decreases as the number of processors increases. As expected, we obtain a linear scaling for the time spent by the simulation in the stochastic computation, while the time spent in the deterministic computation remains constant (see Figure 4.6).

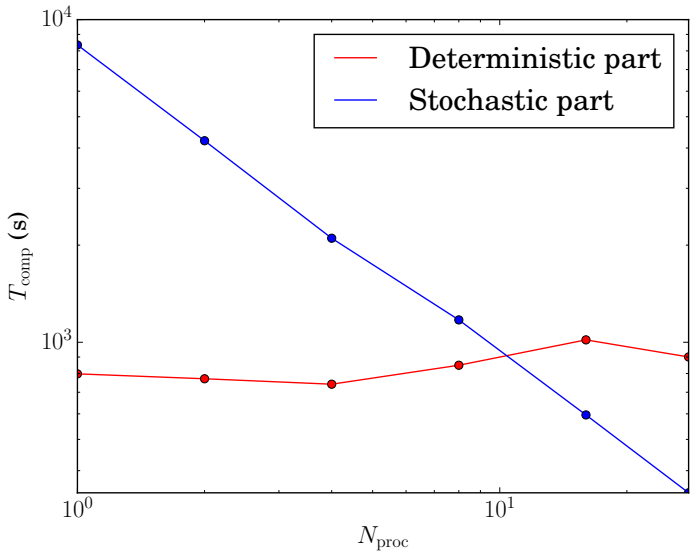



**Fig. 4.6:** Comparison of the time spent by the algorithm computing the deterministic part and the stochastic part.

Then, we compare the wall-clock time of the simulations for different algorithms. We used CVODES for solving the full CD, then two different algorithms for solving the dynamics using the Fokker–Planck approximation. The first algorithm is based on a Chang-Cooper scheme for discretizing the Fokker–Planck equation. This scheme is extremely efficient but generates a numerical diffusion in the distribution. The second algorithm is based on the more precise MP5 scheme, which is much slower. Our hybrid algorithm with $N_{\mathrm{part}} = 2.8 \times 10^6$ particles is less efficient than the Chang-Cooper based algorithm, but more precise and it is competitive in comparison of the MP5 based algorithm (see Table 4.1).

We finally illustrate the results of our simulations with our algorithm and compare it with a reference solution obtained by solving the rate equations of cluster dynamics using the CVODES solver. In Figure 4.7, we observe a good agreement between the reference solution and the one obtained with our algorithm. We still observe some noise with 28

| | |
|---|---|
| Full CD with CVODES | 28h03 |
| Hybrid Deterministic/Stochastic | 21h53 |
| Chang-Cooper | 13min |
| MP5 | 287h40 |

**Tab. 4.1:** Comparison of the performance of various algorithms for the simulation of a FeHe system.

millions of particles, but with an increased number of processors we could reduce this noise with the same wall-clock time.

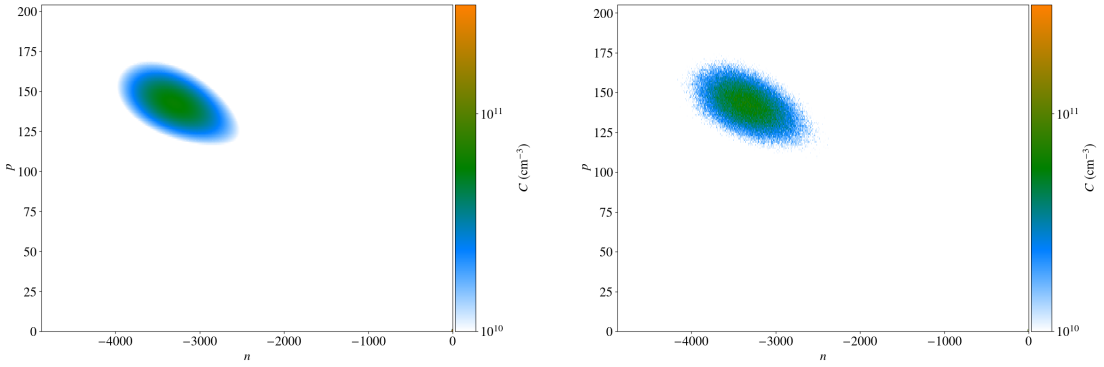



**Fig. 4.7:** Left: Deterministic simulation using CVODES with CD rate equations. Right: Hybrid deterministic/stochastic simulation.

### Discussion

The system of ODEs is integrated deterministically through time-stepping using an implicit integrator based on multi-step backward differential formulae (BDF) and resorting to CVODES library. The implemented BDF integrator is very efficient for purely deterministic simulations. The efficiency however decreases dramatically in hybrid deterministic/stochastic simulations wherein a large fraction of the CD equations are propagated stochastically. The stochastic propagation introduces statistical fluctuations on cluster concentrations. These concentration equations being input data of the BDF scheme, the superimposed statistical noise is understood as the signature of spurious fast modes. As a result, the CVODE solver is automatically reinitialized. As a result, the internal time step of the solver is initially reduced by several orders of magnitude at the beginning of each integration cycle, *i.e* every $\Delta t_{\text{splitting}}$. The mean internal time step is therefore drastically reduced, by a factor of about 100 and the overall performance of the hybrid method is reduced by the same amount. When the stochastic propagation of the particles is parallelized on a large number of processes, the numerical integration of the deterministic part of the CD equations becomes the limiting factor. Future developments should therefore focus on the numerical solver used for integrating the system of ODEs.

## 4.4 Improving Cluster Dynamics with stochastic coefficients

As stated in the Section 1.3, CD is the mean field counterpart of OkMC, leading to results in very good agreement if spatial correlations between defects can be neglected. Nevertheless, it has been shown that distributions obtained with OkMC can be broader than those obtained with CD [JC18].

A recent work based on OkMC simulations has demonstrated that there is a dispersion of absorption and emission coefficients [Car+]. This dispersion should be taken into account in order to make CD more precise. We first consider a very simple model of cluster dynamics, still paradigmatic of the dynamics, where we only consider absorption coefficients. We then present in which way cluster dynamics is changed by introducing the dispersion of these coefficients and explain why the hybrid deterministic/stochastic algorithm is useful for the computations. We finally present some results confirming the validity of this new model.

### 4.4.1 Introducing sink strength dispersion in cluster dynamics simulations

The simplified CD model corresponding to the OKMC calculations, where a dispersion of the absorption coefficients (and therefore emission coefficients) is observed, reads

$$\frac{dC_n}{dt} = \beta_{n-1}C_{n-1}C_1 - \beta_n C_n C_1 \qquad n \geqslant 2 \tag{4.3}$$

$$\frac{dC_1}{dt} = G_1 - \beta_1 C_1^2 - \sum_{n\geqslant 1} \beta_n C_n C_1, \tag{4.4}$$

where $C_n$ is the concentration of a cluster containing $n$ monomers and

$$\beta_n = 4\pi r_n D_1,$$

where $D_1$ is the diffusion coefficient of a monomer and $r_n$ the radius of a cluster of size $n$. In this specific case, we consider interstitial defects, namely SIA for Self Interstitial Atoms, and the radius reads

$$r_n = \sqrt{\frac{nV_{\text{at}}}{\pi b}}$$

where $V_{\text{at}}$ is the atomic volume and $b$ the norm of the Burgers vector. To introduce sink strength dispersion in these equations, the absorption coefficient $\beta_n$ is assumed to depend on the normalized Voronoi volume $v$ for a size larger than $n^*$, with $n^* \geqslant 2$, and two parameters $\alpha$ and $\beta$ (see [Car+]):

$$\beta_n(v) = 4\pi r_n \left(v^\alpha + \beta\right) D_1. \tag{4.5}$$

The cluster concentration of a given class $n$ now also depends on $v$, so it is noted $C_n(v)$. It has been noted [Car+] that large clusters are more present in large Voronoi volumes than in small ones. This means that the change of neighbourhood of a cluster, or in other words the change of its Voronoi volume, due to the creation of a cluster nearby, must happen over timescales which are sufficiently large with respect to the growth process. Accordingly, the neighbourhood of a cluster is assumed to remain the same. This approximation can

lead to an overestimation of the dispersion, which can be checked a posteriori on cluster distributions. Finally, the dispersion on the absorption coefficient is given by the dispersion on $v$ which is distributed according to a probability law $P(v)$, namely the distribution of normalized Voronoi volumes of a Poisson-Voronoi tessellation [Kum+92], where

$$P(v) = \frac{v^{\zeta-1}}{\lambda^{\gamma}\Gamma(\zeta)} \exp\left(-\frac{v}{\lambda}\right),$$

and the parameters $\lambda$ and $\zeta$ have been determined in [Laz+13]. Note that the conservation law now reads

$$\frac{d}{dt}\left(\sum_{n=1}^{n^*-1} nC_n + \sum_{n\geqslant n^*}\int_0^{\infty} nC_n(v)P(v)dv\right) = 0.$$

With the choice (4.5) introducing a dispersion, Equations (4.3)–(4.4) become

$$\begin{aligned}
\frac{dC_n}{dt} &= \beta_{n-1}C_{n-1}C_1 - \beta_n C_n C_1, && 2 \leqslant n \leqslant n^*-1 \qquad (4.6)\\
\frac{dC_n(v)}{dt} &= P(v)\beta_{n-1}C_{n-1}C_1 - \beta_n(v)C_n(v)C_1, && n = n^*,\ v \in ]0,\infty[\\
\frac{dC_n(v)}{dt} &= \beta_{n-1}(v)C_{n-1}(v)C_1 - \beta_n(v)C_n(v)C_1, && n > n^*,\ v \in ]0,\infty[\\
\frac{dC_1}{dt} &= G_1 - \beta_1 C_1 C_1 - \sum_{n=1}^{n^*-1}\beta_n C_n C_1 - \int_0^{\infty}\sum_{n\geqslant n^*}\beta_n(v)C_n(v)C_1P(v)\,dv. && (4.7)
\end{aligned}$$

To solve these equations numerically, two possible methods can be used, one purely deterministic and another one based on our hybrid deterministic/stochastic algorithm.

### Deterministic approach

The first method consists in working with an empirical measure of $P(v)$. In practice, we consider a discretized version of $P(v)$, namely

$$P_N(v) = \Delta v \sum_{i=1}^{N} \delta_{v_i}(v)P(v_i)$$

where $N > 0$ represents the number of equally spaced possible normalized Voronoi volumes, $\Delta v$ is the spacing between two values of $v$, so that $v_i = v_{\min} + i\Delta v$, and $\delta_{v_i}$ is the usual Kronecker symbol. Since the values of $v$ are discretized, so are the concentrations which can be noted $C_{n,i}$, where $i = 1,\dots,N$. Values $v_i$ between $v_{\min} = 0$ and 5 are sufficient to accurately sample the Poisson-Voronoi distribution (see [Car+]). We also use the notations $\beta_{n,i} = \beta_n(v_i)$ and $P_i = P(v_i)$. Finally, let us note that $\sum_{1\leqslant i\leqslant N} P_i\Delta v \simeq 1$. Then, we obtain a new set of equations to solve:

$$\begin{aligned}
\frac{dC_n}{dt} &= \beta_{n-1}C_{n-1}C_1 - \beta_n C_n C_1, && 2 \leqslant n \leqslant n^*-1 \qquad (4.8)\\
\frac{dC_{n,i}}{dt} &= \Delta v P_i \beta_{n-1}C_{n-1}C_1 - \beta_{n,i}C_{n,i}C_1, && n = n^*,\ i \in [1,N]\\
\frac{dC_{n,i}}{dt} &= \beta_{n-1,i}C_{n-1,i}C_1 - \beta_{n,i}C_{n,i}C_1, && n > n^*,\ i \in [1,N]
\end{aligned}$$

$$\frac{dC_1}{dt} = G_1 - \beta_1 C_1 C_1 - \sum_{1 \leqslant n \leqslant n^* - 1} \beta_n C_n C_1 - \sum_{i=1}^{N} \Delta v P_i \sum_{n \geqslant n^*} \beta_{n,i} C_{n,i} C_1. \tag{4.9}$$

This set of equations can be readily introduced in the CD code CRESCENDO, which enables the use of different cluster populations coupled together only through the mobile species [Jou+14]. Here the different cluster populations correspond actually to the same clusters, but in different environments. The numerical cost can increase substantially, since the number of equations is roughly multiplied by $N$ compared to a classical calculation involving a single population.

### Using the new algorithm for CD

A more elegant way to solve equations (4.6)–(4.7) is to resort to the hybrid deterministic/stochastic scheme introduced in Chapter 3. In this method, small clusters ($n < n^*$) are treated deterministically, while cluster dynamics equations are solved stochastically for larger sizes, using either a jump process approach or the Langevin process associated to the Fokker-Planck approximation. The deterministic and stochastic regions are separated by a buffer region where the transfer between deterministic cluster density and stochastic particles is performed. The dispersion of absorption coefficients can be naturally introduced in this method. Each time a stochastic particle is created, due to the flux of clusters from the deterministic region to the stochastic region, a normalized Voronoi volume, drawn according to the Poisson-Voronoi distribution $P(v)$, is associated to this particle. The particle then evolves according to the value of the absorption coefficient corresponding to the normalized Voronoi volume.

### Results

Results are compared to OKMC simulations for $t = 10^{-3}$ s in Figure 4.8. The deterministic part in the hybrid approach was chosen equal to $N_{\text{front}} = 20$ and the size of the buffer region is $N_{\text{buff}} = 30$. This ensures good performance and precision for the hybrid algorithm. For the fully deterministic simulation, $N = 25$ was used. It was checked, by varying this value, that in the conditions considered the cluster distributions are accurately simulated. For too low values of $N$ (namely $N \lesssim 15$), cluster distributions become distorted. The other parameters used for the simulations are given in Table 4.2.

| | |
|---|---|
| Atomic volume ($V_{\text{at}}$) | $1.66 \times 10^{-29}$ m$^3$ |
| First parameter of the Poisson-Voronoi distribution ($\lambda$) | 0.179 |
| Second parameter of the Poisson-Voronoi distribution ($\zeta$) | 5.586 |
| First parameter of the dispersion ($\alpha$) | 0.25 |
| Second parameter of the dispersion($\beta$) | 0.07 |
| SIA diffusion coefficient ($D_1$) | $8.611 \times 10^{-8}$ m$^2$/s |
| SIA creation rate ($G_1$) | $10^{-1}$ dpa/s |

**Tab. 4.2:** Parameters for the simulation of CD with dispersion

Rate equation of CD calculations including the dispersion are both in good agreement with the reference OKMC calculation, at variance with the classical CD calculation which

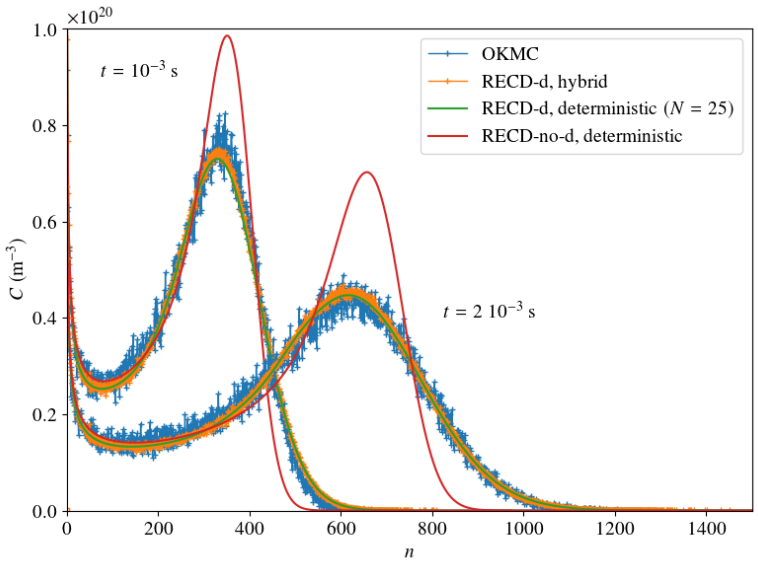



**Fig. 4.8:** Cluster distributions at $t = 10^{-3}$ s and $t = 2 \times 10^{-3}$ s, obtained with OKMC and different RECD models: hybrid deterministic-stochastic calculation using sink strength dispersion in the stochastic region (RECD-d, hybrid), deterministic calculation with sink strength dispersion based on Eqs. (4.8)–(4.9) with $N = 25$ (RECD-d, deterministic), deterministic calculation without dispersion (RECD-no-d, *i.e.* $P(v) = \delta_{v=0}$). Hybrid calculations were performed with 2 million stochastic particles.

leads to a too peaked distribution. If the calculation is continued up to $t = 2\ 10^{-3}$ s, the discrepancy with OKMC increases, whereas CD simulations including dispersion remain very close to OKMC. This result is rather encouraging concerning the generality of our approach.

The normalized sink strengths of the stochastic particles $\kappa = \beta / D_1$ is shown in Figure 4.9 as a function of the cluster radius, for $t = 10^{-3}$ s. The overall shape of the distribution is very similar to the one observed in [Car+], with large clusters mostly present in large Voronoi volumes. This is due to the fact that the environment of particles does not change with time in CD calculations. The similarity of sink strength distributions in CD and OKMC tends to confirm this hypothesis. A more refined model could be envisaged, by resampling some of the absorption rates of stochastic particles depending on the nucleation rate of clusters. For the present case, this additional complexity proved to be unessential.

The value of $n^*$ was fixed according to performance and accuracy constraints in the hybrid calculation. It can be easily changed in the deterministic calculation, to study the influence of the sink strength dispersion of small immobile clusters (Figure 4.10). By decreasing the value of $n^*$, the agreement is slightly worse than for $n^* = 20$, but it remains far better than with the RECD calculation without dispersion.

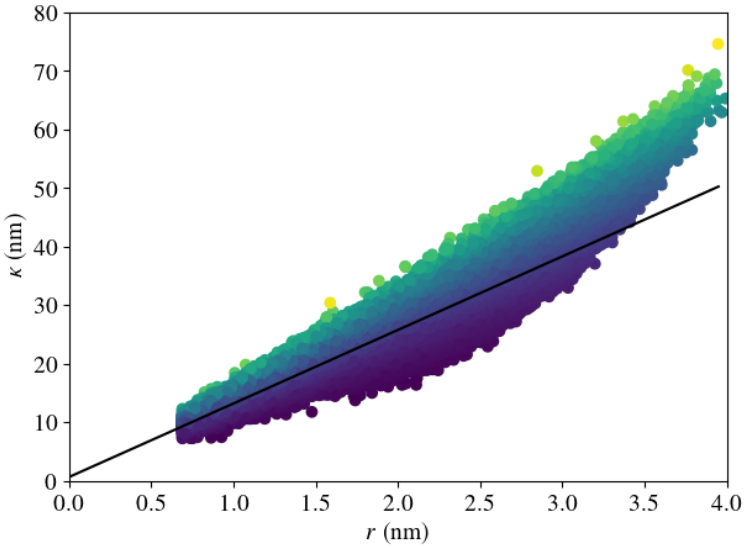



**Fig. 4.9:** Normalized sink strengths at $t = 10^{-3}$ s, extracted from the values ascribed to stochastic particles in RECD hybrid calculations.

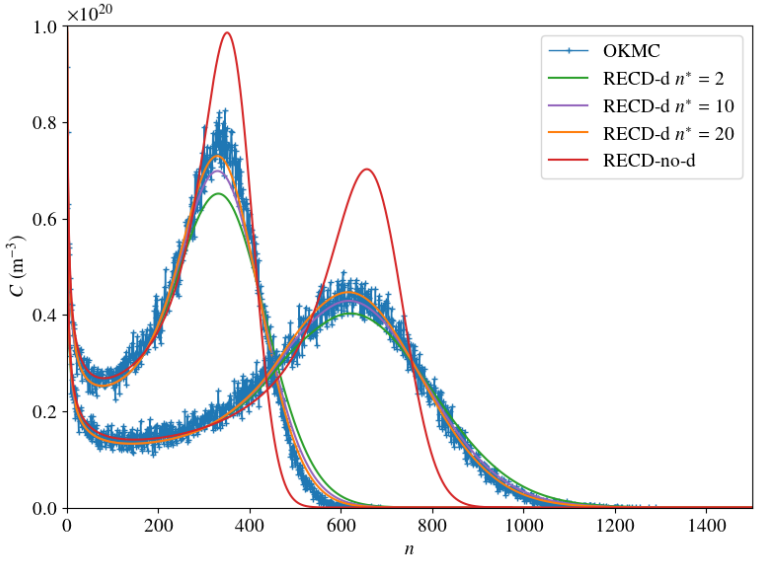



**Fig. 4.10:** Cluster distributions at $t = 10^{-3}$ s and $t = 2\ 10^{-3}$ s, obtained with OKMC, deterministic RECD calculation without sink strength dispersion (RECD-no-d) and RECD calculation including dispersion for clusters larger than $n^*$ (RECD-d).

# Conclusion and perspectives

# 5

## General conclusion

In this thesis, we have studied Cluster Dynamics from several points of view: we explored some mathematical properties of CD, we developed a new algorithm to simulate CD and we used it to improve the accuracy of CD in comparison to OkMC methods. The CD model has been used in the community of nuclear materials for several decades. It is considered to be efficient and accurate in comparison with other kinetic models. In particular, CD allows to study the evolution of defects in materials under irradiation over long periods of time, typically more than 40 years, the operating lifetime of nuclear power plants. This time scale is hardly achievable with AkMC or OkMC methods.

To our knowledge, CD was never studied from a mathematical perspective. This work tries to remedy this issue by proving several results in Chapter 2. We first prove that CD is well-posed in a precise mathematical framework. To prove the existence and uniqueness of a solution, we rely on techniques developed for the study of semigroups of operators and their applications to differential equations. We confirm the physical properties which are expected from CD: the sign of the solution and the quantity of matter are both preserved. We also introduce a splitting of the dynamics, used in particular in Chapter 3, and prove it to be consistent of order 1. Then, we study an approximation of the dynamics, namely an advection-diffusion partial differential equation, called the Fokker–Planck approximation. This approximation is still questioned [KW16] in the community even though it proved to be accurate and very efficient [Jou+14]. We hope that bringing a mathematical viewpoint to the question will help settling the argument. In particular, we rigorously relate CD and its Fokker–Planck approximation by resorting to stochastic tools enabling us to estimate the errors of the approximation. In particular, we relate CD and its Fokker–Planck approximation through a diffusion equation, whose solution is given by the Feynman–Kac representation formula.

The connection between ODEs, PDEs and stochastic processes is recurrent throughout this thesis and is best illustrated by the algorithm introduced in Chapter 2. Based on the splitting introduced in Chapter 2, we developed a hybrid deterministic/stochastic coupling algorithm for simulating CD equations. Thanks to the splitting, the evolution equations of immobile clusters are linear. Then, EDOs of CD can be seen as the forward-Kolmogorov equations of a Markov process, while the Fokker–Planck approximation is directly related to a stochastic process, namely a Langevin process. A stochastic approach, coupled with the underlying linearity, numerically reduces the number of equations to solve and allows a direct use of parallel computing.

Reducing the wall-clock computation time and memory usage was the primary motivation for developing a new hybrid deterministic/stochastic algorithm. Due to a curse of dimensionality which results in combinatorial explosion, the simulation of the evolution of complex materials might be challenging. Then, coupling a stochastic approach, which reduces the number of equations to solve, with a deterministic one which avoids the drawbacks of a fully stochastic approach — *i.e.* the simulation of high frequent events which might increase the computation time — in a versatile framework is promising. The

implementation of the algorithm in a code, called CRESCENDO, dedicated to the production of physical results is investigated in Chapter 4. While some technical constraints required to modify the original algorithm, the implementation and use of our algorithm produce promising and encouraging results. Moreover, the versatility of the algorithm proved to be very helpful to further improve the model of CD. Recent developments have shown the importance of simulating the dispersion of absorption and emission coefficients, which can be done by resorting to a specific probability law.

## Discussion and perspectives

This thesis is organized as a three-part work around Cluster Dynamics: a mathematical analysis, an algorithmic development and physical applications. Each part is still open for investigation and improvement.

The well-posedness of CD is proved in a simple but paradigmatic framework. We do not provide a proof of the well-posedness for more complex cases but the techniques we relied on can be extended to such situations. On the contrary, the decay estimates on the derivatives of the diffusion equation (see Chapter 2, Theorem 10) we have proved for studying the Fokker–Planck approximation heavily rely on a "one-dimensional" technique, namely a Lamperti transform. Therefore, we cannot generalize our results for more complex systems with different types of defects. Moreover, the relative error of the approximation is hard to quantify and can only be estimated numerically. Given these limitations, some questions are still open for investigation: can we mathematically quantify the relative error of the approximation? Is there a better mathematical approach to handle these equations?

Concerning the new algorithm and its applications, we have shown some limitations. Despite its versatility, the algorithm could be more efficient and accurate. Since the algorithm couples both deterministic and stochastic computations, let us describe how each part could be improved.

— Deterministic part. For historic reasons, the numerical solver used for computing the ODEs is CVODES and this imposes some constraints, in particular the restarting of the solver at each splitting time step. The use of an exponential integrator in combination with iterative Krylov subspace solvers should alleviate the difficulties encountered by the sequential multi-step integrator used in the hybrid CD simulations presented in Chapter 4. Assessing the performance of exponential integrators in traditional deterministic CD simulations is already an instructive study in itself. Given the currently observed poor performance of the sequential and multi-step integrator in the presence of statistical noise, we expect a substantial improvement in the efficiency of hybrid simulations owing to the use of parallel solvers with good scalability.

— Stochastic part. The accuracy of this part could be improved. Vacancies and interstitials are currently handled without distinction. However, the concentrations of vacancies and interstitials can differ by several orders of magnitude in physical models. Therefore, a certain type of defect can be poorly described since it won't be sufficiently sampled given a fixed number of particles. One solution is to handle these types of defects with dedicated classes of particles.

Finally, the actual structure of CRESCENDO only allows to describe systems with 2 types of defects or less. The implementation of the algorithm in another version of CRESCENDO, more suited for the study of complex materials, represents the next step of development in the coming years.

# Bibliography


[AB14] M. Athènes and V.V. Bulatov. „Path Factorization Approach to Stochastic Simulations“. In: *Physical Review Letters* 113.23 (2014), p. 230601 (cit. on p. 67).

[Bal+86] J. M. Ball, J. Carr, and O. Penrose. „The Becker–Döring cluster equations: basic properties and asymptotic behaviour of solutions“. In: *Communications in Mathematical Physics* 104.4 (1986), pp. 657–692 (cit. on pp. 11, 21).

[Ban22] S. Banach. „Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales“. In: *Fundamenta Mathematicae* 3.1 (1922), pp. 133–181 (cit. on p. 39).

[BC07] A. Barbu and E. Clouet. „Cluster dynamics modeling of materials: advantages and limitations“. In: *Solid State Phenomena* 129 (2007), pp. 51–58 (cit. on pp. 19, 22).

[BD35] R. Becker and W. Döring. „Kinetische Behandlung der Keimbildung in übersättigten Dämpfen“. In: *Annalen der Physik* 416.8 (1935), pp. 719–752 (cit. on p. 10).

[Bec+10] C.S. Becquart, A. Barbu, J.L. Bocquet, et al. „Modeling the long-term evolution of the primary damage in ferritic alloys using coarse-grained methods“. In: *Journal of nuclear materials* 406.1 (2010), pp. 39–54 (cit. on p. 3).

[Bel27] E.T. Bell. „Partition polynomials“. In: *Annals of Mathematics* (1927), pp. 38–46 (cit. on p. 58).

[Ber+11] F. Berthier, E. Maras, I. Braems, and B. Legrand. „Multiscale modelling of the ageing kinetics of a 2D deposit“. In: *Solid State Phenomena*. Vol. 172. Trans Tech Publications. 2011, pp. 664–669 (cit. on p. 13).

[Bor+75] A.B. Bortz, M.H. Kalos, and J.L. Lebowitz. „A new algorithm for Monte Carlo simulation of Ising spin systems“. In: *Journal of Computational Physics* 17.1 (1975), pp. 10–18 (cit. on pp. 3, 5, 19).

[Bur77] J. J. Burton. „Nucleation Theory“. In: *Statistical Mechanics*. Springer, 1977, pp. 195–234 (cit. on p. 11).

[Car+] D. Carpentier, T. Jourdan, P. Terrier, M. Athènes, and Y. Le Bouar. „Effect of sink strength dispersion on particle size distributions simulated by cluster dynamics“. In preparation (cit. on pp. 89, 99, 100, 102).

[Cat+00] M.J. Caturla, N. Soneda, E. Alonso, et al. „Comparative study of radiation damage accumulation in Cu and Fe“. In: *Journal of Nuclear Materials* 276.1 (2000), pp. 13–21 (cit. on p. 67).

[Cer01] S. Cerrai. *Second Order PDE's in Finite and Infinite Dimension: A Probabilistic Approach*. Vol. 1762. Lecture Notes in Mathematics. Springer-Verlag Berlin Heidelberg, 2001 (cit. on pp. 33, 57, 58).

[Coh+96] S.D. Cohen, A.C. Hindmarsh, and P.F. Dubois. „CVODE, a stiff/nonstiff ODE solver in C“. In: *Computers in physics* 10.2 (1996), pp. 138–143 (cit. on p. 91).

[DG+16] G. Di Gesu, T. Lelievre, D. Le Peutrec, and B. Nectoux. „Jump Markov models and transition state theory: the quasi-stationary distribution approach“. In: *Faraday Discussions* 195 (2016), pp. 469–495 (cit. on p. 4).

[DLR+97] T.D. De La Rubia, N. Soneda, M.J. Caturla, and E.A. Alonso. „Defect production and annealing kinetics in elemental metals and semiconductors“. In: *Journal of Nuclear Materials* 251 (1997), pp. 13–33 (cit. on p. 9).

[Dom+04] C. Domain, C.S. Becquart, and L. Malerba. „Simulation of radiation damage in Fe alloys: an Object kinetic Monte Carlo approach“. In: *Journal of Nuclear Materials* 335.1 (2004), pp. 121–145 (cit. on pp. 6, 67).

[Dra03] S.S. Dragomir. *Some Gronwall Type Inequalities and Applications*. Nova Science Publishers New York, 2003 (cit. on p. 28).

[Dun+16] A. Dunn, R. Dingreville, E. Martínez, and L. Capolungo. „Synchronous parallel spatially resolved stochastic cluster dynamics“. In: *Computational Materials Science* 120 (2016), pp. 43–52 (cit. on p. 67).

[Dup+02] A. H. Duparc, C. Moingeon, N. Smetniansky-de Grande, and A. Barbu. „Microstructure modelling of ferritic alloys under high flux 1 MeV electron irradiations“. In: *Journal of Nuclear Materials* 302.2 (2002), pp. 143–155 (cit. on p. 67).

[Eva98] W.D. Evans. *Partial Differential Equations*. Vol. 19. Graduate Studies in Mathematics. American Mathematical Society, 1998 (cit. on p. 27).

[Fri08] A. Friedman. *Partial Differential Equations of Parabolic Type*. Courier Dover Publications, 2008 (cit. on p. 27).

[Fri12] A. Friedman. *Stochastic Differential Equations and Applications*. Courier Corporation, 2012 (cit. on p. 32).

[GB00] M.A. Gibson and J. Bruck. „Efficient exact stochastic simulation of chemical systems with many species and many channels“. In: *The Journal of Physical Chemistry A* 104.9 (2000), pp. 1876–1889 (cit. on p. 7).

[Ghe+12] M. Gherardi, T. Jourdan, S. Le Bourdiec, and G. Bencteux. „Hybrid deterministic/stochastic algorithm for large sets of rate equations“. In: *Computer Physics Communications* 183.9 (2012), pp. 1966–1973 (cit. on pp. 14, 67, 68).

[Gho99] N.M. Ghoniem. „Clustering theory of atomic defects“. In: *Radiation Effects and Defects in Solids* 148.1-4 (1999), pp. 269–318 (cit. on p. 12).

[Gil00] D.T. Gillespie. „The chemical Langevin equation“. In: *Journal of Chemical Physics* 113.1 (2000), pp. 297–306 (cit. on p. 67).

[Gil07] D.T. Gillespie. „Stochastic simulation of chemical kinetics“. In: *Annual Review of Physical Chemistry* 58 (2007), pp. 35–55 (cit. on pp. 7, 8).

[Gil76] D.T. Gillespie. „A general method for numerically simulating the stochastic time evolution of coupled chemical reactions“. In: *Journal of computational physics* 22.4 (1976), pp. 403–434 (cit. on pp. 3, 6).

[Gil92] D.T. Gillespie. „A rigorous derivation of the chemical master equation“. In: *Physica A: Statistical Mechanics and its Applications* 188.1-3 (1992), pp. 404–425 (cit. on p. 7).

[Gol+01] S.I. Golubov, A.M. Ovcharenko, A.V. Barashev, and B.N. Singh. „Grouping method for the approximate solution of a kinetic equation describing the evolution of point-defect clusters“. In: *Philosophical Magazine A* 81.3 (2001), pp. 643–658 (cit. on pp. 11, 67).

[Goo64] F.C. Goodrich. „Nucleation rates and the kinetics of particle growth. II. The birth and death process“. In: *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*. Vol. 277. 1369. The Royal Society. 1964, pp. 167–182 (cit. on pp. 3, 12, 20, 25, 74, 77).

[GS80] N.M. Ghoniem and S. Sharafat. „A numerical solution to the Fokker-Planck equation describing the evolution of the interstitial loop microstructure during irradiation“. In: *Journal of Nuclear Materials* 92.1 (1980), pp. 121–135 (cit. on pp. 11, 12, 19, 22).

[Hai11] M. Hairer. „On Malliavin's proof of Hörmander's theorem“. In: *Bulletin des sciences mathématiques* 135.6-7 (2011), pp. 650–666 (cit. on pp. 32, 55).

[Har02] P. Hartman. *Ordinary Differential Equations*. Vol. 38. Classics in Applied Mathematics. SIAM edition, 2002 (cit. on p. 40).

[Hin+05] A.C. Hindmarsh, P.N. Brown, K.E. Grant, et al. „SUNDIALS: Suite of nonlinear and differential/algebraic equation solvers“. In: *ACM Transactions on Mathematical Software (TOMS)* 31.3 (2005), pp. 363–396 (cit. on p. 91).

[HY17] E. Hingant and R. Yvinec. „Deterministic and Stochastic Becker–Döring Equations: Past and Recent Mathematical Developments“. In: *Stochastic Processes, Multiscale Modeling, and Numerical Methods for Computational Cellular Biology*. Springer, 2017, pp. 175–204 (cit. on pp. 10, 11).

[Isi25] E. Ising. „Beitrag zur Theorie des Ferromagnetismus“. In: *Zeitschrift für Physik* 31.1 (1925), pp. 253–258 (cit. on pp. 4, 19).

[JC12] T. Jourdan and J.-P. Crocombette. „Rate theory cluster dynamics simulations including spatial correlations within displacement cascades“. In: *Physical Review B* 86.5 (2012), p. 054113 (cit. on p. 9).

[JC18] T. Jourdan and J.-P. Crocombette. „On the transfer of cascades from primary damage codes to rate equation cluster dynamics and its relation to experiments“. In: *Computational Materials Science* 145 (2018), pp. 235–243 (cit. on p. 99).

[Joh02] W.P. Johnson. „The curious history of Faà di Bruno's formula“. In: *The American Mathematical Monthly* 109.3 (2002), pp. 217–234 (cit. on p. 57).

[Jou+14] T. Jourdan, G. Bencteux, and G. Adjanor. „Efficient simulation of kinetics of radiation induced defects: a cluster dynamics approach“. In: *Journal of Nuclear Materials* 444.1 (2014), pp. 298–313 (cit. on pp. 12, 16, 17, 19, 20, 22, 25, 67, 69, 89, 91, 95, 97, 101, 105).

[Jou+16] T. Jourdan, G. Stoltz, F. Legoll, and L. Monasse. „An accurate scheme to solve cluster dynamics equations using a Fokker–Planck approach“. In: *Computer Physics Communications* 207 (2016), pp. 170–178 (cit. on pp. 12, 13, 20, 25, 67, 77).

[Kir73] M. Kiritani. „Analysis of the clustering process of supersaturated lattice vacancies“. In: *Journal of the Physical Society of Japan* 35.1 (1973), pp. 95–107 (cit. on pp. 11, 67).

[Kum+92] S. Kumar, S.K. Kurtz, J.R. Banavar, and M.G. Sharma. „Properties of a three-dimensional Poisson-Voronoi tesselation: A Monte Carlo study“. In: *Journal of Statistical Physics* 67.3-4 (1992), pp. 523–551 (cit. on p. 100).

[KW16] A.A. Kohnert and B.D. Wirth. „Grouping techniques for large-scale cluster dynamics simulations of reaction diffusion processes“. In: *Modelling and Simulation in Materials Science and Engineering* 25.1 (2016), p. 015008 (cit. on pp. 11, 15, 105).

[Lan74] J.-M. Lanore. „Simulation de l'évolution des défauts dans un réseau par la méthode de Monte-Carlo“. In: *Radiation Effects and Defects in Solids* 22.3 (1974), pp. 153–162 (cit. on pp. 3, 5, 67, 88).

[Laz+13] E.A. Lazar, J.K. Mason, R.D. MacPherson, and D.J. Srolovitz. „Statistical topology of three-dimensional Poisson-Voronoi cells and cell boundary networks“. In: *Physical Review E* 88.6 (2013), p. 063309 (cit. on p. 100).

[LBL08] C. Le Bris and P.-L. Lions. „Existence and Uniqueness of Solutions to Fokker–Planck Type Equations with Irregular Coefficients“. In: *Communications in Partial Differential Equations* 33.7 (2008), pp. 1272–1317 (cit. on pp. 27, 56).

[LM02] P. Laurençot and S. Mischler. „From the Becker–Döring to the Lifshitz–Slyozov–Wagner Equations“. In: *Journal of Statistical Physics* 106.5 (2002), pp. 957–991 (cit. on pp. 11, 13, 22).

[LP06] H. Luschgy and G. Pagès. „Functional quantization of a class of Brownian diffusions: a constructive approach“. In: *Stochastic Processes and their Applications* 116.2 (2006), pp. 310–336 (cit. on p. 56).

[LR91] F. Leyvraz and S. Redner. „Spatial organization in the two-species annihilation reaction A+B→0“. In: *Physical Review Letter* 66 (16 1991), pp. 2168–2171 (cit. on p. 7).

[MA97] H.H. McAdams and A. Arkin. „Stochastic mechanisms in gene expression“. In: *Proceedings of the National Academy of Sciences* 94.3 (1997), pp. 814–819 (cit. on p. 7).

[MB11] J. Marian and V.V. Bulatov. „Stochastic cluster dynamics method for simulations of multispecies irradiation damage accumulation“. In: *Journal of Nuclear Materials* 415.1 (2011), pp. 84–95 (cit. on pp. 7, 12, 16, 67).

[McQ67] D.A. McQuarrie. „Stochastic approach to chemical kinetics“. In: *Journal of applied probability* 4.3 (1967), pp. 413–478 (cit. on p. 6).

[Met+53] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. „Equation of State Calculations by Fast Computing Machines“. In: *Journal of Chemical Physics* 21.6 (1953), pp. 1087–1092 (cit. on pp. 5, 78).

[MI07] S. Malefaki and G. Iliopoulos. „Short Communication: Simulating from a Multinomial Distribution with Large Number of Categories“. In: *Computational Statistics & Data Analysis* 51.12 (Aug. 2007), pp. 5471–5476 (cit. on p. 75).

[Nie03] B. Niethammer. „On the Evolution of Large Clusters in the Becker–Döring Model.“ In: *Journal of Nonlinear Science* 13.1 (2003), pp. 115–122 (cit. on p. 13).

[Nie04] B. Niethammer. „A scaling limit of the Becker–Döring equations in the regime of small excess density“. In: *Journal of Nonlinear Science* 14.5 (2004), pp. 453–468 (cit. on p. 13).

[Nor+98] K. Nordlund, M. Ghaly, R.S. Averback, et al. „Defect production in collision cascades in elemental semiconductors and fcc metals“. In: *Physical Review B* 57.13 (1998), p. 7556 (cit. on p. 9).

[Nor97] J.R. Norris. *Markov Chains*. Vol. 2. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1997 (cit. on pp. 5, 54).

[Nua06] D. Nualart. *The Malliavin Calculus and Related Topics*. Vol. 1995. Probability and Its Applications. Springer-Verlag Berlin Heidelberg, 2006 (cit. on p. 33).

[Ort+07] C.J. Ortiz, M.J. Caturla, C.-C. Fu, and F. Willaime. „He diffusion in irradiated $\alpha$-Fe: An ab-initio-based rate theory model“. In: *Physical Review B* 75.10 (2007), p. 100102 (cit. on p. 67).

[Ovc+03] A.M. Ovcharenko, S.I. Golubov, C.H. Woo, and H. Huang. „GMIC++: grouping method in C++: an efficient method to solve large number of master equations“. In: *Computer Physics Communications* 152.2 (2003), pp. 208–226 (cit. on pp. 9, 11, 22, 30, 68, 69, 83).

[Paz12] A. Pazy. *Semigroups of Linear Operators and Applications to Partial Differential Equations*. Vol. 44. Applied Mathematical Science. Springer, New York, NY, 2012 (cit. on pp. 21, 37, 45, 46).

[Pen+78] O. Penrose, J. L. Lebowitz, J. Marro, M. H. Kalos, and A. Sur. „Growth of clusters in a first-order phase transition“. In: *Journal of Statistical Physics* 19.3 (1978), pp. 243–267 (cit. on p. 13).

[Pen97] O. Penrose. „The Becker–Döring equations at large times and their connection with the LSW theory of coarsening“. In: *Journal of statistical physics* 89.1-2 (1997), pp. 305–320 (cit. on p. 13).

[Pro13] P. Protter. *Stochastic Integration and Differential Equations*. Vol. 21. Stochastic Modelling and Applied Probability. Springer-Verlag Berlin Heidelberg, 2013 (cit. on pp. 57, 62).

[RC13] C. Robert and G. Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013 (cit. on p. 78).

[Red01] S. Redner. *A Guide to First-Passage Processes*. Cambridge University Press, 2001 (cit. on p. 4).

[RT74] M.T. Robinson and I.M. Torrens. „Computer simulation of atomic-displacement cascades in solids in the binary-collision approximation“. In: *Physical Review B* 9.12 (1974), p. 5008 (cit. on p. 9).

[Sam+05] M. Samoilov, S. Plyasunov, and A. Arkin. „Stochastic amplification and signaling in enzymatic futile cycles through noise-induced bistability with oscillations“. In: *Proceedings of the National Academy of Sciences of the United States of America* 102.7 (2005), pp. 2310–2315 (cit. on p. 7).

[Sar03] S.A. Sarra. „The method of characteristics with applications to conservation laws“. In: *Journal of Online Mathematics and its Applications* 3 (2003), pp. 1–16 (cit. on p. 28).

[Sch91] L. Schwartz. *Analyse I. Théorie des Ensembles et Topologie*. Vol. 42. Enseignement des sciences. Hermann, 1991 (cit. on p. 41).

[Soi+10] F. Soisson, C. Becquart, N. Castin, et al. „Atomistic Kinetic Monte Carlo studies of microchemical evolutions driven by diffusion processes under irradiation“. In: *Journal of Nuclear Materials* 406.1 (2010), pp. 55–67 (cit. on pp. 3, 19).

[Ste94] W.J. Stewart. *Introduction to the numerical solutions of Markov chains*. Princeton University Press, 1994 (cit. on p. 75).

[Sur+04] M.P. Surh, J.B. Sturgeon, and W.G. Wolfer. „Master equation and Fokker–Planck methods for void nucleation and growth in irradiation swelling“. In: *Journal of Nuclear Materials* 325.1 (2004), pp. 44–52 (cit. on pp. 14, 67–69, 80, 88).

[Ter+17] P. Terrier, M. Athènes, T. Jourdan, G. Adjanor, and G. Stoltz. „Cluster dynamics modelling of materials: A new hybrid deterministic/stochastic coupling approach“. In: *Journal of Computational Physics* 350 (2017), pp. 280–295 (cit. on pp. 20, 24, 25, 39, 47, 67).

[Tro59] H.F. Trotter. „On the product of semi-groups of operators“. In: *Proceedings of the American Mathematical Society* 10.4 (1959), pp. 545–551 (cit. on p. 24).

[Vas+02] A. Vasseur, F. Poupaud, J.-F. Collet, and T. Goudon. „The Becker–Döring system and its Lifshitz–Slyozov limit“. In: *SIAM Journal on Applied Mathematics* 62.5 (2002), pp. 1488–1500 (cit. on p. 13).

[VK92] N.G. Van Kampen. *Stochastic Processes in Physics and Chemistry*. Vol. 1. Elsevier, 1992 (cit. on p. 5).

[Vot07] A. Voter. „Introduction to the kinetic Monte Carlo method“. In: *Radiation effects in solids*. NATO Science Series II: Mathematics, Physics and Chemistry. Springer, 2007, pp. 1–23 (cit. on pp. 3, 6, 19, 67).

[Vot86] A. Voter. „Classically exact overlayer dynamics: Diffusion of rhodium clusters on Rh (100)“. In: *Physical review B* 34.10 (1986), p. 6819 (cit. on p. 5).

[Wol+77] W.G. Wolfer, L.K. Mansur, and J.A. Sprague. *Theory of swelling and irradiation creep*. Tech. rep. Wisconsin Univ., 1977 (cit. on pp. 3, 12, 25, 67, 77).

[YE66] W.M. Young and E.W. Elcock. „Monte Carlo studies of vacancy migration in binary ordered alloys: I“. In: *Proceedings of the Physical Society* 89.3 (1966), p. 735 (cit. on pp. 3, 19).

[Zei95] E. Zeidler. *Applied Functional Analysis: Main Principles and Their Applications*. Vol. 109. Applied Mathematical Sciences. Springer-Verlag New York, 1995 (cit. on p. 39).