Mathematical models and numerical simulation of photovoltaic devices

Athmane Bakhta

UNIVERSITÉ PARIS-EST

l'École Doctorale Mathématiques et Sciences et Technologies de l'Information et de la Communication (MSTIC)

Thèse de doctorat

Discipline : Mathématiques

présentée par

Athmane Bakhta

Modèles mathématiques et simulation numérique de dispositifs photovoltaïques

---

Thèse préparée au CERMICS, École des Ponts ParisTech

Soutenue le 19 décembre 2017 devant le Jury composé de :

Éric Cancès	École des Ponts ParisTech	Directeur de thèse
Tony Lelièvre	École des Ponts ParisTech	Codirecteur de thèse
Thomas Lepoutre	INRIA Rhône-Alpes	Rapporteur
Julien Salomon	INRIA Paris	Rapporteur
Virginie Ehrlacher	École des Ponts ParisTech	Examinatrice
Ayman Moussa	Université Pierre et Marie Curie	Examinateur
Gabriel Turinici	Université Paris-Dauphine	Président du jury
Julien Vidal	EDF R&D	Examinateur

À l'âme de ma grand-mère Baya,
À ma mère Nouara et à mon père Malek,
À mes frères et à ma petite soeur.

REMERCIEMENTS

Ces quelques lignes comportent une dimension très importante dans une thèse mais -tradition oblige- complètement absente des mansucrits, à savoir la dimension humaine. Par ailleurs, il est statistiquement observé que les pages les plus parcourues d'un manuscrit sont les pages de *remerciements*. Il convient donc de bien les soigner.

C'est Tony Lelièvre (mon codirecteur de thèse et mon ancien professeur) qui a initié ce travail et qui m'a fait découvrir le "merveilleux" monde de la recherche scientifique. Je lui exprime toute ma gratitude et mon profond respect comme je le remercie d'avoir joué un rôle crucial dans mon orientation vers les mathématiques appliquées.

Je ne remercierai jamais assez Eric Cancès qui fut un excèllent directeur de thèse pour moi. Il m'a fait confiance, m'a guidé dans mon travail et m'a beaucoup appris (en maths, en physique, en chimie, ...). Chaque réunion avec Eric est une dose de motivation (à consommer sans modération). Ses qualités humaines n'ont d'égal que ses compétences scientifiques. Je témoigne sincèrement à Eric mon admiration et ma gratitude. En dehors des mathématiques, j'ai eu beaucoup de plaisir à partager des moments conviviaux avec Eric (à Paris, à Roscoff, à Minneapolis, à Atlanta,...). Je le remercie d'ailleurs de m'avoir fait découvrir cet excèllent bain de bouche que les américains appellent *root beer*.

J'ai eu un énorme plaisir à travailler avec Virginie Ehrlacher qui a joué un grand rôle dans l'encadrement de ma thèse. J'ai beaucoup appris à ses côtés et je suis très fier d'être son (premier) élève. Virginie est elle même une ancienne étudiante de Tony et d'Eric, je suis donc doublement fier car en plus d'être mon encadrante, elle est aussi ma grande soeur de thèse. J'espère pouvoir continuer à collaborer avec elle dans le futur.

Je tiens sincèrement à remericer les membres du jury à commencer par les rapporteurs Thomas Lepoutre et Julien Salomon qui ont lu minutieusement ce manuscrit et ont contribué à sa qualité. L'intérêt qu'ils ont manifesté pour mon travail ne peut que m'encourager à faire encore mieux. Un grand merci aussi à Ayman Moussa, Gabriel Turinici et Julien Vidal. Je suis fier et honoré d'avoir vos noms sur la page de garde de ma thèse.

Telle une mère attentionnée, Isabelle Simunic veille à ce qu'on ne manque de rien au CERMICS. Elle m'a beaucoup aidé sur les aspects administratifs (et pas que pour la thèse d'ailleurs). Je peux témoigner -sans l'ombre d'un doute- qu'Isabelle est la meilleure secrétaire de labo au monde. Je n'oublie évidemment pas Fatna Baoudj qui, en plus

d'être efficace, est très gentille ! Franchement, que feraient les chercheurs du CERMICS sans vous ? ! Un énomre merci à toutes les deux.

Puisque la thèse est une aventure individuelle qui se vit en groupe, je tiens à remercier tous les doctorants, post-doc, chercheurs,... avec qui j'ai pu partager un bon moment au CERMICS. Je voudrais d'abord remercier mon grand frère de thèse David Gontier. Je suis heureux d'avoir pu publier un article avec lui. Je remercie également Antoine Levitt et Julien Reygner, deux autres exemples pour moi. Ensuite, la liste est longue et certainement pas exhaustive : Yannick Masson, Boris Nectoux, Simon Lemaire, Tom Hudson, Françoins Madiot, Anis Al Gerbi, Richard Fischer, Laurent Daudet, Houssam Alrachid, Nahia Mourad, Marc Josien, Julien Roussel, Grégoire Ferré, Pierre Loïc Rothé, Gérôme Faure, Laura Silva Lopes, Marion Sciauveau, Henri Gerard, Alexandre Zhou, Etienne de Saint Germain, Oumaïma Bencheikh, Amina Benaceur, LingLing Cao, Karol Cascavita, Gustave Emprin, Florent Edin, Adel Cherchali, Mouad Ramil, Sami Siraj-Dine, Raphaël Coyaud, Adrien Lesage, Fréderic marazzato,... Je pourrais raconter une anectode sur chacun mais je garde cela en off[1]

L'aventure s'étend (fort heureusement) au delà du labo. Je souhaiterais donc remercier Damiano Lombardi pour l'ensemble des discussions sur l'estimateur à postériori. J'espère que cette collaboration portera d'autres fruits dans le futur. Je m'estime très chanceux d'avoir connu Yvon Maday, un des chercheurs[2] les plus dynamiques et productifs que je connaisse. En m'offrant un poste d'ingénieur de recherche à l'institut Carnot, Yvon Maday me témoigne d'une grande confiance dont je suis particulièrement fier. Geneviève Dusson était comme une soeur de thèse pour moi. Il semblerait qu'on ait hérité de l'amitié qui lie nos directeurs Eric et Yvon. J'ai eu beaucoup de plaisir à co-organiser le groupe de travail *"chimie quantitique, de la théorie à la pratique"* avec Geneviève. Mais, j'ai surtout eu beaucoup de plaisir à partager des voyages, des conférences, des repas... et la fameuse *root beer* avec elle. Mon séjour à l'université du Minnesota était très agréable grâce à Paul Cazeaux, Matthias Maier, Daniel Massatt, Andrew Stuart et le professeur Mitchell Luskin. Thank you guys ! It has been a pleasure. Je remercie également les physiciens de Harvard (Shiang Fang, Stephen Carr et le professeur Efthimios Kaxiras) pour l'ensemble des discussions autour des fonctions de Wannier. J'ai rencontré les professeurs Farida et Mohamed Cheriet à la fin de ma thèse[3] et une mystérieuse énergie s'est dégagée de cette rencontre. Je les remercie sincèrement pour leurs encouragements et leur exprime mon profond respect.

J'en arrive aux autres amis qui ne sont pas forcèment mathématiciens. Leur présence me rappelle que la vie ne se résume pas à la science. Je commence d'abord par mes amis d'enfance Farid, Zinou et Sidali. Ils sont un parfait contre-exemple à la phrase "loin des yeux, loin du coeur". Notre amitié reste éternelle malgré la distance. Ensuite il y a les amis de l'UPEC : Johann Nicod, Joelle Faure, Michelle Senn, Elizabeth Da Silva, Mounia Afkir, Marion Grandamy, Emilie Nègre... la liste est très longue ! Je les remercie tous pour m'avoir outenu et encouragé. Je remercie également les membres du bureau du REDOC-Paris-Est[4] et l'ensemble de nos adhérents.

J'en viens maintenant aux amis musiciens, ceux-là sont très importants dans ma vie.

[1] Ce qui se passe au labo reste au labo !
[2] pour ne pas dire LE chercheur
[3] qui n'est clairement pas le meilleur moment de l'aventure !
[4] en particulier, l'irremplaçable Philippe Gambette

En effet, en plus de me rappeler que la vie ne se résume pas aux maths, ils me délivrent de la routine quotidienne pour m'aider à transcender la réalité[5].

Mes remerciements les plus profonds s'adressent à ma famille. En commençant par mon frère aîné Hamza. Il a toujours cru en moi depuis mon jeune âge et m'a toujours encouragé à aller de l'avant. Je lui dois une reconnssaince éternelle ainsi qu'à ma belle soeur Naima. Mes autres frères Farid, Yacine, Belkacem et ma petite soeur Nassima méritent toute ma gratitude. Je les remercie infiniment pour leur soutien inconditionnel. Ce qui nous unit dépasse les liens du sang car les frères sont (aussi) des amis donnés par la nature[6]. Ma grand-mère Baya m'a vu commencer cette thèse mais nous a quitté depuis, je lui dédie ce manuscrit.

Enfin, du fond du coeur, merci maman et papa. Merci pour vos sacrifices, votre éducation, vos principes, votre dévouement et votre soutien inconditionnel. Vous rendre fiers est plus qu'un objectif pour moi, c'est un devoir. J'espère que ce résultat va rendre ne serait-ce qu'une goutte de l'océan de tout ce que vous m'avez donné.

[5] Les initiés à Amar Ezzahi comprendront facilement
[6] Citation de Plutarque

Résumé

Cette thèse comporte deux volets indépendants mais tous deux motivés par la modélisation mathématique et la simulation numérique de procédés photovoltaïques.

La **Partie I** traite de systèmes d'équations aux dérivées partielles de diffusion croisée, modélisant l'évolution de concentrations ou de fractions volumiques de plusieurs espèces chimiques ou biologiques. Nous présentons dans le **chapitre 1** une introduction succincte aux résultats mathématiques connus sur ces systèmes losqu'ils sont définis sur des domaines fixes. Nous présentons dans le **chapitre 2** un système uni-dimensionnel que nous avons introduit pour modéliser l'évolution des fractions volumiques des différentes espèces chimiques intervenant dans le procédé de déposition physique en phase vapeur (PVD) utilisé pour la fabrication de cellules solaires à couches minces. Dans ce procédé, un échantillon est introduit dans un four à très haute température où sont injectées les différentes espèces chimiques sous forme gazeuse, si bien que des atomes se déposent petit à petit sur l'échantillon, formant une couche mince qui grandit au fur et à mesure du procédé. Dans ce modèle sont pris en compte à la fois l'évolution de la surface du film solide au cours du procédé et l'évolution des fractions volumiques locales au sein de ce film, ce qui aboutit à un système de diffusion croisée défini sur un domaine dépendant du temps. En utilisant une méthode récente basée sur l'entropie, nous montrons l'existence de solutions faibles à ce système et nous étudions leur comportement asymptotique dans le cas où les flux extérieurs imposés à la surface du film sont supposés constants. De plus, nous prouvons l'existence d'une solution à un problème d'optimisation sur les flux extérieurs. Nous présentons dans le **chapitre 3** comment ce modèle a été adapté et calibré sur des données expérimentales.

La **Partie II** est consacrée à des questions reliées au calcul de la structure électronique de matériaux cristallins. Nous rappelons dans le **chapitre 4** certains résultats classiques relatifs à la décomposition spéctrale d'opérateurs de Schrödinger périodiques. Dans le **chapitre 5**, nous tentons de répondre à la question suivante : est-il possible de déterminer un potentiel périodique tel que les premières bandes d'énergie de l'opérateur de Schrödinger associé soient aussi proches que possible de certaines fonctions cibles ? Nous montrons théoriquement que la réponse à cette question est positive lorsque l'on considère la première bande de l'opérateur et des potentiels uni-dimensionnels appartenant à un espace de mesures périodiques bornées inférieurement en un certain sens. Nous proposons également une méthode adaptative pour accélérer la procédure numérique de résolution du problème d'optimisation. Enfin, le **chapitre 6** traite d'un algorithme glouton pour la compression de fonctions de Wannier en exploitant leurs symétries. Cette compression permet, entre autres, d'obtenir des expressions analytiques pour certains coefficients de tight-binding intervenant dans la modélisation de matériaux 2D.

Abstract

This thesis includes two independent parts, both motivated by mathematical modeling and numerical simulation of photovoltaic devices.

Part I deals with cross-diffusion systems of partial differential equations, modeling the evolution of concentrations or volume fractions of several chemical or biological species. We present in **Chapter 1** a succinct introduction to the existing mathematical results about these systems when they are defined on fixed domains. We present in **Chapter 2** a one-dimensional system that we introduced to model the evolution of the volume fractions of the different chemical species involved in the physical vapor deposition process (PVD) used in the production of thin film solar cells. In this process, a sample is introduced into a very high temperature oven where the different chemical species are injected in gaseous form, so that atoms are gradually deposited on the sample, forming a growing thin film. In this model, both the evolution of the film surface during the process and the evolution of the local volume fractions within this film are taken into account, resulting in a cross-diffusion system defined on a time-dependent domain. Using a recent method based on entropy estimates, we show the existence of weak solutions to this system and study their asymptotic behavior when the external fluxes are assumed to be constant. Moreover, we prove the existence of a solution to an optimization problem set on the external fluxes. We present in **Chapter 3** how was this model adapted and calibrated on experimental data.

Part II is devoted to some issues related to the calculation of the electronic structure of crystalline materials. We recall in **Chapter 4** some classical results about the spectral decomposition of periodic Schrödinger operators. In **Chapter 5**, we try to answer the following question: is it possible to determine a periodic potential such that the first energy bands of the associated periodic Schrödinger operator are as close as possible to certain target functions? We theoretically show that the answer to this question is positive when we consider the first energy band of the operator and one-dimensional potentials belonging to a space of periodic measures that are lower bounded in a certain sens. We also propose an adaptive method to accelerate the numerical optimization procedure. Finally, **Chapter 6** deals with a greedy algorithm for the compression of Wannier functions into Gaussian-polynomial functions exploiting their symmetries. This compression allows, among other things, to obtain closed expressions for certain tight-binding coefficients involved in the modeling of 2D materials.

List of publications

Here is a list of articles that were written during this thesis:

- [BE16] (with Virginie Ehrlacher[7]) *Cross diffusion equations with non-zero flux and moving boundary conditions.* (accepted for publication in M2AN).

- [BEG17] (with Virginie Ehrlacher and David Gontier[8]) *Numerical reconstruction of the first band(s) in an inverse Hill's problem* (submitted).

- [BL17] (with Damiano Lombardi[9]) *An a posteriori error estimator based on shifts for positive hermitian eigenvalue problems* (submitted).

- [BCC+17] (with Eric Cancès[10], Paul Cazeaux[11], Shiang Fang[12] and Efthimios Kaxiras[13]) *Compression of Wannier functions into Gaussian-type orbitals* (submitted).

[7] Université Paris-Est, CERMICS, Ecole des Ponts ParisTech, France
[8] Université Paris-Dauphine, CEREMADE, France
[9] INIRA-Paris, France
[10] Université Paris-Est, CERMICS, Ecole des Ponts ParisTech, France
[11] Department of Mathematics, University of Kansas, Lawrence, Kansas 66045-7594, USA
[12] Department of Physics, Harvard University, Cambridge, Massachusetts 02138, USA
[13] Department of Physics, Harvard University, Cambridge, Massachusetts 02138, USA

CONTENTS

GENERAL INTRODUCTION

Contents

Solar Cells

A solar cell converts solar energy into an electric current, using semiconducting materials. The efficiency of a solar cell therefore relies on the electronic properties of the semiconducting material. A semiconductor is characterized by a band gap : the difference between the energy of the conduction band and the energy of the valence band. Two types of semiconducting materials can be distinguished: semiconductors of type p, which contain acceptor-type defects leading to the creation of an excess of holes in the valence band; and semiconductors of type n, which are doped with donor defects, leading to to the creation of an excess of electrons in the conduction band.

Most photovoltaic (PV) cells consist in a p-type layer (which will be in contact with the light source) on top of a n-type layer leading to a p-n junction. Thus, the excess holes (positively charged) of the p-type layer and the excess electrons (negatively charged) of the n-type layer are attracted to each other. The electronic movement results in an electric field and forms a depletion zone between the two layers. This region plays the role of a barrier and prevents the electrons and holes from recombining. When the sunlight strikes the cell, the photons excite the electrons on the n-type top layer. Therefore, the electrons leave their original state and become mobile and extra holes are created. Because of the electric field, the mobile electrons stay in the n-layer but the holes move to the p-layer. As a consequence, the n-layer contains an extra negative charge and the p-layer an extra positive charge. Finally, the electrical current is obtained by connecting the two sides with a circuit. A schematic representation of the working principle of a solar cell is given in Figure 1. The reader may refer to [PU+16] for further details on the physics of solar cells.

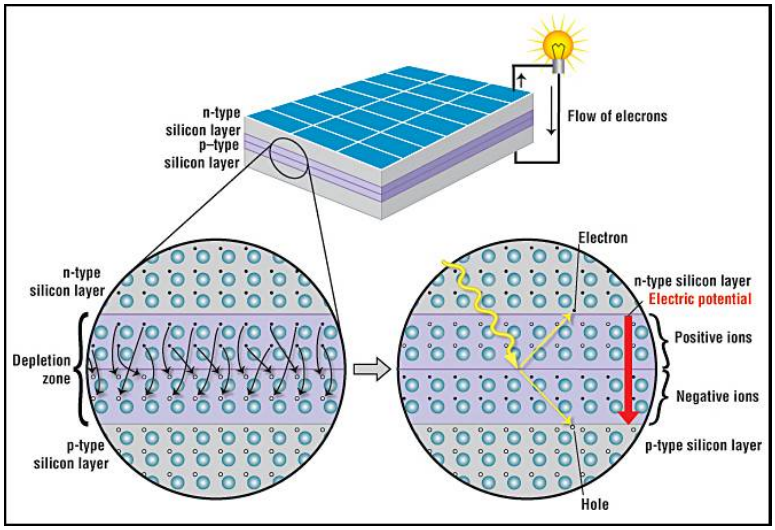


Figure 1 – Schematic representation of a solar cell, showing the n-type and p-type layers, with a close-up view of the depletion zone around the junction between the n-type and p-type layers. Source : Online article of the American Chemical Society: How a Solar Cell Works. https://www.acs.org/content/acs/en/education/resources/

The most commercially available PV technologies are: the ones based on crystalline or multi-crystalline silicon technologies (c-Si) and the ones using thin film technologies (among which the cadmium-telluride (CdTe), amorphous silicon (a-Si) and copper indium gallium diselenide (CIGS) modules). The CIGS-based cells are less efficient than the c-Si cells. However, this technology still presents several competitive advantages : a lower production cost, a lower ecological footprint and a better adaptability to lightweight and flexible substrates [Mol16]. This motivate the many recent efforts for the study and development of CIGS based solar cells [Kli15, Mol16, PWJ+14, JHW+15, JWH+16].

The standard CIGS solar cell structure is shown in Figure 2. The cell is basically composed of a p-type Cu(In,Ga)Se_2 layer, which acts as the main light absorber, in contact with a n-type CdS layer to form a p-n junction. At the frontside, a transparent electrode generally based on a ZnO/ZnO:Al bilayer, collects the electrons. At the rearside, a molybdenum electrode collects the holes. The roles and properties of each layer are discussed in [Kli15, Mol16]. Let us focus here in the CIGS absorber layer, which is the object of interest in this thesis. The Cu(In,Ga)Se_2 material is a semiconductor material with a tetragonal chalcopyrite crystalline structure (see Figure 3).

Two main methods are used for the production of the CIGS layer: the selenization of vacuum-deposited metallic precursors and co-evaporation using the Physical Vapor deposition (PVD) [Kli15, Mat10]. In the co-evaporation approach, the four constituents of the absorber layer are simultaneously evaporated in a high temperature vacuum chamber. As the injected atoms deposit on the substrate, an heterogeneous solid grows upon it forming thus the CIGS layer. Different evaporation senarii, distinguished by different evaporation rates and substrate temperatures, have been developed. The three-stages process (schematically illustrated in Figure 4) allows one to achieve very high cell efficiencies 20% [Kli15].

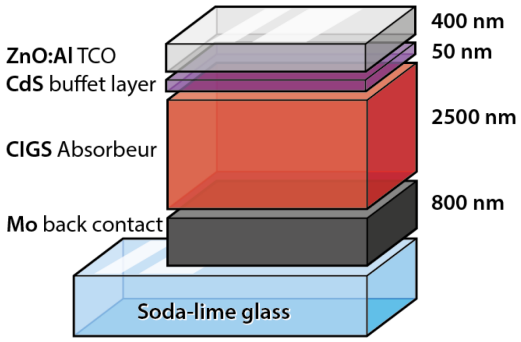


Figure 2 – Typical composition of a standard CIGS based solar cell. Source: [Mol16]

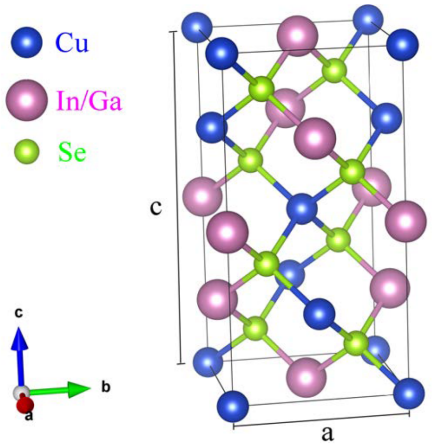


Figure 3 – Unit cell of chalcopyrite Cu(In,Ga)Se2. Source: [Kli15]

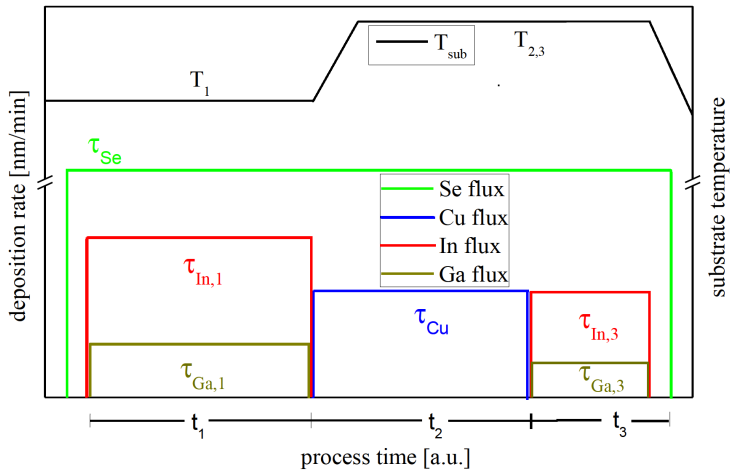


Figure 4 – Evaporation and temperature profile for the 3-stage process. The deposition rates $\tau_{\mathbf{a},i}$ for every component $\mathbf{a} = Cu, In, Ga, Se$ and the temperature of the substrate T_i are given for the three regimes $i = 1, 2, 3$. Source: [Kli15]

Outline of the Thesis

This thesis was originally motivated by a collaboration between the CERMICS lab[14] and the IRDEP lab[15] aiming to present mathematical approaches for the optimization of the

[14] CERMICS is the research center in applied mathematics at Ecole des Ponts ParisTech, France

photovoltaic efficiency of CIGS based solar cells. The last part of the thesis is motivated by a collaboration with physicists from the physics department of Harvard University working on the study of the electronic properties of heterogeneous 2D materials.

This manuscript is organized in two parts. The first part concerns cross-diffusion systems and the second part treats some issues related to electronic structure calculations.

The first contribution of the present thesis concerns the optimization of the co-evaporation production process of the CIGS layer. We propose a one-dimensional mathematical model for the Physical Vapor Deposition process in which two main phenomena are taken into account: the evolution of the surface of the layer and the diffusion of the various species in the bulk, due to the high temperature of the chamber. The proposed model writes under the form of a system of cross-diffusion PDEs defined on a moving domain. We present in **Chapter 1** an introduction to the well-known results about classical cross-diffusion systems on fixed domains. **Chapter 2** gathers the results of [BE16] and is dedicated to the analysis of our proposed model. We show a global-in-time existence of weak solution to the system and investigate their long-time behavior in the case of constant external fluxes. We also formulate an optimization problem set on the external fluxes, for which we prove the existence of a solution. From a numerical point of view, we suggest a numerical scheme for the discretization of the model and present a gradient-based numerical procedure to solve the optimization problem. We finally present in **Chapter 3** some practical improvements of our model and its calibration on experimental measures.

In the second part of the thesis, we consider periodic Schrödinger operators of the form $A = -\Delta + V$ where V is a real-valued periodic potential. We briefly present in **Chapter 4** the standard mathematical tools used in electronic structure calculations. We introduce in particular the Bloch-Floquet transform that allows one to characterize the spectrum of A as the reunion of the spectra of a family of selfadjoint compact resolvent operators A_q indexed by an element $q \in \mathbb{R}^d$ called *quasi-momentum*. The function that maps $q \in \mathbb{R}^d$ with the m^{th} eigenvalue of A_q is the so-called m^{th} energy band associed to A. Then, we focus on the following question: is it possible to determine a periodic potential V such that the lowest energy bands associated to the periodic Schrödinger operator $A = -\Delta + V$ are close to some target functions? In **Chapter 4**, we gather the results of [BEG17] where we formulate the above question as an optimization problem set on the space of one-dimensional periodic potentials that are measure-valued and lower bounded in a certain sense. Moreover, we present an adaptive optimization method which is faster than the standard gradient-based procedures.

Lastly, we consider Wannier functions, which are localized-in-space functions constructed from the Bloch eigenstates of the periodic Schrödinger operator $A = -\Delta + V$. These functions are used in tight-binding models for heterogeneous 2D materials and thus play an essential role in the study of the electronic properties of such structures. In **Chapter 6**, we report present some results of [BCC+17] where we propose a greedy procedure for the compression of Wannier functions into symmetry-adapted Gaussian-polynomials functions. Such a compression has two advantages: i) it allows one to

[15] IRDEP (Institut de recherche et développement sur l'énergie photovoltaïque) is a research lab of Chimie ParisTech, CNRS, EDF R&D, France, working on the new generations of photovoltaic technologies.

characterize a Wannier function by a small number of parameters rather than by its values on a (possibly large) gird, ii) it allows one to accelerate the parametrization of tight-binding Hamiltonians since closed formulas can be obtained for the tight-binding matrix elements.

Part I

Cross-diffusion

CHAPTER 1

CROSS-DIFFUSION SYSTEMS ON FIXED DOMAINS

Cross-diffusion models are systems of Partial Differential Equations (PDEs) describing the time evolution of multicomponent systems. Such models arise naturally in biology, physics and chemistry. In Section 1.1, we give the general form of cross-diffusion systems considered in this thesis along with some classical examples. Some mathematical challenges arising from their analysis are commented in Section 1.2. Section 1.3 will be devoted to the entropy structure admitted by some cross-diffusion systems which is a key-ingredient in the proof of the existence of global-in-time solutions. Three main methods used in the literature to analyze cross-diffusion systems, namely *gradient flow theory*, *the boundedness-by-entropy* method and *the duality* method are discussed respectively in Section 1.3.1, Section 1.3.2 and Section 1.3.3. Remarks on the uniqueness of weak solutions to such systems are reported in Section 1.4. Section 1.5 is dedicated to the long time behavior of the solutions. The contributions of the present thesis to the study of cross-diffusion systems are summarized in Section 1.6.

Contents

1.1 General Form of Cross-diffusion Systems

Let $d \in \mathbb{N}^*$ and let $\Omega \subset \mathbb{R}^d$ be a bounded domain with smooth boundary $\partial\Omega$. We denote by $\mathbf{n}(x)$ the exterior unit normal vector at $x \in \partial\Omega$. Let $T > 0$ denote a final time. We are interested in the dynamics of a multicomponent systems evolving in the domain Ω during the time $[0, T]$. We consider two different cases, which we present separately, but which lead to similar PDE systems, namely the *non-volume filling* case and the *volume filling* case.

Non-volume Filling Case:

Let $n \in \mathbb{N}^*$ denote the number of components in the system and let $u_1, \cdots, u_n$ be real-valued functions defined on $[0, T] \times \Omega$ such that for all $1 \leq i \leq n$, $t \in [0, T]$, $x \in \Omega$, $u_i(t, x)$ describes the local concentration of the species i at time t and point x. In the sequel, we denote by $u := (u_1, \cdots, u_n)^T$ the vector-valued function defined on $[0, T] \times \Omega$. We assume that the evolution of u is ruled by a system of PDEs of the form

$$\partial_t u - \operatorname{div}(A(u)\nabla u) = f(u), \quad \text{for} \quad (t, x) \in [0, T] \times \Omega, \tag{1.1}$$

$$u(0, \cdot) = u^0 \quad \text{in } \Omega, \tag{1.2}$$

$$(A(u)\nabla u) \cdot \mathbf{n} = 0, \quad \text{on } [0, T] \times \partial\Omega, \tag{1.3}$$

where $A : \mathbb{R}^n \mapsto \mathbb{R}^{n \times n}$ is a matrix-valued application, $f : \mathbb{R}^n \to \mathbb{R}^n$ is a vector-valued application and the initial condition $u^0 : \Omega \to \mathbb{R}^n$ is a sufficiently smooth vector-valued function. A system of the form (1.1)-(1.2)-(1.3) is called hereafter a *cross-diffusion* system. The application A (respectively f) is called the *diffusion matrix* (respectively the *reaction term*). The boundary condition (1.3) is a *no-flux boundary condition* whose justification stems from the fact the system is assumed to be isolated.

For all $t \in [0, T]$, $x \in \Omega$ and $1 \leq i \leq n$, $u_i(t, x)$ represents the local concentration of the i^{th} species, it is naturally expected to be non-negative. Thus, the values of u are expected to lie in $\overline{\mathcal{D}}_{\text{non-vf}}$ where

$$\mathcal{D}_{\text{non-vf}} := \{z = (z_1, \cdots, z_n), \quad z_i > 0, \quad 1 \leq i \leq n\} = (\mathbb{R}_+^*)^n. \tag{1.4}$$

Volume Filling Case:

In some applications, the quantities of interest may be the volume fractions of the different components of the system. We refer to this situation as a *volume-filling case*. We assume here that the system is composed on $n + 1$ different species and denote respectively by $u_0(t, x), \cdots, u_n(t, x)$ their local volume fractions at time $t \in [0, T]$ and point $x \in \Omega$. The evolution of $u_0, \cdots, u_n$ can be modeled by a set of PDEs of the following form : for all $0 \leq i \leq n$,

$$\partial_t u_i - \operatorname{div}\left(\sum_{j=0}^{n} G_{ij}(u_0, \cdots, u_n)\nabla u_j\right) = g_i(u_0, \cdots, u_n), \tag{1.5}$$

where for all $0 \leq i \leq n$, $g_i : \mathbb{R}^{n+1} \to \mathbb{R}$ and for all $0 \leq j \leq n$, $G_{ij} : \mathbb{R}^{n+1} \to \mathbb{R}$, along with appropriate initial and no-flux boundary conditions.

As $u_0, \cdots, u_n$ represent the volume fractions of the different species, it is naturally expected that they satisfy the following constraints

$$\forall 0 \leq i \leq n, \quad 0 \leq u_i(t,x) \leq 1 \quad \text{and} \quad \sum_{i=0}^{n} u_i(t,x) = 1 \quad \forall (t,x) \in (0,T) \times \Omega, \tag{1.6}$$

In this case, the following hods

$$u_0(t,x) = 1 - \sum_{i=1}^{n} u_i(t,x), \quad \text{for almost all } (t,x) \in [0,T] \times \Omega.$$

Thus, the evolution of the system with unknown $u := (u_1, \cdots, u_n)^T$ reads under the form (1.1)-(1.2)-(1.3) where for all $u \in \mathbb{R}^n$, the diffusion matrix $A(u) = (A_{ij}(u))_{1 \leq i,j \leq n}$ and the reaction term $f(u) = (f_i(u))_{1 \leq i \leq n}$ are defined as follows : for all $1 \leq i, j \leq n$,

$$A_{ij}(u) = G_{i,j}\left(1 - \sum_{r=1}^{n} u_r, u_1, \cdots, u_n\right) - G_{i,0}\left(1 - \sum_{r=1}^{n} u_r, u_1, \cdots, u_n\right),$$

$$f_i(u) = g_i\left(1 - \sum_{r=1}^{n} u_r, u_1, \cdots, u_n\right).$$

It is thus expected that the values of u lie in the set $\overline{\mathcal{D}_{\text{vf}}}$ where

$$\mathcal{D}_{\text{vf}} := \left\{ (z_1, \cdots, z_n) \in (\mathbb{R}_+^*)^n, \quad \sum_{i=1}^{n} z_i < 1 \right\} \subset [0,1]^n. \tag{1.7}$$

1.1.1 Examples of Cross-diffusion Systems

Let us give some examples of cross-diffusion systems stemming from several applications. Unless it is specified, all the systems presented in this section are written under the form (1.1)-(1.2)-(1.3). The difference between the models lies mainly in the expression of the diffusion matrix and the reaction term.

Example from population dynamics (non volume filling case)

The most standard example of cross-diffusion systems was introduced by Shigesada, Kawasaki and Teramoto [SKT79] to study the spatial segregation of two interacting biological species. In this model of non volume filling type, $n = 2$ and the evolution of $u = (u_1, u_2)^T$ is given by the system (1.1)-(1.2)-(1.3) where the diffusion matrix and the reaction term are respectively given by

$$A : \begin{cases} \overline{\mathcal{D}}_{\text{non-vf}} & \to & \mathbb{R}^{2\times 2} \\ (u_1, u_2) & \mapsto & \begin{pmatrix} d_1 + 2k_{11}u_1 + k_{12}u_2 & k_{12}u_1 \\ k_{21}u_1 & d_2 + k_{21}u_1 + 2k_{22}u_2 \end{pmatrix} \end{cases} \tag{1.8}$$

and

$$f : \begin{cases} \overline{\mathcal{D}}_{\text{non-vf}} & \to & \mathbb{R}^2 \\ (u_1, u_2) & \mapsto & \begin{pmatrix} \beta_1 - b_{11}u_1 - b_{12}u_2 \\ \beta_2 - b_{21}u_1 - b_{22}u_2 \end{pmatrix} \end{cases}, \tag{1.9}$$

where for every $1 \leq i,j \leq 2$, d_i, k_{ij}, β_j and b_{ij} are non-negative parameters. For a given $u \in \overline{\mathcal{D}}_{\text{non}-\text{vf}}$, the diffusion matrix $A(u)$ is in general neither positive definite nor symmetric. Existence of non-negative global-in-time solutions to the SKT system remained an open question for many years. Nevertheless, several works investigated the existence of a solution under suitable assumptions on the diffusion matrix and the reaction terms.

We cite for example the article [Kim84] where the existence of global-in-time non-negative weak solutions was shown under the assumption $k_{11} = k_{22} = 0$ and $k_{12} = k_{21} = k > 0$ and where the initial condition was assumed to satisfy $\|u^0\|_{H^1} \leq M$ for some fixed $M > 0$. Another assumption $k_{21} = 0$ was made in [DT15] and allowed to show existence of global-in-time non-negative weak solutions with initial data satisfying $u_i^0 \geq 0$, for $i = 1, 2$ and $u_1^0 \in L^p(\Omega)$ for some $p > 1$ and $u_2^0 \in L^\infty(\Omega) \cap H^{1+p/d}(\Omega)$. When the cross-diffusion coefficients satisfy $k_{12} < 8k_{11}, k_{21} < 8k_{22}, k_{12} < 8k_{21}$, then the matrix $A(u)$ is positive semi-definite for any $u \in \mathcal{D}_{\text{non}-\text{vf}}$. This latter case was studied in [Yag93] where the existence and uniqueness of non-negative global-in-time weak solutions were proved for initial data satisfying $u^0 \in H^{1+\varepsilon}(\Omega)$ with $\varepsilon > 0$. Using a suitable change of variables (entropy variables), Chen and Jüngel showed in [CJ04, CJ06] a global-in-time existence result for two-components SKT systems under the assumption that $k_{ij} > 0$ for $1 \leq i,j \leq 2$. The initial condition was assumed to lie in an Orlicz space (see the appendix of [CJ04] for a rigorous a definition of the considered Orlicz space) which corresponds to a bounded entropy initial condition.

Several generalizations of the SKT system have been introduced [ZJ15, Lep17] bringing more difficulties in the existence analysis.

Example From Medical Biology (volume filling case)

In the article [JB02], the authors derived a one-dimensional continuous mechanical model for the growth of symmetric avascular tumors. This model describes the evolution of volume fractions of the tumor cells u_1, the extracellular matrix u_2 and the water phases $u_3 = 1 - u_2 - u_1$. The model reads under the form (1.1)-(1.2)-(1.3) where the diffusion matrix and the reaction term are given by

$$A : \begin{cases} \overline{\mathcal{D}}_{\text{vf}} & \to & \mathbb{R}^{2\times 2} \\ (u_1, u_2) & \mapsto & \begin{pmatrix} 2u_1(1-u_1) - \beta\theta^2 u_1 u_2^2 & -2\beta u_1 u_2(1+\theta u_1) \\ -2u_1u_2 + \beta\theta u_2^2(1-u_2) & 2\beta u_2(1-u_2)(1+\theta u_1) \end{pmatrix} \end{cases} \tag{1.10}$$

and

$$f : \begin{cases} \overline{\mathcal{D}}_{\text{vf}} & \to & \mathbb{R}^2 \\ (u_1, u_2) & \mapsto & \begin{pmatrix} \alpha_1 u_1(1-u_1-u_2) - \alpha_2 u_1 \\ \alpha_3 u_1 u_2 (1-u_1-u_2) \end{pmatrix} \end{cases}, \tag{1.11}$$

where $\beta, \theta > 0$ are some positive parameters. The solutions to this system, called the Jackson Byrne tumor growth model, are assumed to satisfy the volume filling constraints (1.6). Yet, proving the existence of a global weak solution satisfying these constraints is not an easy task since the diffusion matrix is in general not positive definite and no maximum/minimum principle applies. The model has nevertheless been studied in [JS12] where existence of global-in-time bounded weak solutions satisfying

the volume filling constrains (1.6) were shown under the assumption that $0 \leq \theta \leq 4/\sqrt{\beta}$ and with an initial condition $u^0 \in L^1(\Omega)$ satisfying the constraints (1.6).

Example from Physics (volume filling case)

Developed independently by James Clerk Maxwell [Max66] for dilute gases and Josef Stefan [Ste71] for fluids, the Stefan-Maxwell equations model the diffusive transport of multicomponent systems such as a mixture of gases. This model is in particular able to predict the experimental results of Duncan and Toor [DT62] on the uphill diffusion phenomena. The Stefan-Maxwell equations for $u_0, \cdots, u_n$ are given by

$$\partial_t u_i + \operatorname{div} J_i = f(u_i), \quad \nabla u_i = -\sum_{i \neq j} \frac{u_j J_i - u_i J_j}{k_{ij}}, \quad \text{for } 0 = 1, \cdots, n. \tag{1.12}$$

where $k_{ij} = k_{ji} > 0$ are the cross-diffusion coefficients between components i and j and where $\sum_{i=0}^{n} u_i = 1$. The system is usually supposed to be physically isolated, thus the reaction term is $f = 0$. Equations (1.12) can be rewritten under the general from (1.1)-(1.2)-(1.3). For instance, the diffusion matrix in the case $n = 2$ is given by

$$A : \begin{cases} \overline{\mathcal{D}}_{\text{vf}} & \to & \mathbb{R}^{2\times 2} \\ (u_1, u_2) & \mapsto & \dfrac{1}{a(u)} \begin{pmatrix} k_{22} + (k_{11} - k_{22})u_1 & (k_{11} - k_{12})u_1 \\ (k_{11} - k_{22})u_2 & k_{12} + (k_{11} - k_{12})u_2 \end{pmatrix} \end{cases} \tag{1.13}$$

with a(u) $= k_{12}k_{22}(1 - u_1 - u_2) + k_{11}(k_{12}u_1 + k_{22}u_2)$. Also, in this system, the diffusion matrix is in general neither symmetric nor positive definite. Thus, it is not obvious to derive suitable a priori bounds for the solutions.

Giovangigli proved in [Gio12] the existence and uniqueness of global-in-time bounded smooth solutions but only when the initial datum u^0 is sufficiently close to the constant equilibrium state u^∞: when $\|u^0 - u^\infty\|_{H^1(\Omega)}$ is sufficiently small. Some results on the existence and uniqueness of local-in-time classical solutions (in the L^p sense) are given in [Bot11, HMPW17] for more general initial condition $u^0 \in H^{2-2/p}, p > (d+2)/2$ satisfying the volume filling constraints (1.6). A three components Stefan-Maxwell system was considered in [BGS12] where it was assumed that the diffusion coefficients are equal, reducing the system to a heat equation for the first component u_1 and an advection-diffusion equation for the second one u_2. In this (simple) situation, existence and uniqueness of global-in-time classical solutions were proved. These solutions were moreover shown to satisfy the volume filling constraints (1.6) and the mass conservation property $\|u(t,\cdot)\|_{L^1(\Omega)} = \|u^0\|_{L^1(\Omega)}$ for $t \in [0, T]$. Based on entropy methods, the first global-in-time existence result of bounded weak solutions (without strong assumptions) was proved in [JS13] for a multi-component Stefan-Maxwell system with general initial condition (measurable functions) satisfying the volume filling constraints (1.6).

Example from Chemistry (volume-filling case)

Let us assume that we are interested in the dynamics of the local concentrations of different chemical species evolving in a cristalline lattice. A model for such a phenomena can be derived from the formal hydrodynamic limit of a stochastic lattice hopping model

(see the appendix of Chapter **??**) resulting in a cross-diffusion system of the form (1.1)-(1.2)-(1.3) with zero reaction term. In the case $n = 2$, the diffusion matrix is given by

$$A : \left\{ \begin{array}{lcl} \overline{\mathcal{D}}_{\rm vf} & \to & \mathbb{R}^{2\times 2} \\ (u_1, u_2) & \mapsto & \begin{pmatrix} (k_{12}-k_{10})u_2 + k_{10} & -(k_{12}-k_{10})u_1 \\ -(k_{21}-k_{20})u_2 & (k_{21}-k_{20})u_1 + k_{20} \end{pmatrix} \end{array} \right. \tag{1.14}$$

where $k_{ij} > 0$ for all $0 \le i \ne j \le 2$. The global-in-time existence of bounded weak solutions to this system is proved in [JZ14] for $n = 2$ and generalized in [BE16] for systems with an arbitrary number $n \ge 2$ of components with initial condition $u^0 \in L^1(\Omega)$ satisfying the volume filling constraints (1.6).

1.2 Limits of Amann's Theory

In this section, we briefly present and comment the main challenges raising from the mathematical analysis of cross-diffusion systems of the form (1.1)-(1.2)-(1.3). Some results that are reported in this section can also be found in [Lep17]. Let us first give some definitions that are useful in our context.

Let $n \in \mathbb{N}^*$ and let $A \in \mathbb{R}^{n\times n}$.

Definition 1.1 (Normal ellipticity)**.** *The matrix A is said to be* elliptic *if the determinant of its symmetric part is positive:*

$$\left| \frac{1}{2}(A + A^T) \right| > 0.$$

The matrix A is said to be normally elliptic *if its eigenvalues have positive real part:*

$$\sigma(A) \subset \{z \in \mathbb{C}, \quad \mathrm{Re}(z) > 0\}$$

where $\sigma(A)$ denotes the spectrum of A.

In the sequel, let $\mathcal{D}$ denote the domain where the solutions of the considered systems are assumed to lie. In the non volume filling case $\mathcal{D} = \mathcal{D}_{\rm non-vf}$ and in the volume filling case $\mathcal{D} = \mathcal{D}_{\rm vf}$.

Definition 1.2 (Normal parabolicity)**.** *A system of the form (1.1) is said to be* parabolic *if its diffusion matrix A is elliptic:*

$$\forall u \in \overline{\mathcal{D}}, \quad \left| \frac{1}{2}(A(u) + A^T(u)) \right| > 0$$

and said to be normally parabolic *if its diffusion matrix A is normally elliptic:*

$$\forall u \in \overline{\mathcal{D}}, \quad \sigma(A(u)) \subset \{z \in \mathbb{C}, \quad r(z) > 0\}.$$

The analysis of cross-diffusion systems is a challenging task from a mathematical point of view [LPR12, Ali79, Kue96, Red89, CJ04, CJ06, DFR08, Jue15a, ZJ15, GR10, Pie10, JS13, Lep17] for the following reasons:

- The equations are *strongly nonlinearly coupled*. As a consequence, standard tools such as the maximum/minimum principle do not apply in general. Besides, there is no regularity theory as in the scalar case. Nice counterexamples are given in [SJ95]: there exist Hölder continuous solutions to certain cross-diffusion systems which are not bounded, and there exist bounded weak solutions which develop singularities in finite time.

- The diffusion matrix is in general not elliptic and may be *degenerate*. Thus, even the local-in-time existence of solutions is not guaranteed. Consider for instance, the two-species SKT system (1.8) with $d_1 = d_2 = 1$ and $k_{11} = k_{22} = 0$ and $k_{12} = k_{21} = k$ and consider $u_1, u_2 \geq 0$. The determinant of the symmetric part of $A(u)$ given by

$$\left|\frac{1}{2}(A + A^T)\right| = (1 + ku_1)(1 + ku_2) - \frac{k^2}{4}(u_1 + u_2)^2$$

  may be negative (e.g. $u_1 = 0$, $u_2 = 1$ and $k = 5$) which means that $A(u)$ is not elliptic. One can easily check that the SKT diffusion matrix (1.8) satisfies

$$\forall u \in \mathcal{D}, \quad \mathrm{Tr}(A(u)) > 0 \quad \text{and} \quad |A(u)| > 0$$

  which fulfills the condition of Definition 1.2. The same verification can be made for the tumor growth and the Stefan-Maxwell models [Jue15b]. This normal parabolicity property is exploited in Amann's works [Ama88, A$^+$90] to prove the existence and uniqueness of local-in-time classical solutions. Yet, the existence of global-in-time solutions still represents a challenge.

- The solution u models concentrations, mass fractions, densities,... thus *upper and/or lower bounds* must be shown to be satisfied. But, as already mentioned, the standard tools as maximum/minimum principle do not apply in general.

Several attempts have been proposed to overcome these difficulties and prove global-in-time existence.

Amann developed a theory of parabolic systems in [Ama88, A$^+$90, Ama89] where he used the normal parabolicity property to prove existence of local-in-time classical solutions for initial conditions in $W^{1,p}$. He also showed that the existence of global-in-time solutions is reduced to deriving suitable $W^{1,p}$ bounds for the local solutions. In particular, the following alternative holds : *either the $W^{1,p}$ norm of the local-in-time solutions explodes in finite time, or the global-in-time solutions exist.* In several works on the SKT system, the global-in-time existence is obtained under assumptions on the cross-diffusion coefficients $(k_{ij})_{1\leq i,j\leq n}$. A typical example is to consider lower or upper triangular diffusion matrices. This gives raise to so-called *triangular systems*. This kind of approach is adpoted for instance in [CLY04, Deu87, HNP15, Kim84, LW15, LZ05, VT08, Wan05].

The question of regularity of the solutions is also a difficult problem. As remarked in [SJ95, Dun00] and unlike the scalar case, one cannot expect in general that bounded weak solutions to cross-diffusion systems are Hölder continuous everywhere. For some particular systems with smooth diffusion matrices, partial regularity results were established in [GS82]. The everywhere Hölder continuity was investigated in [JS98] for

only low dimensional systems $d \leq 2$ and in [Wie92] for an arbitrary space dimension $d \in \mathbb{N}^*$ but with rather restrictive structural conditions. The everywhere regularity of the weak solutions to possibly degenerate systems of the form (1.1)-(1.2)-(1.3) was investigated in [LN06]. Sufficient conditions for the everywhere Hölder continuity of the solutions are given for arbitrary space dimension under several structural assumptions of the diffusion matrix. We refer the reader to [LN06] for the details of these assumptions. Roughly speaking, the strategy of the proof consists in introducing for each weak solution $u : (0,T) \times \Omega \to \mathbb{R}^n$ a set

$$\Sigma(u) := \left\{ (t,x) \in (0,T) \times \Omega : \quad \liminf_{R \to 0} \iint_{Q_R(t,x)} |u(\tau,y) - \overline{u}(t,x)|^2 \mathrm{d}y\mathrm{d}\tau > 0 \right\}$$

with

$$\overline{u}(t,x) = \frac{1}{|Q_R(t,x)|} \iint_{Q_R(t,x)} u(\tau,y)\mathrm{d}y\mathrm{d}\tau.$$

where $Q_R(t,x) = (t-R^2,t) \times B_R(x)$ with $B_R(x)$ being the ball centered at x with radius $R > 0$ and then proving that the $d-$dimensional Hausdorff measure of the set $\Sigma(u)$ is zero for every solution u. Roughly speaking, the set $\Sigma(u)$ contains points $(t,x) \in (0,T) \times \Omega$ where the spread of the solution u is positive for arbitrary small neighborhood of (t,x) meaning that the solution is not continuous at that point.

1.3 Entropy Structure

It appears that several cross-diffusion systems of the from (1.1)-(1.2)-(1.3) admit an *entropy structure* that can be exploited to prove existence of global-in-time bounded weak solutions. Let us first give here a suitable definition of the notion of entropy in our context. Let $\mathcal{D} = \mathcal{D}_{\mathrm{vf}}$ if the considered system is of volume filling type and $\mathcal{D} = \mathcal{D}_{\mathrm{non-vf}}$ if the considered system is of non volume filling type.

Definition 1.3 (Entropy)**.** *We call a function $h : \overline{\mathcal{D}} \to \mathbb{R}$ an* entropy density *associated to the system (1.1)-(1.2)-(1.3) if*

1. *$h \in \mathcal{C}^2(\mathcal{D};\mathbb{R})$,*
2. *h is convex on $\mathcal{D}$,*
3. *the derivative $Dh : \mathcal{D} \to \mathbb{R}^n$ and the Hessian $D^2h : \mathcal{D} \to \mathbb{R}^{n\times n}$ are well defined and invertible,*
4. *the matrix $D^2h(u)A(u)$ is positive semidefinite for every $u \in \mathcal{D}$.*

In this case, we define the *entropy functional* $\mathcal{E}$ of the system as follows

$$\mathcal{E} : \left\{ \begin{array}{ccc} L^\infty((0,T)\times\Omega;\mathcal{D}) & \to & \mathbb{R} \\ u & \mapsto & \int_\Omega h(u)\mathrm{d}x \end{array} \right. \tag{1.15}$$

and we introduce the *entropy variables* $w_1, \cdots, w_n$ as follows

$$(w_1, \cdots, w_n) = w := Dh(u). \tag{1.16}$$

If there exists an entropy functional in the sense of Definition 1.3, then the system under consideration can be formally rewritten with a *gradient flow* structure of the form

$$\partial_t u - \operatorname{div}\left(B(w)\nabla w\right) = f(u), \quad t > 0, \quad u(0,\cdot) = u^0 \quad \text{in } \Omega \tag{1.17}$$

where the matrix $B : \mathbb{R}^n \mapsto \mathbb{R}^{n\times n}$ is called *the mobility matrix* of the system and is defined for every $w \in \mathbb{R}^n$ as $B(w) = A(u(w))(D^2h)^{-1}(u(w))$ with, by definition of the entropy variables, $u(w) = Dh^{-1}(w)$. The terminology "gradient flow" will be justified in Section 1.3.1.

Systems admitting such a formal gradient flow formulation are said to have an *entropy structure*. We will see in the next sections how this property can be exploited to prove the existence of global-in-time weak solutions and to investigate their long time behavior. At this point, let us make the following remark: formally, if one assumes in addition that

$$\forall u \in \mathcal{D}, \quad f(u)\cdot Dh(u) \leq 0, \tag{1.18}$$

then the entropy functional $\mathcal{E}$ is necessarily a Lyapunov functional for the system. Indeed, a simple calculation using (1.18) and the positivity of the mobility matrix B leads to the conclusion:

$$\begin{aligned} \frac{d}{dt}\mathcal{E}(u) &= -\int_\Omega \nabla u \cdot D^2h(u)A(u)\cdot\nabla u \mathrm{d}x + \int_\Omega f(u)\cdot Dh(u)\mathrm{d}x \\ &= -\int_\Omega \nabla w \cdot B(w)\cdot \nabla w \mathrm{d}x + \int_\Omega f(u)\cdot Dh(u)\mathrm{d}x \\ &\leq 0. \end{aligned} \tag{1.19}$$

This *entropy dissipation inequality* is a key-point in the analysis of most of the cross-diffusion systems.

The reader is certainly concerned at this point with the following question: how can one a priori determine if a system of the from (1.1)-(1.2)-(1.3) admits en entropy structure and how to identify an associated entropy density h? This adds actually one more item to our list of mathematical challenges. It is not obvious in the general case to answer this question. Nevertheless, it is observed that many systems modeling tumor-growth, gases mixtures, and ion transport with volume filling constraints (1.6) have an entropy structure induced by the entropy density

$$u \in \mathcal{D}_{\mathrm{vf}} \mapsto h(u) = \sum_{i=1}^n u_i \log u_i - u_i + \rho_u \log \rho_u - \rho_u, \quad \text{with} \quad \rho_u = 1 - \sum_{i=1}^n u_i. \tag{1.20}$$

Note that this function is equvalent (up to the sign minus) to the statistical Boltzmann-Shannon notion of entropy [Jay57]. In this case, the entropic variables w_i can be written and inverted explicitly for all $1 \leq i \leq n$:

$$w_i(u) = \log\left(\frac{u_i}{\rho_u}\right), \quad u_i(w) = \frac{e^{w_i}}{1+\sum_{j=1}^n e^{w_j}}. \tag{1.21}$$

We shall now describe three different methods that exploit the entropy structure to prove the existence of global-in-time weak solutions. Namely, gradient flow theory, boundedness-by-entropy and duality approach.

1.3.1 Elements of Gradient Flow Theory

Let us first mention that a short introduction to *gradient flows* in the general setting of metric spaces is given in the appendix for the reader's convenience. The results gathered in the appendix are mainly extracted from [LAS08, San17].

In summary, one can see the notion of gradient flow as the generalization, to the framework of metric (functional) spaces, of an ordinary differential equation of the form $u'(t) + \nabla\mathcal{E}(u(t)) = 0$, where $\mathcal{E} : \mathbb{R}^n \to \mathbb{R}$ is a functional defined on an Euclidian space, say $\mathbb{R}^n$. Indeed, under suitable smoothness assumptions on $\nabla\mathcal{E}$ (Lipschitz), the associated Cauchy problem, with initial condition $u(t=0) = u^0 \in \mathbb{R}^n$:

$$\begin{cases} u'(t) = -\nabla\mathcal{E}(u(t)) & \text{for } t > 0, \\ u(0) = u^0. \end{cases} \tag{1.22}$$

admits a unique solution $u : [0,T] \to \mathbb{R}^n$. Moreover, this solution can be constructed from a discretization scheme such as the *implicit Euler scheme*.

It is shown in [LAS08], that existence and uniqueness results can also be obtained for Cauchy problems of type (1.22) where the Euclidian space $\mathbb{R}^n$ endowed with the Euclidian distance is replaced by an arbitrary *complete metric space* $\mathcal{M}$ endowed with a distance $d_{\mathcal{M}}$. In this case, the classical notion of gradient $\nabla\mathcal{E}$ which can not be rigorously defined (unless $\mathcal{M}$ is a vector space) is replaced by the notion of *descending slope*. When a (the) curve $u \in [0,T] \to \mathcal{M}$ solution to the gradient system (1.22) in the metric space $(\mathcal{M}, d_{\mathcal{M}})$ exists, we call it a (the) *gradient flow* associated to the functional $\mathcal{E}$.

The authors in [LAS08] proposed three different characterizations of gradient flows in metric spaces. Namely, *gradient flows in the EDE sense*, *gradient flows in the EVI sense* and *gradient flows in the GMM sense* (see Definitions 1.16, 1.17 and 1.15 of the appendix).

Without giving details, the main point of gradient flow theory developed in [LAS08] is that the existence and uniqueness are, in several cases, consequences of the *geodesic* $\lambda-$*convexity*, for some $\lambda \in \mathbb{R}$, of the functional $\mathcal{E} : \mathcal{M} \to \mathbb{R}$ with respect to the distance $d_{\mathcal{M}}$.

Let us now consider a cross-diffusion system of the form (1.1)-(1.2)-(1.3) with $f = 0$ and assume that it admits an entropy structure given by an entropy functional $\mathcal{E}$ defined as in (1.15) and thus reads under the form (1.17). It was observed [LM13, ZM15] that such a system is a formal gradient flow. In other words, a solution u to the problem (1.17) may be seen as a curve of steepest descent, starting from the initial datum u^0, on a manifold $\mathcal{M}$ endowed with a metric $d_{\mathcal{M}}$ induced by the mobility matrix B. More precisely, consider the manifold $\mathcal{M}$ defined by

$$\mathcal{M} = \{v \in H^1(\Omega;\mathbb{R}^n), \quad v(x) \in \overline{\mathcal{D}}, \text{ for almost all } x \in \Omega\} \tag{1.23}$$

and for every $u, v \in \mathcal{M}$, let us define the set of smooth parametric curves that link u to v as follows

$$\mathcal{C}(u,v) := \left\{\gamma \in \mathcal{C}^1\left([0,1];\mathcal{M}\right), \quad \gamma : [0,1] \to \mathcal{M}, \quad \gamma(0) = u, \gamma(1) = v\right\}. \tag{1.24}$$

Following the same steps as in [LM13], one can introduce the metric induced by the mobility matrix B using the Benamou-Brenier formulation as follows: for all $u \in \mathcal{M}$, let $G(u) : H^1(\Omega;\mathbb{R}^n) \to H^1(\Omega;\mathbb{R}^n)$ be the linear operator defined by

$$\forall v \in H^1(\Omega;\mathbb{R}^n), \quad \langle G(u)v, v\rangle = \inf\left\{\int_\Omega \nabla\Psi : B^{-1}(u)\nabla\Psi^T \mathrm{d}x, \quad \mathrm{div}\,\Psi = v \text{ on } \Omega\right\}. \tag{1.25}$$

The associated *optimal transport metric* $d_{\mathcal{M}}$ is therefore given by

$$d_{\mathcal{M}}(u,v) := \left(\inf\left\{\int_0^1 \langle G(\gamma(t))\gamma'(t),\gamma'(t)\rangle \mathrm{d}t, \quad \gamma \in \mathcal{C}(u,v)\right\}\right)^{1/2}. \tag{1.26}$$

The existence/uniqueness of a solution to the cross-diffusion system (1.1)-(1.2)-(1.3) admitting such an entropy structure is then formally equivalent to the existence/uniqueness of a gradient flow $u : [0,T] \to \mathcal{M}$ associated to the Cauchy problem

$$\begin{cases} u'(t) = -\,\mathrm{div}\,(B(u)\nabla Dh(u)) & \text{for } t > 0, \\ u(0) = u_0. \end{cases} \tag{1.27}$$

defined on the metric space $(\mathcal{M}, d_{\mathcal{M}})$. Thus, the existence/uniqueness of such a solution can be obtained, in several cases, from the geodesic $\lambda-$convexity of the entropy functional $\mathcal{E} : \mathcal{M} \to \mathbb{R}$ with respect to the metric $d_{\mathcal{M}}$.

In the case of scalar problem (when $n = 1$), the geodesic $\lambda-$convexity can be characterized in terms of optimal transport problems of Monge-Kantorovitch type [LAS08, OW05, DS08]. Such a tool is no longer at hand in the case of systems (when $n \geq 2$). Nevertheless, other approaches based on the differential characterization of the geodesic convexity property were developed in [LM13, DS08]. Several sufficient conditions (mainly based on the EVI property) are given for $\mathcal{E}$ to be geodesically $\lambda-$convex for some $\lambda \in \mathbb{R}$. The analysis carried in [LM13] and later in [ZM15] allows one to handle several examples. We quote for instance the case of a volume filling type cross diffusion system of n components proposed in [BDFPS10] that reads under the form (1.17) with a mobility matrix given by

$$B(u) = \begin{pmatrix} u_1 - u_1^2 & -u_1u_2 & \cdots & -u_1u_n \\ -u_1u_2 & u_2 - u_2^2 & \cdots & -u_2u_n \\ \vdots & & \ddots & \vdots \\ -u_1u_n & -u_2u_n & \cdots & u_n - u_n^2 \end{pmatrix} \tag{1.28}$$

driven by the entropy (1.20). The following result is obtained for this system:

Theorem 1.4 (Theorem 4.8 of [LM13]). *If $\Omega \subset \mathbb{R}^d$ is bounded, convex and has smooth boundary $\partial\Omega$. Then, the entropy functional $\mathcal{E} : \mathcal{M} \to \mathbb{R}$ defined in the metric space $\mathcal{M}$ (1.23) and given for every $u \in \mathcal{M}$ by $\mathcal{E}(u) = \int_\Omega h(u)$ where h is defined in (1.20) is geodesically $0-$convex with respect to the distance $d_{\mathcal{M}}$ defined in (1.26).*

Several other results of this type are given in [LM13] for scalar problems and weakly coupled reaction-diffusion systems. Proposition 5.3 of [ZM15][1] gives more general conditions on the mobility matrix B for the entropy $\mathcal{E}$ to be geodesically convex, allowing

[1] Proposition 5.3 of [ZM15] is a generalization to the case of systems ($n \geq 2$) of the MacCann's condition that characterizes the $\lambda-$convexity in scalar problems [McC97].

one to treat other systems. However, it is noticed[2] that these conditions are rather restrictive. Unfortunately, most of the cross-diffusion systems with arbitrary diffusion coefficients do not satisfy these assumptions and the geodesic convexity property is not clear in general. To overcome the lack of geodesic convexity, an alternative method was proposed in [ZM15]. The idea is to use the GMM definition of gradient flows. The authors showed in particular that the existence of weak solutions (as limits of the GMM scheme) can still be obtained even if $\mathcal{E}$ is not geodesically convex.

In the sequel, we present two recent alternative approaches to show global-in-time existence of bounded weak solutions: *the duality approach* and *the boundedness-by-entropy technique*. Both methods make use of the entropy structure described below and allow to handle more general cases than the ones covered by the classical results of gradient flow theory.

1.3.2 Boundedness-by-Entropy Method

The main idea of the boundedness-by-entropy method is to use the entropy variables w introduced in (1.21) instead of the classical variables u. To the best of my knowledge, the first introduction of entropy variables in the context of nonlinear coupled systems was in [KS88]. Later, the authors of [DGJ97] used the entorpy variblaes to study a coupled parabolic system describing a multicomponent mixture of charged gases exposed to an electrical field (modeling the electronic transport in semiconductors). The mathematical change of variables $u \mapsto Dh(u)$ is closely motivated by the physical notion of electrochemical potentials. Later, this entropic transformation was used in [CJ04, CJ06] to analyze the SKT system. The authors in [BDFPS10] employ the entropy structure for the analysis of a continuum model describing the transport of two types of particles u_1, u_2 under the influence of electrical fields V and W:

$$\begin{aligned} \partial_t u_1 &= \operatorname{div}\left(k_1(1-u_2)\nabla u_1 + k_1 u_1 \nabla u_2 + k_1 u_1(1-u_1-u_2)\nabla V\right) \\ \partial_t u_2 &= \operatorname{div}\left(k_2(1-u_1)\nabla u_2 + k_2 u_2 \nabla u_1 + k_2 u_2(1-u_1-u_2)\nabla W\right) \end{aligned} \tag{1.29}$$

They proved in particular the existence and uniqueness of strong solutions when the initial data are sufficiently close (in the H^2 norm sense) to the constant steady state. Moreover, they proved existence of global-in-time weak bounded solutions for general initial data (in $L^2(\Omega)$). The method was later analyzed, extended and named *boundedness-by-entropy* method by Jüngel in [Jue15a].

Let us also point out here that the alternative approach, mentioned in the previous section, proposed in [ZM15] to overcome the lack of geodesic convexity can be seen as a particular case of the boundedness-by-entropy method. Indeed, the proofs follow similar arguments in both cases.

We first present in this section the main result of the boundedness-by-entropy approach and make some comments on the assumptions together with a brief sketch of the proof. Then, we discuss some advantages of the approach through the second result of the method which is more adapted to volume filling cases. We will lastly underline some limitations and pathological cases where the method is not helping enough. The reader will find further details in the original paper [Jue15a] and in Chapter 4 of [Jue15b].

[2] as summarized by the following sentence taken from [ZM15] : "$\lambda-$convexity in transportation metrics is a very rare property"

Theorem 1.5 (Theorem 2 of [Jue15a]). *Let $\mathcal{D} = (a,b) \subset \mathbb{R}^n$ be the domain defined in (1.7) if the considered system is of volume filling type and in(1.4) if the considered system is of non volume filling case. Let $A : u \in \overline{\mathcal{D}} \mapsto A(u) := (A_{ij}(u))_{1\leq i,j\leq n} \in \mathbb{R}^{n\times n}$ be a matrix-valued functional defined on $\overline{\mathcal{D}}$ satisfying $A \in \mathcal{C}^0(\overline{\mathcal{D}};\mathbb{R}^{n\times n})$ and let $f : u \in \overline{\mathcal{D}} \mapsto (f_i(u))_{1\leq 1\leq n} \in \mathbb{R}^n$ satisfying $f \in \mathcal{C}^0(\mathcal{D};\mathbb{R}^n)$. Assume in addition that*

(HE1) *there exists a bounded from below convex function $h \in \mathcal{C}^2(\mathcal{D},\mathbb{R})$ such that its derivative $Dh : \mathcal{D} \to \mathbb{R}^n$ is invertible on $\mathbb{R}^n$;*

(HE2) *for all $1 \leq i \leq n$, there exist $\alpha_i^* > 0$ and $1 \geq m_i > 0$ such that for all $z = (z_1,\cdots,z_n)^T \in \mathbb{R}^n$ and $u = (u_1,\cdots,u_n)^T \in \mathcal{D}$,*

$$z^T D^2h(u)A(u)z \geq \sum_{i=1}^n \alpha_i(u_i)z_i^2,$$

where either $\alpha_i(u_i) = \alpha_i^*(u_i-a)^{m_i-1}$ or $\alpha_i(u_i) = \alpha_i^*(b-u_i)^{m_i-1}$.

(HE3) *there exists $a^* > 0$ such that for all $u \in \mathcal{D}$ and $1 \leq i,j \leq n$ for which $m_j > 1$,*

$$|A_{ij}(u)| \leq a^*|\alpha_j(u_j)|.$$

(HE4) *there exists a constant $C_f > 0$ such that*

$$f(u)\cdot Dh(u) \leq C_f\left(1+h(u)\right), \quad \forall u \in \mathcal{D}.$$

Let $u^0 \in L^1(\Omega;\mathcal{D})$ so that $w^0 := Dh(u^0) \in L^\infty(\Omega;\mathbb{R}^n)$. Then, there exists a weak solution u with initial condition u^0 to (1.1)-(1.2)-(1.3) such that for almost all $(t,x) \in \mathbb{R}_+^* \times \Omega$, $u(t,x) \in \overline{\mathcal{D}}$ with

$$u \in L^2_{\text{loc}}((0,T);H^1(\Omega,\mathbb{R}^n)) \text{ and } \partial_t u \in L^2_{\text{loc}}((0,T);(H^1(\Omega;\mathbb{R}^n))').$$

Assumptions **(HE1)** and **(HE2)** mean that the system under consideration admits an entropy structure (in the sense of Definition 1.3). This implies in particular that the matrix $D^2h(u)A(u)$ is positive semi-definite for any $u \in \mathcal{D}$. Hypothesis **(HE3)** is needed to derive uniform bounds for the time derivative $\partial_t u$. Jüngel observed in [Jue15b] that **(HE3)** is only technical and not restrictive. The latter assumption **(HE4)** used to control the reaction term and guarantee the entropy inequality (1.19) is a common assumption in the analysis of reaction-diffusion phenomena.

The strategy of the proof follows the following steps:

S1. A regularization term of the form $\varepsilon((-\Delta)^m + I_d)$ is added to equation (1.17) where $\varepsilon > 0, m > d/2$. This allows one in particular to work in the Sobolev space $H^m(\Omega;\mathbb{R}^n)$ which is embedded in $L^\infty(\Omega;\mathbb{R}^n)$.

S2. The weak formulation of the regularized problem is discretized in time using an implicit Euler scheme with a time step $\tau = T/N$ for some $N \in \mathbb{N}^*$ and $T > 0$.

S3. The existence of a discrete regularized weak entropy solution $w_{\varepsilon,\tau}^k$ to the following iterative implicit scheme is proved : $w_{\varepsilon,\tau}^0 = Dh(u^0)$ and for all $1 \leq k \leq N$ and any test function $\psi \in H^m(\Omega;\mathbb{R}^n)$,

$$\left[\begin{array}{l} \int_\Omega \dfrac{u(w^k_{\varepsilon,\tau}) - u(w^{k-1}_{\varepsilon,\tau})}{\tau}\psi + \int_\Omega \nabla\psi B(w^k_{\varepsilon,\tau})\nabla w^k_{\varepsilon,\tau} \\ \\ \qquad +\varepsilon\int_\Omega \left(\sum\limits_{|\alpha|=m} D^\alpha w^k_{\varepsilon,\tau}\cdot D^\alpha\psi + w^k_{\varepsilon,\tau}\psi\right) \end{array}\right] = \int_\Omega f(u(w^k_{\varepsilon,\tau}))\cdot\psi \quad (1.30)$$

This is done in two steps: First, equation (1.30) is linearized, i.e. the term $B(w^k_{\varepsilon,\tau})$ is replaced by $B(g)$ for some $g \in H^m(\Omega;\mathbb{R}^n)$. The Lax-Millgram lemma is then sufficient to obtain the existence of a unique solution $w^k_{\varepsilon,\tau}(g)$ to the linearized weak problem. Second, an operator $S : H^m(\Omega;\mathbb{R}^n) \to H^m(\Omega;\mathbb{R}^n)$ mapping any function g to the solution $w^k_{\varepsilon,\tau}(g)$ is introduced. The operator S is shown to satisfy the assumptions of the Leray-Shauder fixed point theorem, which allows one to conclude the existence of the weak solution $w^k_{\varepsilon,\tau}$ to the original problem (1.30).

S4. Let $w_{\varepsilon,\tau}$ denote the piecewise constant-in-time interpolation of the sequence $(w^k_{\varepsilon,\tau})_{1\le k\le N}$. Suitable uniform bounds for the weak solutions $u(w^{\varepsilon,\tau})$ are derived using the assumptions **(HE2)**,**(HE3)** and **(HE4)**. Typically, the two following quantities are controlled uniformly in τ and ε:

$$\|\nabla u(w_{\varepsilon,\tau})\|_{L^2((0,T),L^2(\Omega;R^n))} \quad \text{and} \quad \|\tau^{-1}\left(u(w_{\varepsilon,\tau}) - u(w_{\varepsilon,\tau})\right)\|_{L^2((0,T),(H^m(\Omega;R^n))')}$$

S5. The final step consists in passing to the limit $\tau, \varepsilon \to 0$. The main tool to perform this limit is a version of the Aubin-Lions lemma proposed in [DJ12].

Remark 1.6. *Burger and co-authors adopted another strategy in [BDFPS10]. The key ingredients are basically the same : regularization, discretization, uniform bounds using the entropy dissipation property and passing to the limit. The main difference lies in the discretization step. Indeed, Burger and coauthors left the system continuous in time and considered space Galerkin discretization instead. This reduces the problem to proving existence of a solution to a system of ordinary differential equations.*

The boundedness-by-entropy method has been successfully used in several works. In his original paper [Jue15a], Jüngel discussed the applicability of the technique to several examples. We mention here for instance the tumor growth model (1.10) with $\beta = \theta = 1$ which possesses an entropy structure with h defined in (1.20). Moreover, in this case, assumptions (HE1)-(HE2)-(HE3) are automatically satisfied. Indeed,

$$\begin{aligned} z^T D^2h(u)A(u)z &= z_1^1 + (1+u_1)z_2^2 + u_1z_1z_2 \\ &\ge \frac{1}{2}z_1^2 + \left(1 + u_1 - \frac{u_2^2}{2}\right)z_2^2 \\ &\ge \alpha_1(u_1)z_1^2 + \alpha_2(u_2)z_2^2 \end{aligned}$$

with $\alpha_1(u_1) = 1/2$ and $\alpha_2(u_2) = (1-u_2)/2$. The existence is a direct corollary of Theorem 1.5 as soon as the assumption (HE4) on the reaction term is satisfied. The same remarks may be made for the Stefan-Maxwell system. For instance, it is verified for the case $n = 2$ that assumption (HE2) is satisfied with $m_1 = m_2 = 0$. The generalization to the case $n \ge 3$ is done in [JS13]. It seems that the boundedness-by-entropy method is

very favorable to the volume filling case. This remark has been explored deeply in [ZJ15] where generalizations of Theorem 1.5 were proposed to cover more general systems with diffusion matrices of the form

$$A(u) := \begin{cases} \forall 1 \leq i \leq n, \quad A_{ii}(u) = a_i(u)b_i(u_0) + u_i a_i(u) b_i'(u_0) + u_i b_i(u_0)\partial_i a_i(u) \\ \\ \forall 1 \leq i \neq j \leq n, \quad A_{ij}(u) = u_i a_i(u) b_i'(u_0) + u_i b_i(u_0)\partial_j a_i(u) \end{cases} \tag{1.31}$$

where $u = (u_1, \cdots, u_n)^T$ and $u_0 = 1 - \sum_{i=1}^n u_i$. They showed in particular that if there exists functions $\beta : [0,1] \to \mathbb{R}$, $\gamma : \overline{\mathcal{D}} \to \mathbb{R}$ and a real number $\eta > 0$ such that for all $1 \leq i \leq n$,

$$\begin{array}{c} \beta(s) := b_i(s) > 0, \quad \text{for } s \in [0,1] \\ \beta'(s) \geq \eta\beta(s), \quad \text{for } s \in [0,1] \\ \beta(0) = 0, \\ \beta \in \mathcal{C}^3([0,1];\mathbb{R}) \end{array} \tag{1.32}$$

and

$$\begin{array}{c} a_i(u) = \exp(\partial_{u_i}\gamma(u)), \quad \text{for } u \in \mathcal{D}, \\ \gamma \quad \text{convex on } \overline{\mathcal{D}}, \\ \gamma \in \mathcal{C}^3(\mathcal{D};\mathbb{R}), \end{array} \tag{1.33}$$

then the system (with zero reaction term $f = 0$) admits a global-in-time weak solution $u : (0,T) \times \Omega \to \overline{\mathcal{D}}$ satisfying in addition

$$u \in L^\infty((0,T); L^\infty(\Omega, \mathbb{R}^n)) \text{ and } \partial_t u \in L^2((0,T); (H^1(\Omega;\mathbb{R}^n))').$$

The proof of this result follows the same strategy as above with the modified entropy density

$$h(u) = \sum_{i=1}^n u_i(\log u_i - 1) + \int_a^{u_0} \log(\beta(s))\mathrm{d}s + \gamma(u) + (n-1). \tag{1.34}$$

where $a \in (0,1]$ given by

$$a = \begin{cases} 1 & \text{if} \quad \beta(1) \leq 1 \\ \beta^{-1}(1) & \text{if} \quad \beta(1) > 1 \end{cases}$$

Note that the SKT system (1.8) is a particular case of (1.31) where $a_i(u) = k_{ii} + k_{i1}u_1 + k_{i2}u_2$ and $b_i(u_3) = 1$ for $1 \leq i \leq 2$. Additional progress was made in [CDA16] for multicomponent systems of non volume filling type having diffusion matrices of the form

$$A_{ij}(u) = \delta_{ij}a_i(u) + u_i\partial_i a_j(u), \quad a_i(u) = k_{i0} + \sum_{r=1}^n k_{ir}u_r^m \tag{1.35}$$

where $k_{i0}, k_{ij} \geq 0$ and $m > 0$. The main idea of [CDA16] is to introduce the entropy

$$\mathcal{E}(u) = \int_\Omega \sum_{i=1}^n \pi_i h_m(u_i)$$

where $\pi_i > 0$ are some well chosen numbers and where h_m has the form

$$h_m(z) = \begin{cases} z\log z - z + 1 & \text{if} \quad m = 1 \\ \dfrac{z^m - mz}{m-1} + 1 & \text{if} \quad s \neq 1 \end{cases} \tag{1.36}$$

The system admits an entropy structure (in the sense of Definition 1.3) with the functional $\mathcal{E}(u)$ as soon as the numbers π_i satisfy the *detailed balance* property:

$$\pi_i k_{ij} = \pi_j k_{ji}, \quad \forall 1 \leq i, j \leq n. \tag{1.37}$$

Moreover, the mobility matrix is symmetric in this case. The authors observed in particular that there is a relation between condition (1.37) and the symmetry of the mobility matrix. This, together with a positiveness assumption on the diagonal coefficients k_{ii} allowed to prove two global-in-time existence results: the first result concerning linear diffusion rates $m = 1$ and the second one treats nonlinear diffusion rates when $m > \max(0, 1 - d/2)$.

Despite the success of the method, the entropy may sometimes fail to provide the right gradient estimates. Let us try to clarify this point through an example following the arguments of [Lep17]. Let us consider a two species system with a diffusion matrix

$$u \in \mathcal{D}_{\text{non}-\text{vf}} \mapsto A(u) = \begin{pmatrix} k_1 + a_2(u_2) & u_1 a_2'(u_2) \\ u_2 a_1'(u_1) & k_2 + a_1(u_1) \end{pmatrix}, \tag{1.38}$$

where the coefficients k_i and the diffusion rates $a_i : \mathbb{R}_+ \to \mathbb{R}$ satisfy the condition

$$\begin{aligned} k_i > 0, \quad a_i' > 0 \\ \forall (u_1, u_2) \in \mathcal{D}_{\text{non}-\text{vf}}, \quad a_1(u_1)a_2(u_2) - u_1 u_2 a_1'(u_1) a_2'(u_2) \geq 0 \end{aligned} \tag{1.39}$$

This condition ensures in particular that the diffusion matrix A is normally elliptic and allows to show existence of local-in-time solutions. Let us introduce the functions $\phi_1, \phi_2 : \mathbb{R}_+ \to \mathbb{R}$ as follows

$$z \in \mathbb{R}_+ \mapsto \phi_i(z) = \int_0^z \int_0^x \frac{a_i'(y)}{y} \mathrm{d}y \mathrm{d}x$$

with

$$\phi_i'(z) = \int_0^z \frac{a_i'(y)}{y} \mathrm{d}y + C, \quad \phi_i''(z) = \frac{a_i'(z)}{z} + C$$

Thus, a possible choice for the entropy density h associated to the diffusion matrix (1.38) is given by

$$h(u) = \phi_1(u_1) + \phi_2(u_2).$$

Introducing the entropy variables

$$w_i = D_{u_i} h(u) = \phi_i'(u_i) = \int_0^{u_i} \frac{a_i'(y)}{y} \mathrm{d}y = \int_0^{u_i} \phi_i''(y) dy$$

allows one to write the system under the gradient flow structure

$$\partial_t u = \operatorname{div} \left(B(u(w)) \nabla w \right)$$

where the mobility matrix is explicitly given by

$$B(u) = \begin{pmatrix} \dfrac{\alpha_1 u_1 + a_2(u_2) u_1}{a_1'(u_1)} & u_1 u_2 \\ u_1 u_2 & \dfrac{\alpha_2 u_2 + a_1(u_1) u_2}{a_1'(u_1)} \end{pmatrix}$$

and where the gradient of the entropy variables are

$$\nabla w_i = \nabla \left(\int_1^{u_i(x)} \phi_i''(y) dy \right) = \phi_i''(u_i) \nabla u_i.$$

The entropy dissipation term $\int_\Omega (\nabla w)^T B(w) \nabla w$ involves gradient terms of the type

$$\int_\Omega \frac{a_i'(u_i)}{u_i} |\nabla u_i|^2. \tag{1.40}$$

As mentioned previously in the strategy of the proof, the entropy dissipation property is an essential ingredient to derive uniform estimates for the discrete solutions allowing one to pass to the limit. It becomes clear that the efficiency of the method is subject to the control of the terms (1.40) which depend on the expression of the diffusion rates a_i. This remark has been explained by Lepoutre in [Lep17]. In particular, he distinguished between three main forms of entropy densities and reported the expression of the entropy variables and the typical gradient terms (1.40) that need to be controlled in the L^2 norm. For the sake of completeness, we report in Table 1.1 the examples given in [Lep17].

	case 1	case 2	case 3
$a_i(u_i)$	u_i	u_i^m with $m > 0$	$1 - \exp(-u_i)$
$\phi_i(u_i)$	$u_i(\log u_i - 1) + 1$	$\dfrac{u_i^{m-1} - m u_i + m - 1}{m-1}$	$\int \int \dfrac{a_i'(y)}{y} \mathrm{d}y \mathrm{d}u_i$
$w_i(u_i)$	$\log u_i$	$\dfrac{u_i^{m-1} - 1}{m-1}$	$\int \dfrac{a_i'(y)}{y} \mathrm{d}y$
Gradient term	$\dfrac{1}{u_i} \lvert\nabla u_i\rvert^2$	$\dfrac{1}{u_i^{m-2}} \lvert\nabla u_i\rvert^2$	$\dfrac{\exp(-u_i)}{u_i} \lvert\nabla u_i\rvert^2$

Table 1.1 – Three main cases of entropy structures reported from the literature of population dynamics models. Case 1 corresponds to the classical SKT system (1.8) where the diffusion rates are linear [SKT79, Kim84, Jue15a, ZJ15, JZ14, DT15, LM17]. Case 2 corresponds to the generalized SKT system (1.38) studied in [DLMT15, LM17].

This observation shows the limits of the entropy structure to derive the suitable gradient estimates adding one more difficulty to the list of mathematical challenges. This difficulty is not present in several articles using the boundedness-by-entropy technique because of the logarithmic form of the entropy density (1.20). In the more general setting such as cases 2 and 3 of Table 1.1, one needs to invoke additional tools. That brings us to the second method based on *duality* estimates.

1.3.3 Duality method

The main idea of the duality method is to adapt the a priori duality estimates proved in [PS00] in order to obtain L^2 uniform bounds in addition to the entropy bounds stemming from the entropy dissipation property. This method was mainly developed by Desvillettes, Lepoutre, Moussa and collaborators in order to analyze generalized SKT systems. It is therefore more adapted to the non volume filling case. Let then $\mathcal{D} = \mathcal{D}_{\text{non-vf}}$ in this section where $\mathcal{D}_{\text{non-vf}}$ is defined in (1.4).

Let us assume that system (1.1) can be written under the *laplacian* formulation

$$\partial_t u - \Delta(Q(u)) = R(u), \quad t > 0, \quad u(0, \cdot) = u^0 \quad \text{in } \Omega \tag{1.41}$$

with the boundary condition

$$(\nabla Q(u)) \cdot \mathbf{n} = 0 \quad \text{in } \partial\Omega \tag{1.42}$$

where the diffusion term Q and the reaction term R are given respectively by

$$Q : \left\{ \begin{array}{rcl} \mathcal{D} & \to & \mathbb{R}_+^n \\ u & \mapsto & Q(u) = (q_i(u)u_i)_{1\le i\le n} \end{array} \right. \tag{1.43}$$

$$R : \left\{ \begin{array}{rcl} \mathcal{D} & \to & \mathbb{R}_+^n \\ u & \mapsto & R(u) = (r_i(u)u_i)_{1\le i\le n} \end{array} \right. \tag{1.44}$$

with some measurable functions q_i, r_i for $1 \le i \le n$ whose regularity will be made more precise later in the assumptions of the existence result.

Note that all the systems of the form (1.1) cannot in general be written under the laplacian formulation (1.41). However, this is the case for the SKT system which is treated in [ZJ15, CDA16] with the divergence-gradient formulation (1.1) and in [LM17, Lep17] with the laplacian formulation (1.41).

The method presented in this section is mainly based on the duality estimates shown by Pierre and Schmitt in [PS00] for reaction diffusion systems. We give here a version of their result that is suitable to our context.

Lemma 1.7 (Duality estimate, [PS00, LM17]). *Let $\lambda \ge 0$ and let $\varphi^0 : \Omega \to \mathbb{R}_+$ be a measurable nonnegative function and denote by $\overline{\varphi^0}$ its average on Ω. Consider an integrable function $a : (0,T)\times\Omega \to \mathbb{R}$ satisfying $a(t,x) \ge \nu > 0$ for every $(t,x) \in [0,T]\times\Omega$. Let φ be a smooth solution to the inequation*

$$\begin{array}{rcll} \partial_t\varphi - \Delta[a\varphi] & \le & \lambda\varphi & \text{on } (0,T)\times\Omega \\ \varphi(0,\cdot) & = & \varphi^0 & \text{on } \Omega \\ \partial_x(a\varphi)\cdot\boldsymbol{n} & = & 0 & \text{on } [0,T]\times\partial\Omega. \end{array} \tag{1.45}$$

Then, the following estimate holds

$$\int_0^T\int_\Omega a\varphi^2 \le \mathrm{e}^{2\lambda T}\left(\|\varphi^0 - \overline{\varphi^0}\|_{(H^1(\Omega;\mathbb{R}))'} + \overline{\varphi^0}\int_0^T\int_\Omega a \right) \tag{1.46}$$

where $(H^1(\Omega;\mathbb{R}))'$ denotes the dual space of $H^1(\Omega;\mathbb{R})$.

The original proof of Pierre and Schmitt is based on a dual formulation of the problem. Nevertheless, Lepoutre suggests another proof in [Lep17] based on direct computations: multiply equation (1.45) by $a\varphi$ and integrate first in space and then integrate in time. This direct calculation can be discretized, which is useful in the proof of the global-in-time existence result that will be stated later.

Roughly speaking, Lemma 1.7 tells us that the solution φ can be controlled if we have suitable controls on the terms involving the function a. This property is employed in the cross-diffusion system (1.41) as follows. For the sake of simplicity, let us (temporarily) consider that $r_i = 0$ for all $1 \le i \le n$. Then, summing up all the equations of the system yields to

$$\partial_t\varphi - \Delta[a\varphi] = 0, \quad \text{with} \quad \varphi := \sum_{i=1}^n u_i \quad \text{and} \quad a := \frac{\sum_{i=1}^n q_i(u)u_i}{\varphi}.$$

Thus, in virtue of Lemma 1.7 and since $\varphi \geq 0$ by construction, the following estimate holds

$$\int_0^T \int_\Omega a\varphi^2 \leq C\left(1 + \int_0^T \int_\Omega a\right)$$

where the constant C is given by $C := \mathrm{e}^{2\lambda T} \max\left(\|\varphi^0 - \overline{\varphi^0}\|_{(H^1(\Omega,\mathbb{R}))'}, \overline{\varphi^0}\right)$. The important point is that C does not depend on the solution φ but only on the parameters T, λ and the initial condition φ^0. If we assume in addition that the diffusion terms are continuous, i.e $q_i \in \mathcal{C}^0(\mathcal{D})$ for any $1 \leq i \leq n$, then the following control in obtained (see the appendix of [LM17]) for the solution $u = (u_i)_{1\leq i\leq n}$ to (1.41):

$$\int_0^T \int_\Omega \left(\sum_{i=1}^n u_i\right)\left(\sum_{i=1}^n q_i(u)u_i\right) \leq C \tag{1.47}$$

where the constant C depends on Ω, T, u^0, λ and the diffusion rates q_i. This (formal) estimate is a key-point in the proof of the global-in-time existence of weak solutions to systems of the form (1.41).

Let us state here a version of the main existence result of the duality method proposed in [LM17]. We comment on the assumptions and give the main arguments of the proof right after.

Theorem 1.8 (Existence by duality approach, [LM17])**.** *Consider a cross-diffusion system of the form (1.41)-(1.42) and let $\mathcal{D} = \mathcal{D}_{\mathrm{non-vf}}$ be the non volume filling domain defined in (1.4). Assume that the diffusion rates q_i and the reaction terms r_i satisfy the following assumptions:*

(HD1) *For every $1 \leq i \leq n$,*

$$q_i \in \mathcal{C}^0(\overline{\mathcal{D}}; \mathbb{R}_+) \cap \mathcal{C}^1(\mathcal{D}; \mathbb{R}_+), \quad r_i \in \mathcal{C}^0(\mathcal{D}; \mathbb{R}).$$

(HD2) *There exist positive constants $\alpha, \lambda > 0$ such that for every $0 \leq i \leq n$,*

$$p_i \geq \alpha, \quad \text{and} \quad r_i \leq \lambda.$$

(HD3) *Q is a self-homeomorphism[3] on $\mathcal{D}$.*

(HD4) *There exists an entropy density h (in the sense of Definition 1.3). In addition,*

(HD4)' *there exists continuous functions $\sigma_i : \mathbb{R}_+^* \to \mathbb{R}_+^*$ such that for every $z \in \mathbb{R}^n$ and every $u \in \mathcal{D}$,*

$$z^t D^2 h(u) \nabla Q(u) z \geq z^t \mathrm{Diag}(\sigma_i(u_i)) z.$$

(HD4)" *for some $C_R > 0$ and any $Z \in \mathcal{D}$, $Dh(z) \cdot R(z) \leq C_R(1 + h(z))$.*

(HD5) *The reaction term R satisfies[4]*

$$R(z) = o\left(\left(\sum_{i=1}^n q_i(z) z_i\right)\left(\sum_{i=1}^n z_i\right) + h(z)\right) \quad \text{as} \quad \|z\| \to \infty$$

[3] A homeomorphism $Q : E \to F$ between two topological spaces E and F is a continuous bijection with a continuous inverse. When $E = F$, Q is called self-homeomorphism

[4] All the norms being equivalent in $\mathbb{R}^n$, it suffices to choose an arbitrary norm $\|\cdot\|$.

Then, for any integrable initial condition $u^0 : \Omega \to \mathcal{D}$, such that $h(u^0) \in L^1(\Omega)$, there exists a weak solution $u \in L^1([0,T] \times \Omega; \mathcal{D})$ to the system (1.41)-(1.42) such that $Q(u), R(u) \in L^1([0,T] \times \Omega; \mathcal{D})$. Moreover, for every $t \in [0,T]$,

$$\int_\Omega h(u(t,\cdot)) + \int_0^t \int_\Omega (\nabla u(t,\cdot))^T (D^2 h(u(t,\cdot)) \nabla Q(u(t,\cdot)))(\nabla u(t,\cdot)) \leq (1+\mathrm{e}^{2C_R T}) \int_\Omega h(u^0)$$

Let us make some brief comments on the assumptions of the theorem before we give the main ideas of the proof. First, the continuity assumption **(HD1)** is essential in the proof and the result may fail if this assumption is removed. The bounds in **(HD2)** allow one to invoke Lemma 1.7 and it was remarked by the authors that they probably can be weakened. Hypothesis **(HD3)**, which seems restrictive at a first sight, was investigated in Section 4 of [LM17] and shown to be a consequence of the entropy structure in several cases. The structural assumptions **(HD4)**,**(HD4)'** and **(HD4)"** are of the same family as **(HE2)**,**(HE3)** and **(HE4)** appearing in Theorem 1.5. Lastly, hypothesis **(HD5)** is a technical assumption that can probably be weakened as well.

The proof of Theorem 1.8 is structured in three main steps : an implicit time discretization is first introduced, for which existence of (discrete) solutions is shown. Then, the entropy dissipation property and the duality a priori estimate are exploited to derive suitable uniform bounds. Lastly, the weak solutions to the continuous system are obtained as the limit of the discrete ones when the time step goes to zero. More precisely, The following implicit scheme is introduced. Let $N \in \mathbb{N}^*$ and let $\tau = T/N$ and consider the iterative problem, for $1 \leq k \leq N$

$$\frac{u_\tau^k - u_\tau^{k-1}}{\tau} - \Delta[Q(u_\tau^k)] = R(u^k) \text{ on } \Omega, \tag{1.48}$$

$$\nabla Q(u_\tau^k) \cdot \mathbf{n} = 0 \text{ on } \partial\Omega, \tag{1.49}$$

with the initialization u_τ^0, which is a suitably chosen approximation of the continuous initial condition u^0. The semi-discrete system (1.48)-(1.49) was studied in [DLMT15]. It is proven in Theorem 2.2 of [DLMT15] that, under the assumptions **(HD1)**-**HD2)**-**(HD3)**, there exists a nonnegative sequence $(u_\tau^k)_{1 \leq k \leq N-1}$ belonging the space $L^\infty(\Omega)$ solving (1.48)-(1.49). Moreover, the a priori duality estimate (1.47) is preserved in the discrete level : there exists a constant $C > 0$ depdending on $\Omega, u^0, Q, \lambda, N$ such that

$$\sum_{k=0}^{N-1} \tau \int_\Omega \left(\sum_{i=1}^n u_{\tau,i}^k \right) \left(q_i(u^k) u_{\tau,i}^k \right) \leq C, \quad \forall 1 \leq i \leq n. \tag{1.50}$$

Using the convex character of the entropy density h and the assumption **(HD4)"** allows to obtain a discrete version of the entropy dissipation :

$$\int_\Omega (h(u_\tau^k) - h(u_\tau^{k-1})) + \tau \int_\Omega (\nabla u_\tau^k)^T D^2 h(u_\tau^k) \nabla Q(u_\tau^k)(\nabla u_\tau^k) \leq \tau C \left(1 + \int_\Omega h(u_\tau^k) \right)$$

which, together with the assumption **(HD4)'**, provide suitable L^2 uniform bounds. The passing to the limit is rather technical and uses a non linear variant of the Aubin-Lions lemma proposed in Proposition 3 of [Mou16].

1.4 Uniqueness of Solutions

The methods presented in the previous sections do not allow one to obtain uniqueness of solutions. Other approaches must be employed. Only few uniqueness results can be found in the literature. To the best of my knowledge, a general uniqueness result, -or at least a robust method of investigating uniqueness- still remains an open question. We report in this section some particular cases where uniqueness of the global-in-time weak solutions can be shown.

1.4.1 Fully Decoupled Systems

The easiest scenario is obviously the fully decoupled case. More precisely, we consider here systems of the form (1.1)-(1.2)-(1.3) having diagonal diffusion matrices $A(u) = \mathrm{diag}[a_i(u_i)]_{1\le i\le n}$ and decoupled reaction terms $f(u) = (f_i(u_i))_{1\le i\le n}$ where $a_i : \mathbb{R} \to \mathbb{R}$ and $f_i : \mathbb{R} \to \mathbb{R}$ are smooth (enough) functions. In this case, the system is reduced to a set of decoupled reaction diffusion scalar equations

$$\partial_t u_i = \mathrm{div}(a_i(u_i)\nabla u_i) + f_i(u_i), \quad 1 \le i \le n. \tag{1.51}$$

The analysis of such scalar equations has a much longer history. The reader may refer for example to [Eva98, BCL99, EG02]. As already mentioned in Section 1.3, fully decoupled cross-diffusion systems having an entropy structure may also be treated with the gradient flow theory tools [ZM15].

1.4.2 H^{-1} Method

Consider an isolated ($f = 0$) cross-diffusion system of the form (1.1)-(1.2)-(1.3) to which the existence of a global-in-time solution u is proved. Uniqueness of the solution can be shown if we assume that there exists a function $\Psi : \mathbb{R}^n \to \mathbb{R}^n$ satisfying the monotony property

$$\forall w, v \in \mathcal{D}, \quad (\Psi(w) - \Psi(v)) \cdot (w - v) \ge 0 \tag{1.52}$$

and such that for every solution u to (1.1)-(1.2)-(1.3), $A(u)\nabla u = \Psi(u)$. Indeed, consider two weak solutions u and v with the same initial data u^0 and let $\theta \in L^2(0,T;H^1(\Omega;\mathbb{R}^n))$ be the weak solution to the Neumann Poisson problem

$$\begin{array}{rcll} -\Delta\theta & = & u - v & \text{on} \quad \Omega, \\ \nabla\theta \cdot \mathbf{n} & = & 0 & \text{on} \quad \partial\Omega, \end{array} \tag{1.53}$$

which is unique (up to an additive constant). Then, we formally have,

$$\begin{aligned} \frac{d}{dt}\int_\Omega |\nabla\theta|^2 \mathrm{d}x &= \langle \partial_t(-\Delta\theta), \theta\rangle \\ &= \langle \partial_t(u - v), \theta\rangle \\ &= \langle \mathrm{div}\left(\nabla\Psi(u) - \nabla\Psi(v)\right), \theta\rangle \\ &= -\langle \nabla\Psi(u) - \nabla\Psi(v), \nabla\theta\rangle \\ &= -\langle \Psi(u) - \Psi(v), -\Delta\theta\rangle \\ &= -\langle \Psi(u) - \Psi(v), u - v\rangle \\ &\le 0 \end{aligned}$$

This allows to conclude that θ is constant and thus necessarily $u = v$. This structural assumption is rather strong and in general not satisfied. If the cross-diffusion system admits en entropy structure and reads under the form (1.17), then the following slightly weaker assumption allows to obtain the uniqueness of the weak solutions using the same arguments : assume that there exists a function $\Phi : \mathbb{R}^n \to \mathbb{R}^n$ such that the composition $\Psi \circ Dh$ satisfies the monotony property (1.52) and such that for every entropy solution $w = Dh(u)$ to (1.1)-(1.2)-(1.3), $B(w)\nabla w = \nabla\Phi(w)$. Unfortunately, even this assumption is rather restrictive in practice.

1.4.3 Gajewski Method

In the volume filling case where the diffusion matrix has the form (1.31) with assumptions (1.32)-(1.33) and $f = 0$, uniqueness can also be shown when the diffusion rates a_i are supposed to be constant and equal to one, i.e. for all $0 \leq 0 \leq n$, $a_i = 1$. In this case, every component u_i for $0 \leq i \leq n$ solves the equation

$$\partial_t u_i = \operatorname{div}(\beta(u_0)\nabla u_i - u_i \nabla\beta(u_0)),$$

where β is the function coming from assumption (1.32). Summing up all the equations for $i = 1, \cdots, n$ yields to a simple diffusion equation for the last component u_0 which, we recall that from the volume filling constraint, is given by $u_0 = 1 - \sum_{i=1}^n u_n$:

$$\begin{aligned}
\partial_t u_0 &= -\partial_t \left(\sum_{i=1}^n u_i \right) \\
&= -\sum_{i=1}^n \operatorname{div}\left(\beta(u_0)\nabla u_i - u_i\nabla\beta(u_0)\right) \\
&= -\operatorname{div}\left(-\beta(u_0)\nabla u_0 - (1-u_0)\nabla\beta(u_0)\right) \\
&= \operatorname{div}\left(\beta(u_0)\nabla u_0 + (1-u_0)\nabla\beta(u_0)\right) \\
&= \operatorname{div}\left(\nabla\Psi(u_0)\right)
\end{aligned}$$

where the non linear function $\Psi : [0,1] \to \mathbb{R}$ is by construction defined for every $z \in [0,1]$ by $\Psi(z) = \int_0^z \beta(x) + (1-x)\nabla\beta(x)\mathrm{d}x$. Moreover, It follows from assumption (1.32) that Ψ is non decreasing on $[0,1]$. Thus, the uniqueness of u_0 can immediately be obtained using the arguments of the H^{-1} method presented previously.

The uniqueness of the remaining solutions $u_1, \cdots, u_n$ is shown by the *E-monotonicity* method[5] proposed firstly by Gajewski in [Gaj94b, Gaj94a] in the context of drift diffusion models for semiconductors and then developed in [ZJ15] for volume filling cross-diffusion systems. Briefly speaking, let $\eta > 0$ be a positive parameter and introduce a semi-metric[6] d_η on the space $L^\infty([0,T]\times\Omega;\mathcal{D}_{\mathrm{vf}})$ (denoted simply by L^∞ to shorten the

[5] using the terminology of [Jue15b]

[6] A semi-metric is a function that satisfies the positiveness, the positive definiteness and the symmetry properties but not necessarily the triangular inequality.

notation) as follows

$$d_\eta : \left\{ \begin{array}{ccc} L^\infty \times L^\infty & \to & \mathbb{R}_+ \\ u, v & \mapsto & d_\eta(u,v) = \sum_{i=1}^n \int_\Omega \left(h_\eta(u_i) + h_\eta(v_i) - 2h_\eta\left(\frac{u_i+v_i}{2}\right)\right) \mathrm{d}x \end{array} \right. \tag{1.54}$$

where h_η is a regularized entropy defined also on X as follows

$$h_\eta : \left\{ \begin{array}{ccc} L^\infty \to & \mathbb{R} \\ u & \mapsto & h_\eta(u) = (u+\eta)\log(u+\eta) - (u+\eta) + 1. \end{array} \right. \tag{1.55}$$

The regularization parameter $\eta > 0$ is necessary for the term $\log\left((u_i+v_i)/2\right)$ to be well defined when the solutions u_i and v_i vanish. Note first that thanks to the convexity of the entropy h_η, it follows that

$$d_\eta(u,v) \geq 0, \quad \forall u, v \in X.$$

Furthermore, using Taylor expansions and the fact that the function $[0,1] \ni z \mapsto h_\eta''(z)$ is bounded from below by $1/2$ allows one to obtain the following estimate for every $1 \leq i \leq n$,

$$h_\eta(u_i) + h_\eta(v_i) - 2h_\eta\left(\frac{u_i+v_i}{2}\right) \geq \frac{1}{8}(u_i - v_i)^2. \tag{1.56}$$

Moreover, some elementary algebraic manipulations (we do not report all the details here but the reader may refer to [ZJ15]) yields:

$$\begin{aligned} d_\eta(u,v) = -4\int_0^T \sum_{i=1}^n \int_\Omega \left(|\nabla\sqrt{u_i+\eta}|^2 + |\nabla\sqrt{v_i+\eta}|^2 - |\nabla\sqrt{u_i+v_i+2\eta}|^2 \right) \beta(u_{n+1}) \mathrm{d}x\mathrm{d}t \\ + 2\int_0^T \sum_{i=1}^n \int_\Omega \left(\frac{u_i}{u_i+\eta} - \frac{u_i+v_i}{u_i+v_i+2\eta} \right) \sqrt{\beta(u_0)}\nabla\sqrt{\beta(u_0)}\nabla u_i \mathrm{d}x\mathrm{d}t \\ + 2\int_0^T \sum_{i=1}^n \int_\Omega \left(\frac{u_i}{v_i+\eta} - \frac{u_i+v_i}{u_i+v_i+2\eta} \right) \sqrt{\beta(u_0)}\nabla\sqrt{\beta(u_0)}\nabla v_i \mathrm{d}x\mathrm{d}t. \end{aligned}$$

The first integral of the right hand side can be shown to be nonnegative thanks to the subadditivity property of the Fischer information $F(u) := \int_\Omega |\nabla\sqrt{u}|^2$ (see Lemma 9 of [ZJ15]). Furthermore, the two remaining integrals of the right hand side tend to zero as η goes to 0 via the dominated convergence theorem since all the terms of the integrands are bounded. Hence, it holds that for every $1 \leq i \leq n$,

$$h_\eta(u_i) + h_\eta(v_i) - 2h_\eta\left(\frac{u_i+v_i}{2}\right) \to 0 \quad \text{a.e. in} \quad (0,T)\times\Omega. \tag{1.57}$$

Finally, estimate (1.56) together with the limit (1.57) allow to infer $(u_1, \cdots, u_n) = (v_1, \cdots, v_n)$ which concludes the proof of uniqueness. Details of this proof can be found in [ZJ15]. Unfortunately, the assumption $a_i = 1$ is rather strong and this strategy does not seem to apply for weaker assumptions on the diffusion rates.

Let us finally mention that other non-general uniqueness results can be obtained in some particular cases. We mention for example [Bot11, HMPW17] where the uniqueness of local-in-time solutions to the Stefan-Maxwell system were proved. In [Gio12] and [BDFPS10] the uniqueness of the global-in-time weak solutions is obtained for initial condition that is sufficiently close to the constant steady states.

1.5 Long-time Behavior

The long-time behavior of the solutions is also an important feature in the study of cross-diffusion systems of the form (1.1)-(1.2)-(1.3). Let us assume in this section that $f = 0$. The steady states of such systems are usually given by the constant profiles

$$u_i^\infty = \frac{1}{|\Omega|}\int_\Omega u_i^0(x)\mathrm{d}x, \quad \forall 1 \leq i \leq n. \tag{1.58}$$

Neverthless, different steady states may co-exist for the same system. This phenomenon is particularly observed in the (different variants of the) SKT system where several works investigated the question [GQZQXL08, CP04, Wen13, BLMP09]. The formation of patterns (called *Turing patterns* and corresponding to non-constant equilibrium profiles) is for example theoretically studied and numerically characterized in [BLMP09]. Additional conditions on the cross-diffusion coefficients that lead to the existence of non-constant steady states bifurcating from the constant ones were given in [LM14] and assessed numerically.

The authors in [BDFPS10] investigated the convergence of the solutions of the two species ion transport model (1.29) to the constant steady states (1.58). They showed in particular a strong L^1 convergence but did not give a rate for it. Later, Jüngel and Zamponi investigated in [ZJ15] the long-time behavior of volume filling systems having diffusion matrices of the form (1.31) with assumptions (1.32)-(1.33) and $f = 0$. They were able to prove an exponential convergence for all the species under the additional assumptions that β' is strictly positive and β/β' is concave on $(0, 1)$.

The mathematical arguments used in [BDFPS10, ZJ15, JS12] are standard arguments in the asymptotic analysis of PDEs solutions. We can summarize the strategy of the proof by the following points.

- Introduce a suitable relative entropy $\overline{\mathcal{E}}(u, u^\infty)$ for the system. A suitable choice of $\overline{\mathcal{E}}$ for cross-diffusion systems that have logarithmic entropy density (1.20) is
$$\overline{\mathcal{E}}(u, u^\infty) = \mathcal{E}\left(\frac{u}{u^\infty}\right)$$
and different forms may be more convenient in other cases.

- Estimate from below the entropy dissipation term by means of the relative entropy.i.e. find $\lambda > 0$ such that,
$$\int_\Omega \nabla w B(w)\nabla w \geq \lambda\overline{\mathcal{E}}(u, u^\infty), \quad \text{with} \quad \lambda > 0, \tag{1.59}$$
which is equivalent to
$$\int_\Omega \nabla u D^2h(u)A(u)\nabla u \geq \lambda\overline{\mathcal{E}}(u, u^\infty).$$
Obtaining such estimate is subject to the assumptions made for the diffusion matrix. For instance, when $A(u)$ satisfies hypotheses (HE2) of Theorem 1.5 with $\alpha_i^* = m_i = 1/2$ then the question is reduced to prove that
$$\sum_{i=1}^n \int_\Omega |\nabla\sqrt{u_i}|^2\mathrm{d}x \geq \lambda\sum_{i=1}^n \int_\Omega u_i \log\left(\frac{u_i}{u_i^\infty}\right)\mathrm{d}x$$
which can be easily done via a Logarithmic Sobolev inequality [ABL00].

- Exploiting the entropy dissipation inequality

$$\frac{d}{dt}\mathcal{E}(u) + \int_\Omega \nabla w B(w) \nabla w \leq 0$$

satisfied by the system allows to obtain

$$\frac{d}{dt}\mathcal{E}(u) + \lambda \overline{\mathcal{E}}(u, u^\infty) \leq 0.$$

Thus, the Gronwall lemma leads to the exponential convergence of the relative entropy

$$\overline{\mathcal{E}}(u, u^\infty) \leq \overline{\mathcal{E}}(u^0, u^\infty) \exp(-\lambda t)$$

- The Csizár Kullback inequality [ABL00] allows to conclude the proof

$$\|u - u^\infty\|_{L^1(\Omega)} \leq C \exp\left(-\frac{\lambda}{2} t\right).$$

1.6 Contributions of the Thesis

This section is a summary of our main contributions related to the study of cross-diffusion systems.

A one-dimensional cross-diffusion system in a moving domain

Consider a multicomponent system composed of $n+1$ different species ($n \geq 2$) and consider functions $(\phi_0, \cdots, \phi_n)$ belonging to $L^\infty_{\text{loc}}(\mathbb{R}_+; \mathbb{R}_+^{n+1})$, which we refer to in the sequel as *external fluxes*. Let $e_0 > 0$ and for every $t \in \mathbb{R}_+$, let $e(t) := e_0 + \int_0^t \sum_{i=0}^n \phi_i(s)\, ds$. For every $0 \leq i \leq n$, denote by $u_i(t,x)$ the volume fraction of the species i at time t and point $x \in (0, e(t))$. Consider an initial condition given by the integrable functions $u_0^0, \cdots u_n^0$ satisfying the volume filling constraints (1.6). For every $0 \leq i \neq j \leq n$, let $K_{ij} = K_{ji} > 0$ denote the cross-diffusion coefficient between species i and j. Consider the matrix $A : [0,1]^n \to \mathbb{R}^{n \times n}$ defined for every $u \in \mathcal{D}_{\text{vf}}$ by

$$\begin{cases} \forall 1 \leq i \leq n, \quad A_{ii}(u) = \displaystyle\sum_{1 \leq j \neq i \leq n} (K_{ij} - K_{i0}) u_j + K_{i0}, \\ \forall 1 \leq i \neq j \leq n, \quad A_{ij}(u) = -(K_{ij} - K_{i0}) u_i. \end{cases} \tag{1.60}$$

Let us lastly denote by $u = (u_1, \cdots, n_n)^T$ and by $\varphi = (\phi_1, \cdots, \phi_n)^T$. The system that we mainly analyze in the first part of this thesis reads

$$\begin{cases} e(t) = e_0 + \int_0^t \sum_{i=0}^n \phi_i(s)\, ds, & \text{for } t \in \mathbb{R}_+^*, \\ \partial_t u - \partial_x \left(A(u) \partial_x u\right) = 0, & \text{for } t \in \mathbb{R}_+^*,\ x \in (0, e(t)), \\ \left(A(u)\partial_x u\right)(t,0) = 0, & \text{for } t \in \mathbb{R}_+^*, \\ \left(A(u)\partial_x u\right)(t, e(t)) + e'(t) u(t, e(t)) = \varphi(t), & \text{for } t \in \mathbb{R}_+^*, \\ u(0,x) = u^0(x), & \text{for } x \in (0, e_0). \end{cases} \tag{1.61}$$

Let us mention that we initially introduced this system to model the PVD process used in the production of thin film solar cells. The function $\mathbb{R}_+ \mapsto e(t)$ models the thickness of the thin film and the functions $\mathbb{R}_+ \mapsto \phi_i(t)$ model the external atomic fluxes injected

in the chamber during the process. A formal derivation of the diffusive term from a stochastic lattice hopping model is given in Section 2.7 of Chapter 2.

For all $0 \leq i \leq n$, $t \geq 0$ and $y \in (0,1)$, we denote by $v_i(t,y) := u_i(t, e(t)y)$. Thus, u is a solution to (1.61) if and only if v is a solution to the following system:

$$\begin{cases} e(t) = e_0 + \int_0^t \sum_{i=0}^n \phi_i(s)\, ds, & \text{for } t \in \mathbb{R}_+^*, \\ \partial_t v - \frac{1}{e(t)^2}\partial_y \left(A(v)\partial_y v\right) - \frac{e'(t)}{e(t)} y \partial_y v = 0, & \text{for } (t,y) \in \mathbb{R}_+^* \times (0,1), \\ \frac{1}{e(t)}(A(v)\partial_y v)(t,1) + e'(t)v(t,1) = \varphi(t), & \text{for } (t,y) \in \mathbb{R}_+^* \times (0,1), \\ \frac{1}{e(t)}(A(v)\partial_y v)(t,0) = 0, & \text{for } (t,y) \in \mathbb{R}_+^* \times (0,1) \\ v(0,y) = v^0(y) := u^0(e_0 y), & \text{for } y \in (0,1). \end{cases} \tag{1.62}$$

This rescaled version, which is equivalent to (1.61) allows to get rid of the moving boundary. But, the drawback is the presence of the advection term $\frac{e'(t)}{e(t)} y \partial_y v$. The system (1.62) is a *cross-diffusion-advection* system with *mixed boundary conditions* which does not fall in the classical framework (1.1)-(1.2)-(1.3). To the best of my knowledge, the presence of such drift terms and boundary conditions has never been considered for strongly coupled cross-diffusion systems.

Note that in the case where the external fluxes vanish, the system perfectly falls in the general framework (1.1)-(1.2)-(1.3) and writes in an arbitrary (smooth enough) domain $\Omega \subset \mathbb{R}^d, d \geq 1$, as follows

$$\begin{cases} \partial_t u - \operatorname{div}\left(A(u)\nabla u\right) = 0, & \text{for } t \in \mathbb{R}_+^*,\ x \in \Omega, \\ (A(u)\nabla u) \cdot \mathbf{n} = 0, & \text{for } t \in \mathbb{R}_+^*,\ x \in \partial\Omega, \\ u(0,x) = u^0(x), & \text{for } x \in \Omega. \end{cases} \tag{1.63}$$

Existence

As a first preliminary result, we show that the zero-fluxes system (1.63) with the diffusion matrix (1.60) satisfies the assumptions of Theorem 1.5 and admits thus global-in-time bounded weak solutions. Then, our main result concerns one-dimensional non-zero fluxes systems of the form (1.62) with an arbitrary diffusion matrix A. The existence theorem is proved in Section 2.4.2 of Chapter 2.

Long-time Behavior for Constant fluxes

When the external fluxes φ are constant-in-time and the entropy density h associated to the system (1.62) is of the logarithmic form (1.20), we show that the weak solutions to the system converge (for the L^1-norm) in the long-time limit to constant steady profiles at a rate inversely proportional to the square root of time. This asymptotic result is proved in Section 2.4.3 of Chapter 2.

Optimization of the external fluxes

Our initial motivation for studying system (1.61) is the control of the external atomic fluxes injected during a PVD process in order to achieve a certain thickness and certain final concentration profiles. To this aim, we formulate the following optimization problem:

Let $T > 0$ denote the time duration of the process. Moreover, let $F > 0$ and denote by $\Xi := \left\{\Phi \in L^\infty((0,T);\mathbb{R}_+^{n+1}), \quad \|\Phi\|_{L^\infty} \leq F\right\}$ the set of admissible external fluxes profiles. For each profile $\Phi := (\phi_0, \cdots, \phi_n) \in \Xi$, denote by $e_\Phi : t \in [0,T] \mapsto e_0 + \int_0^t \sum_{i=0}^n \phi_i(s)\, ds$ the time-dependent thickness of the film, and by v_Φ a solution to (1.62) associated with Φ. Let $e_{\text{opt}} > e_0$ and $v_{\text{opt}} \in L^2((0,1);\overline{\mathcal{D}})$ denote respectively the target thickness and the target final concentration profiles for the different chemical species and consider the cost function $\mathcal{J} : \Xi \to \mathbb{R}$ defined by

$$\forall \Phi \in \Xi, \quad \mathcal{J}(\Phi) := |e_\Phi(T) - e_{\text{opt}}|^2 + \|v_\Phi(T,\cdot) - v_{\text{opt}}\|^2_{L^2(0,1)}. \tag{1.64}$$

The optimization problem of interest reads

$$\Phi^* \in \underset{\Phi \in \Xi}{\operatorname{argmin}}\, \mathcal{J}(\Phi). \tag{1.65}$$

If we assume that for any $\Phi \in \Xi$ there exists a unique global weak solution v_Φ to system (1.62), then $\mathcal{J}$ is well-defined and there exists a minimizer $\Phi^* \in \Xi$ to (1.65). The proof of this result is detailed in Section 2.4.4 of Chapter 2.

Numerical Results

From a numerical point of view, we propose a fully implicit unconditionally stable scheme for the discretization of the system (1.62) and an iterative procedure based on an adjoint formulation associated to the discretization scheme for the optimization problem.

As part of the collaboration work with IRDEP lab, we also propose a few practical improvements for the model (1.61) taking into account the temperature evolution of the system and the surface absorption rates of the different chemical species. Then, we calibrate the adapted model on experimental measures. Details of this work along with some numerical results are presented in Chapter 3.

1.7 Appendix: Brief Introduction to Gradient Flows

We give in this appendix a very short introduction to the theory of gradient flows in metric spaces. For the reader's convenience, we first recall in Section 1.7.1 some basic notions of metric spaces that are essential to the remaining sections. We present in Section 1.7.2 the well-known case of Euclidian spaces. Then, we report in 1.7.3 three characterizations for gradient flows in metric spaces generalizing the properties satisfied in the Euclidian case. We mention that all the notions an results gathered in this appendix are extracted from [LAS08, San17].

1.7.1 Basic Notions in Metric Spaces

Let X be metric space endowed with a distance d. A curve $\gamma : [0,1] \to X$ is a continuous function defined on $[0,1]$ and valued in the considered metric space (X,d). Note that the derivative of a curve $\gamma'(t)$ can be defined only if X is a vector space. Nevertheless, one can define the modulus $|\gamma'|(t)$ instead .

Definition 1.9 (Metric derivative, [LAS08]-1.1)**.** *The* metric derivative *of a curve $\gamma : [0,1] \to X$ at time t, is denoted by $|\gamma'|(t)$ and defined as*

$$|\gamma'|(t) := \lim_{h\to 0} \frac{d(\gamma(t+h),\gamma(t))}{|h|},$$

provided this limit exists.

Definition 1.10 (Absolute continuous curve [LAS08]-1.1)**.** *A curve $\gamma : [0,1] \to X$ is said to be absolutely continuous whenever there exists a function $g \in L^1([0,1];\mathbb{R})$ such that $d(\gamma(t_0),\gamma(t_1)) \leq \int_{t_0}^{t_1} g(s)\mathrm{d}s$ for every $0 \leq t_0 < t_1 \leq 1$. The set of absolutely continuous curves defined on $[0,1]$ and valued in X is denoted by* $\mathrm{AC}(X)$.

The length of an absolute continuous curve γ is denoted by $\mathrm{Length}(\gamma)$ and defined as follows:

Definition 1.11 (Length of a curve [LAS08]-1.1)**.** *For a curve $\gamma : [0,1] \to X$,*

$$\mathrm{Length}(\gamma) := \sup\left\{\sum_{k=0}^{n-1} d(\gamma(t_k),\gamma(t_{k+1})) \,:\, n \geq 1,\, 0 = t_0 < t_1 < \cdots < t_n = 1\right\}.$$

Some notions involving geodesics are gathered in the following definition:

Definition 1.12 (Geodesics [LAS08]-1.1)**.** *A curve $\gamma : [0,1] \to X$ is said to be a* geodesic *between y_0 and $y_1 \in X$ if $\gamma(0) = y_0$, $\gamma(1) = y_1$ and*

$$\mathrm{Length}(\gamma) = \min\{\mathrm{Length}(\omega) \,:\, \omega(0) = y_0,\, \omega(1) = y_1\}.$$

A space (X,d) is said to be a length space *if for every y and z we have*

$$d(y,z) = \inf\{\mathrm{Length}(\gamma) \,:\, \gamma \in \mathrm{AC}(X),\, \gamma(0) = y,\, \gamma(1) = z\}.$$

A space (X,d) is said to be a geodesic space *if for every y and z we have*

$$d(y,z) = \min\{\mathrm{Length}(\gamma) \,:\, \gamma \in \mathrm{AC}(X),\, \gamma(0) = y,\, \gamma(1) = z\},$$

In a length space, a curve $\gamma : [0,1] \to X$ is said to be a constant-speed geodesic *between $\gamma(0)$ and $\gamma(1) \in X$ if it satisfies*

$$d(\gamma(t),\gamma(s)) = \frac{|t-s|}{t_1-t_0} d(\gamma(t_0),\gamma(t_1)) \quad \text{for all } t,s \in [t_0,t_1].$$

1.7.2 Gradient flows in Euclidean spaces

Let $n \in \mathbb{N}^*$ and let us consider the Euclidean space $\mathbb{R}^n$ endowed with the standard Euclidian metric. Let $\mathcal{E} : \mathbb{R}^n \to \mathbb{R}$ be a differentiable functional defined on $\mathbb{R}^n$ and let $u_0 \in \mathbb{R}^n$ and $T > 0$. We consider the following Cauchy problem

$$\begin{cases} u'(t) = -\nabla \mathcal{E}(u(t)) & \text{for } t > 0, \\ u(0) = u_0. \end{cases} \tag{1.66}$$

In virtue of the Cauchy Lipschitz theorem, the classical Cauchy problem (1.66) admits a unique solution if ∇F is Lipschitz continuous. A definition of a *gradient flow* in this case is simply given by

Definition 1.13 (Gradient flow as solution of an ODE)**.** *We call a (the)* gradient flow *associated to $\mathcal{E}$, a (the) solution to the Cauchy problem (1.66). In other words, it is the curve $u : [0, T] \to \mathbb{R}^n$ starting at time $t = 0$ from a point u_0, which moves along the* steepest descent *direction.*

Let us now relax the differentiability assumption and replace the classical gradient $\nabla \mathcal{E}$ by the sub-differtiential of $\mathcal{E}$ denoted $\partial \mathcal{E}$ and defined as follows: for every $y \in \mathbb{R}^n$,

$$\partial \mathcal{E}(y) := \{p \in \mathbb{R}^n \ : \ \mathcal{E}(z) \geq \mathcal{E}(y) + p \cdot (z - y) \ \text{ for all } z \in \mathbb{R}^n\}. \tag{1.67}$$

Consider, instead of the classical Cauchy problem (1.66), the following differential inclusion: search for an absolutely continuous curve $u : [0, T] \to \mathbb{R}^n$ such that

$$\begin{cases} u'(t) \in -\partial \mathcal{E}(u(t)) & \text{for a.e. } t > 0, \\ u(0) = u_0, \end{cases} \tag{1.68}$$

In this case, existence and uniqueness of a solution to (1.68) can be shown under *convexity* assumptions on F. For instance, when $\mathcal{E}$ is supposed to be *convex*. Indeed, if we consider two solutions y_1 and y_2 of (1.68), it suffices to differentiate the quantity $\frac{1}{2}|y_1(t) - y_2(t)|^2$ with respect to t and use the convexity of $\mathcal{E}$ to obtain $|y_1(t) - y_2(t)| \leq |y_1(0) - y_2(0)|$ for every time $t \in [0, T]$ which implies in particular the uniqueness of the solution [San17]. A second, more general case is when $\mathcal{E}$ is assumed to be λ-*convex* for some $\lambda \in \mathbb{R}$. We recall that $\mathcal{E}$ is said to be λ-convex if the function $\mathbb{R}^n \ni y \mapsto \mathcal{E}(y) - \frac{\lambda}{2}|y|^2$ is convex. Also in this case and using the same arguments, one can deduce uniqueness of the solution to (1.68). [San17]. Then, we can immediately extend the Definition 1.13 of gradient flows to the differential inclusion (1.68).

Let us now present (at a formal level) three properties satisfied by gradient flows in the sense of Definition 1.13. The reader may refer to [San17, LAS08] for rigorous justification of the calculations. The interest of these properties is that they involve quantities that have counterparts in metric spaces. The generalization of these properties serves as characterizations of the notion of gradient flows in metric spaces.

Minimizing Movement

Let us fix a small time step $\tau > 0$ and look for a sequence $(u_k^\tau)_{k \in \mathbb{N}^*}$ defined through the iterative *Minimizing Movement* scheme:

$$u_{k+1}^\tau \in \operatorname{argmin}_{u \in \mathbb{R}^n} \left[\mathcal{E}(u) + \frac{|u - u_k^\tau|^2}{2\tau} \right]. \tag{1.69}$$

It results, in particular, from the first optimality condition that for every $k \in \mathbb{N}$,

$$u_{k+1}^\tau \in \operatorname{argmin}\left[\mathcal{E}(u) + \frac{|u-u_k^\tau|^2}{2\tau}\right] \quad \Rightarrow \quad \nabla\mathcal{E}(u_{k+1}^\tau) = -\frac{u_{k+1}^\tau - u_k^\tau}{\tau},$$

which is equivalent to discretize the Cauchy problem (1.66) using an *Euler scheme*. Thus, one can interpret the sequence $(u_k^\tau)_{k\in\mathbb{N}^*}$ as the values of the curve $u(t)$ at the discrete times $t = 0, \tau, 2\tau, \dots, k\tau, \dots, T$. This gives a constructive way to obtain the gradient flow u. Moreover, in this case, even weaker assumptions of $\mathcal{E}$ allow to show the existence of solutions for small enough τ. It suffices for example to suppose that $\mathcal{E}$ is lower semi-continuous and is lower bounded : for every $y \in \mathbb{R}^n$, $\mathcal{E}(y) \geq C_1 - C_2|y|^2$ for some $C_1, C_2 \in \mathbb{R}$. See [San17].

Energy Dissipation Equality

Let $\mathcal{E} : \mathbb{R}^n \to \mathbb{R}$ be a differentiable functional defined on $\mathbb{R}^n$ and let $u : [0,T] \to \mathbb{R}^n$ be a differentiable curve. For every $t \in [0,T]$, we have $\frac{d}{dt}\mathcal{E}(u(t)) = u'(t)\nabla F(u(t))$. Thus, for any $0 \leq s < t \leq T$, the following holds [7]

$$\begin{aligned}
\mathcal{E}(u(s)) - \mathcal{E}(u(t)) &= -\left(\mathcal{E}(u(t)) - \mathcal{E}(u(s))\right) \\
&= \int_s^t -\nabla\mathcal{E}(u(r)) \cdot u'(r)\, \mathrm{d}r \\
&\leq \int_s^t |\nabla\mathcal{E}(u(r))||u'(r)|\, \mathrm{d}r \\
&\leq \int_s^t \left(\frac{1}{2}|u'(r)|^2 + \frac{1}{2}|\nabla\mathcal{E}(u(r))|^2\right) \mathrm{d}r.
\end{aligned}$$

Note that the first inequality is an equality if and only if there exists $\alpha < 0$ such that $u'(r) = -\alpha\nabla\mathcal{E}(u(r))$ for almost every r, and the second inequality is an equality if and only if $|\nabla\mathcal{E}(u(r))| = |u'(r)|$ for almost every r. Consequently, when u is a solution to the Cauchy problem (1.66), the following condition, called EDE (*Energy Dissipation Equality*), is satisfied.

$$\mathcal{E}(u(s)) - \mathcal{E}(u(t)) \leq \int_s^t \left(\frac{1}{2}|u'(r)|^2 + \frac{1}{2}|\nabla\mathcal{E}(u(r))|^2\right) \mathrm{d}r, \quad \forall 0 \leq s < t \leq T. \tag{1.70}$$

Evolution Variational Inequality

Let $\mathcal{E} : \mathbb{R}^n \to \mathbb{R}$ be a functional defined on $\mathbb{R}^n$ (not necessarily differentiable) and let $u : [0,T] \to \mathbb{R}^n$ be differentiable curve. Let $\lambda \in \mathbb{R}$ and assume that $\mathcal{E}$ is λ-convex. In the one hand, for any $y \in \mathbb{R}^n$, from the definition of the sub-differential of $\mathcal{E}(y)$, the following holds for every $p \in \partial\mathcal{E}(y)$,

$$\mathcal{E}(z) \geq \mathcal{E}(y) + \frac{\lambda}{2}|y-z|^2 + p \cdot (z-y) \quad \text{for all } z \in \mathbb{R}^n,$$

which implies

$$p \cdot (z-y) \leq \mathcal{E}(z) - \mathcal{E}(y) - \frac{\lambda}{2}|y-z|^2 \quad \text{for all } z \in \mathbb{R}^n. \tag{1.71}$$

[7] Young inequality is used in the last line : $2ab \leq a^2 + b^2$, for any $a, b \in \mathbb{R}_+$ which is an equality if $a = b$.

In the other hand, the following holds for any curve $u(t)$ and any vector $z \in \mathbb{R}^n$,

$$\frac{d}{dt}\frac{1}{2}|u(t)-z|^2 = (u'(t))\cdot(u(t)-z) = (z-u(t))\cdot(-u'(t)). \tag{1.72}$$

As a result, if u is a solution to the differential inclusion Cauchy problem (1.68), then (1.71) and (1.72) imply the following property, called *Evolution Variational Inequality*,

$$\frac{d}{dt}\frac{1}{2}|u(t)-z|^2 \leq \mathcal{E}(z) - \mathcal{E}(u(t)) - \frac{\lambda}{2}|u(t)-z|^2, \quad \forall z \in \mathbb{R}^n \tag{1.73}$$

1.7.3 Gradient Flows in Metric Spaces

Let (X,d) be a metric space endowed with a distance d. Let us consider a functional $\mathcal{F}: X \to \mathbb{R}\cup\{+\infty\}$ defined on X and a point $u_0 \in X$. We would like to give a suitable sense to the following formal Cauchy problem

$$\begin{cases} u'(t) = -\nabla\mathcal{E}(u(t)) & \text{for } t>0, \\ u(0) = u_0. \end{cases} \tag{1.74}$$

whose solution $u:[0,T]\to X$ is a curve valued in the metric space X. Note first, that unless X is a vector space (which is not assumed to be true in general) the classical notion of the gradient $\nabla\mathcal{E}$ must be adapted. We call *upper gradient* every function $g: X \to \mathbb{R}$ such that, for every Lipschitz curve $u:[0,1]\to\mathbb{R}$, we have

$$|\mathcal{E}(u(0)) - \mathcal{E}(u(1))| \leq \int_0^1 g(u(t))|u'|(t)\mathrm{d}t.$$

A suitable choice of the upper gradient which is adapted to lower semi-continuous functionals is the *descending slope*[8] proposed in [LAS08, San17] and abusively denoted by $\nabla\mathcal{E}$:

$$\nabla\mathcal{E}(u) := \limsup_{z\to u} \frac{[\mathcal{E}(u)-\mathcal{E}(z)]_+}{d(u,z)} \tag{1.75}$$

Let us now give a suitable generalization to the notion of $\lambda-$convexity in order to be able to reproduce the arguments of the Euclidian case. The appropriate notion in metric spaces is the *geodesic convexity* which can only be defined in a geodesic metric spaces. On such a space, we have the following definition:

Definition 1.14 (Geodesic convexity [LM13]). *Let $\lambda \in \mathbb{R}$. A functional $\mathcal{E}: X \to \mathbb{R}\cup\{+\infty\}$ is said to be* geodesically λ-convex *with respect to the metric d if an only if : for every geodesic $\gamma:[t_0,t_1]\to X$ with constant speed and every $\theta\in[0,1]$, the following holds*

$$\mathcal{E}(\gamma((1-\theta)t_0+\theta t_1)) \leq (1-\theta)\mathcal{E}(\gamma(t_0)) + \theta\mathcal{E}(\gamma(t_1)) - \lambda\frac{\theta(1-\theta)}{2}d^2(t_0,t_1). \tag{1.76}$$

where $0 \leq t_0 < t_1 \leq T$.

[8] Other choices are possible if we assume more regularity on $\mathcal{E}$. For instance, if the functional $\mathcal{E}$ is assumed to be Lipschitz continuous, then a possible choice fo the upper gradient is the *local Lipschitz constant* [San17]. Nevertheless, the notion of descending slope offers the "most" general framework since the only assumption on $\mathcal{E}$ is a lower semi continuity.

We have now all the ingredients needed to "define" or characterize the notion of gradient flows in metric spaces. In the theory, mainly developed in [LAS08], three main characterizations are proposed to define a gradient flow in a metric space : gradient flow as a limit (when the time step goes to 0) of the discrete solutions to a minimizing movement scheme of the form (1.69), gradient flow as a curve satisfying the Energy Dissipation Equality (1.70), gradient flow as a curve satisfying the Evolutional Variational Inequality (1.73). Let us give a brief description and comments for each case.

Gradient Flow as a Generalized Movement Scheme

Let $\mathcal{E} : X \to \mathbb{R} \cup \{+\infty\}$ be a functional defined on the metric space (X, d) and assume that $\mathcal{E}$ is lower semi-continuous. Let $u_0 \in X$ and consider the iterative problem

$$u_{k+1}^\tau \in \operatorname{argmin}_{u \in X} \left[\mathcal{E}(u) + \frac{d(u, u_k^\tau)^2}{2\tau} \right] \tag{1.77}$$

together with the piecewise constant interpolation

$$u^\tau(t) := u_k^\tau \quad \text{for every } t \in](k-1)\tau, k\tau]. \tag{1.78}$$

This approximation scheme was introduced by De Giorgi [DG93] as a generalization of the minimizing movement (1.69). In particular, the limit of u^τ (when the time step τ goes to 0) is shown [LAS08, San17] to solve the gradient system (1.74). Thus, a first definition of gradient flows in metric spaces is given thorough the *Generalized Minimizing Movement* (1.77) as follows:

Definition 1.15 (Gradient Flow in the GMM sense, [San17, DG93])**.** *Let $\mathcal{E} : X \to \mathbb{R} \cup \{+\infty\}$ be a lower semi-continuous functional on the metric space (X, d). A curve $u : [0, T] \to X$ is called* Generalized Minimizing Movements *(GMM) associate to $\mathcal{E}$ if there exists a sequence of time steps $\tau_j \to 0$ such that the sequence of curves u^{τ_j} defined in (1.78) using the iterated solutions of (1.77) uniformly converges to u in $[0, T]$. In this case we say that u is a* gradient flow *associated to $\mathcal{E}$.*

Gradient Flow in the EDE Sense

Let $\mathcal{E} : X \to \mathbb{R} \cup \{+\infty\}$ be a functional defined on the metric space (X, d) and assume that $\mathcal{E}$ is lower semi-continuous and let $u_0 \in X$. Consider the formal Cauchy problem (1.74) where $\nabla\mathcal{E}$ is defined by the descending slope (1.75). A second definition of gradient flows in metric spaces can be obtained from the formal equality (that was shown to be satisfied in the Euclidian case): if $u : [0, T] \to X$ solves (1.74) then

$$\mathcal{E}(u(s)) - \mathcal{E}(u(t)) \leq \int_s^t \left(\frac{1}{2}|u'(r)|^2 + \frac{1}{2}|\nabla\mathcal{E}(u(r))|^2 \right) \mathrm{d}r, \quad \forall 0 \leq s < t \leq T. \tag{1.79}$$

Definition 1.16 (Gradient Flow in the EDI sense)**.** *Let $\mathcal{E} : X \to \mathbb{R} \cup \{+\infty\}$ be a lower semi-continuous functional on the metric space (X, d). A curve $u : [0, T] \to X$ is called* gradient flow in the EDE sense *starting at $u_0 \in X$ if $u \in AC(X)$ and u satisfies the EDE property (1.79) with $|\nabla F|$ defined in (1.75).*

Gradient Flow in the EVI Sense

If we assume in addition that the entropy is geodesically $\lambda-$convex for some $\lambda \in \mathbb{R}$, then we can give an additional characterization to the associated gradient flow in terms of the foraml inequality (that was shown to be satisfied in the Euclidian case): if $u : [0, T] \to X$ solves (1.74) then

$$\frac{d}{dt}\frac{1}{2}|u(t) - z|^2 \leq \mathcal{E}(z) - \mathcal{E}(u(t)) - \frac{\lambda}{2}|u(t) - z|^2, \quad \forall z \in \mathbb{R}^n \tag{1.80}$$

Definition 1.17 (Gradient Flow in the EVI sense,)**.** *Let $\mathcal{E} : X \to \mathbb{R} \cup \{+\infty\}$ be a lower semi-continuous functional on the metric space (X, d). A curve $u : [0, T] \to X$ is called* gradient flow in the EDE sense *starting at $u_0 \in X$ if $y \in AC(X)$ and u satisfies the EVI property (1.80) with $|\nabla F|$ defined as in (1.75).*

CHAPTER 2

CROSS-DIFFUSION SYSTEMS IN A MOVING DOMAIN

We report in this chapter the results of [BE16] obtained with Virginie Ehrlacher.

Abstract. We propose and analyze a one-dimensional multi-species cross-diffusion system with non-zero-flux boundary conditions on a moving domain, motivated by the modeling of a Physical Vapor Deposition process. Using the boundedness by entropy method introduced and developped in [BDFPS10, Jue15a], we prove the existence of a global weak solution to the obtained system. In addition, existence of a solution to an optimization problem defined on the fluxes is established under the assumption that the solution to the considered cross-diffusion system is unique. Lastly, we prove that in the case when the imposed external fluxes are constant and positive and the entropy density is defined as a classical logarithmic entropy, the concentrations of the different species converge in the long-time limit to constant profiles at a rate inversely proportional to time. These theoretical results are illustrated by numerical tests.

Contents

2.1 Introduction

The aim of this work is to propose and analyze a mathematical model for the description of a Physical Vapor Deposition (PVD) process, the different steps of which are described in details for instance in [Mat10]. Such a technique is used in several contexts, for instance for the fabrication of thin film crystalline solar cells. The procedure works as follows: a wafer is introduced in a hot chamber where several chemical elements are injected under a gaseous form. As the latter deposit on the substrate, an heterogeneous solid layer grows upon it. Two main phenomena have to be taken into account: the first is naturally the evolution of the surface of the film; the second is the diffusion of the various species in the bulk, due to the high temperature conditions. Experimentalists are interested in controlling the external gas fluxes that are injected into the chamber, so that, at the end of the process, the spatial distributions of the concentrations of the diverse components inside the new layer are as close as possible to target profiles.

In this article, a one-dimensional model which takes into account these two factors is studied. We see this work as a preliminary step before tackling more challenging models in higher dimensions, including surfacic diffusion effects for instance. This will be the object of future work. Our main motivation for the study of such a model concerns the optimization of the external fluxes injected in the chamber during a PVD process.

More precisely, let us assume that at a time $t \geq 0$, the solid layer is composed of $n+1$ different chemical species and occupies a domain of the form $(0, e(t)) \subset \mathbb{R}_+$, where $e(t) > 0$ denotes the thickness of the film. The evolution of $e(t)$ is determined by the fluxes of atoms that are absorbed at the surface of the layer. At time $t > 0$ and point $x \in (0, e(t))$, the local volumic fractions of the different species are denoted respectively by $u_0(t,x), \cdots, u_n(t,x)$. Let us point out that if the molar volume of the solid is uniform in the thin film layer and constant during all the process, then $u_i(t,x)$ is also equal (up to a multiplicative constant) to the local concentration of the i^{th} species at time $t > 0$ and point $0 \leq x \leq e(t)$. Up to some renormalization condition, it is natural to expect that these functions are non-negative and satisfy a volumic constraint which reads as follows:

$$\forall 0 \leq i \leq n, \quad u_i(t,x) \geq 0 \text{ and } \sum_{i=0}^{n} u_i(t,x) = 1. \tag{2.1}$$

Because of the constraint (2.1), it holds that $u_0(t,x) = 1 - \sum_{i=1}^{n} u_i(t,x)$ for all $t > 0$ and $x \in (0, e(t))$. Thus, the knowledge of the n functions $u_1, \cdots, u_n$ is enough to determine the dynamics of the whole system. Replacing u_0 by $1 - \sum_{i=1}^{n} u_i$, and denoting by u the vector-valued function $(u_1, \cdots, u_n)$, the evolution of the concentrations inside the bulk of the solid layer is modeled through a system of cross-diffusion equations of the form

$$\partial_t u - \partial_x \left(A(u) \partial_x u \right) = 0, \quad \text{for } t > 0,\ x \in (0, e(t)), \tag{2.2}$$

with approriate boundary and initial conditions, where $A : [0,1]^n \to \mathbb{R}^{n\times n}$ is a matrix-valued function encoding the cross-diffusion properties of the different species.

Such systems have received much attention from the mathematical community in the case when no-flux boundary conditions are imposed on a fixed domain [LS67, Ama89, LN06, GR10]. Then, in arbitrary dimension $d \in \mathbb{N}^*$, the system reads

$$\partial_t u - \mathrm{div}_x \left(A(u)\nabla_x u\right) = 0, \qquad \text{for } t > 0,\ x \in \Omega,$$

for some fixed bounded regular domain $\Omega \subset \mathbb{R}^d$ and boundary conditions

$$\left(A(u)\nabla_x u\right) \cdot \mathbf{n} = 0 \text{ on } \partial\Omega \text{ and } t \geq 0,$$

where $\mathbf{n}$ denotes the outward normal unit vector to Ω.

Such systems appear naturally in the study of population's dynamics in biology, and in chemistry, for the study of the evolution of chemical species concentrations in a given environment [Pai09, HP09]. The analysis of these systems is a challenging task from a mathematical point of view [LPR12, Ali79, Kue96, Red89, CJ04, CJ06, DFR08]. Indeed, the obtained system of parabolic partial differential equations may be degenerate and the diffusion matrix A is in general not symmetric and/or not positive definite. Besides, in general, no maximum principle can be proved for such systems. Nice counterexamples are given in [SJ95]: there exist Hölder continuous solutions to certain cross-diffusion systems which are not bounded, and there exist bounded weak solutions which develop singularities in finite time.

It appears that some of these cross-diffusion systems have a formal gradient flow structure. Recently, an elegant idea, which consists in introducing an entropy density that appears to be a Lyapunov functional for these systems, has been introduced by Burger et al. in [BDFPS10]. This analysis strategy, which was later extended by Jüngel in [Jue15a] and named *boundedness by entropy* technique, enables to obtain the existence of global in time weak solutions satisfying (2.1) under suitable assumptions on the diffusion matrix A. It was successfully applied in several contexts (see for instance [JS13, JS12, ZJ15, JZ14]).

However, there are very few works which focus on the analysis of such cross-diffusion systems with non zero-flux boundary conditions and moving domains. To our knowledge, only systems containing at most two different species have been studied, so that $n = 1$ and the evolution of the concentrations inside the domain are decoupled and follow independent linear heat equations [PP08].

The one-dimensional model (2.2) we propose and analyze in this paper describes the evolution of the concentration of $n+1$ different atomic species, with external flux boundary conditions, in the case when the diffusion matrix A satisfies similar assumptions to those needed in the no-flux boundary conditions case studied in [Jue15a].

The article is organized as follows: the results of [Jue15a] in the case of no-flux boundary conditions in arbitrary dimension are recalled in Section 2.2. We illustrate them on a prototypical example of diffusion matrix A, which is introduced in Section 2.2.1.

Our results in the case of a one-dimensional moving domain with non-zero flux boundary conditions are gathered in Section 2.3. We prove the existence of a global in time weak solution to (2.2) with appropriate boundary conditions and evolution law

for $e(t)$ in Section 2.3.2. The long time behaviour of a solution is analyzed in the case of constant external absorbed fluxes in 2.3.2 and an optimization problem is studied in 2.3.2. The proofs of these results are gathered in Section 5.3.

A numerical scheme used to approximate the solution of such systems is described in Section 5.4 and our theoretical results will be illustrated by several numerical tests.

2.2 Case of no-flux boundary conditions in arbitrary dimension

In Section 2.2.1, a particular cross-diffusion model on a fixed domain with no-flux boundary conditions is presented. The latter is a prototyical example of the systems of equations considered in this paper. Its formal gradient flow structure is highlighted in Section 2.2.1. Using slight extensions of results of [ZJ15, JZ14], it can be seen that this system can be analyzed using the theoretical framework developped in [Jue15a, BDFPS10], which is recalled in Section 2.2.2.

Throughout this section, let us denote by $d \in \mathbb{N}^*$ the space dimension, $\Omega \subset \mathbb{R}^d$ the regular bounded domain occupied by the solid. The local concentrations at time $t > 0$ and position $x \in \Omega$ of the $n+1$ different atomic species entering in the composition of the material are respectively denoted by $u_0(t,x), \cdots, u_n(t,x)$. We also denote by $\mathbf{n}$ the normal unit vector pointing outwards the domain Ω.

2.2.1 Example of cross-diffusion system

Presentation of the model

As mentioned above, we have one particular example of system of cross-diffusion equations in mind, which is used to illustrate more general theoretical results. This system, with no-flux boundary conditions, reads as follows : for any $0 \leq i \leq n$,

$$\begin{cases} \partial_t u_i - \operatorname{div}_x \left(\displaystyle\sum_{0 \leq j \neq i \leq n} K_{ij}(u_j \nabla_x u_i - u_i \nabla_x u_j) \right) = 0, & \text{for } (t,x) \in \mathbb{R}_+^* \times \Omega, \\ \left(\displaystyle\sum_{0 \leq j \neq i \leq n} K_{ij}(u_j \nabla_x u_i - u_i \nabla_x u_j) \right) \cdot \mathbf{n} = 0, & \text{for } (t,x) \in \mathbb{R}_+^* \times \partial\Omega, \end{cases} \tag{2.3}$$

where for all $0 \leq i \neq j \leq n$, the positive real numbers K_{ij} satisfy $K_{ij} = K_{ji} > 0$. They represent the cross-diffusion coefficients of atoms of type i with atoms of type j. This set of equations can be formally derived from a discrete stochastic lattice hopping model, which is detailed in the Appendix.

The initial condition $(u_0^0, \cdots, u_n^0) \in L^1(\Omega; \mathbb{R}^{n+1})$ of this system is assumed to satisfy:

$$\forall 0 \leq i \leq n, \quad u_i^0(x) \geq 0, \quad \sum_{i=0}^{n} u_i^0(x) = 1 \text{ and } u_i(0,x) = u_i^0(x) \quad \text{a.e. in } \Omega. \tag{2.4}$$

The relationship $\sum_{i=0}^n u_i^0(x) = 1$ is a natural volumic constraint which encodes the fact that each site of the crystalline lattice of the solid has to be occupied (vacancies being treated as a particular type of atomic species).

Summing up the $n+1$ equations of (2.3), we observe that a solution $(u_0, \cdots, u_n)$ must necessarily satisfy $\partial_t \left(\sum_{i=0}^n u_i\right) = 0$. It is thus expected that the following relationship should hold:

$$\forall 0 \leq i \leq n, \quad u_i(t,x) \geq 0, \quad \sum_{i=0}^{n} u_i(t,x) = 1, \quad \text{a.e. in } \mathbb{R}_+^* \times \Omega. \tag{2.5}$$

Plugging the expression $u_0(t,x) = 1 - \sum_{i=1}^n u_i(t,x)$ in (2.3), it holds that for all $1 \leq i \leq n$,

$$\begin{aligned}
0 &= \partial_t u_i - \operatorname{div}_x \left[\sum_{1 \leq j \neq i \leq n} K_{ij} \left(u_j \nabla_x u_i - u_i \nabla_x u_j \right) \right] \\
&\quad - \operatorname{div}_x \left[K_{i0} \left(\left(1 - \sum_{1 \leq j \neq i \leq n} u_j - u_i \right) \nabla_x u_i - u_i \nabla_x \left(1 - \sum_{1 \leq j \neq i \leq n} u_j - u_i \right) \right) \right] \\
&= \partial_t u_i - \operatorname{div}_x \left[\sum_{1 \leq j \neq i \leq n} (K_{ij} - K_{i0}) \left(u_j \nabla_x u_i - u_i \nabla_x u_j \right) + K_{i0} \nabla_x u_i \right].
\end{aligned}$$

Thus, the system can be rewritten as a function of $u := (u_1, \cdots, u_n)^T$ as follows

$$\begin{cases}
\partial_t u - \operatorname{div}_x \left(A(u) \nabla_x u \right) = 0, & \text{for } (t,x) \in \mathbb{R}_+^* \times \Omega, \\
\left(A(u) \nabla_x u \right) \cdot \mathbf{n} = 0, & \text{for } (t,x) \in \mathbb{R}_+^* \times \partial\Omega, \\
u(0,x) = u^0(x), & \text{for } x \in \Omega,
\end{cases} \tag{2.6}$$

where $u^0 := (u_1^0, \cdots, u_n^0)^T$ and the matrix-valued application

$$A : \begin{cases} [0,1]^n & \to & \mathbb{R}^{n \times n} \\ u := (u_i)_{1 \leq i \leq n} & \mapsto & (A_{ij}(u))_{1 \leq i,j \leq n} \end{cases}$$

is defined by

$$\begin{cases}
\forall 1 \leq i \leq n, \quad A_{ii}(u) = \displaystyle\sum_{1 \leq j \neq i \leq n} (K_{ij} - K_{i0}) u_j + K_{i0}, \\
\forall 1 \leq i \neq j \leq n, \quad A_{ij}(u) = -(K_{ij} - K_{i0}) u_i.
\end{cases} \tag{2.7}$$

Despite their importance in chemistry or biology, it appears that the mathematical analysis of systems of the form (2.6), taking into account constraints (2.5), is quite recent [BDFPS10, GR10, Jue15a, LM13]. Let us point out here that the non-negativity of the solutions to (2.6) through time is a mathematical issue, linked to the absence of a maximum principle for such systems.

At least up to our knowledge, the first proof of existence of global weak solutions of (2.6) satisfying constraints (2.5) with non-identical cross-diffusion coefficients is given in [BDFPS10] for $n = 2$ with coefficients K_{ij} such that $K_{i0} > 0$ for $i = 1, 2$ and $K_{12} = K_{21} = 0$. These results were later extended in [JZ14] to a general number of species $n \in \mathbb{N}^*$ with cross-diffusion coefficients satisfying $K_{i0} > 0$ and $K_{ij} = 0$ for all $1 \leq i \neq j \leq n$; the authors of the latter article proved in addition the uniqueness

of such weak solutions. In [ZJ15], the case $n = 2$ with arbitrary positive coefficients $K_{ij} > 0$ is covered, though no uniqueness result is provided. The main difficulty of the mathematical analysis of such equations relies in the bounds (2.5), which are not obvious since no maximum principle can be proved for these systems in general. In all the articles mentioned above, the analysis framework used by the authors is the so-called *boundedness by entropy method*. The main idea of this technique is to write the above system of equations as a formal gradient flow and derive estimates on the solutions $(u_0, \cdots, u_n)$ using the decay of some well-chosen entropy functional. We present in Section 2.2.1 the formal gradient flow structure of (2.6) and recall the results of [Jue15a] in Section 2.2.2.

Remark 2.1. *This model is linked to the so-called* Stefan-Maxwell *model, studied in [JS13, BGS12]. Indeed, the model considered in the latter paper reads*

$$\begin{cases} \partial_t u - \operatorname{div}_x \left(A(u)^{-1} \nabla_x u \right) = 0, & \text{for } (t,x) \in (0,T] \times \Omega, \\ (A(u)\nabla_x u) \cdot \boldsymbol{n} = 0, & \text{for } (t,x) \in (0,T] \times \partial\Omega, \\ u(0,x) = u^0(x), & \text{for } x \in \Omega, \end{cases} \tag{2.8}$$

where A is defined by (2.7).

Formal gradient flow structure of (2.6)

We detail in this section the formal gradient flow structure of the system (2.6).

Let $\mathcal{D} \subset \mathbb{R}^n$ be defined by

$$\mathcal{D} := \left\{ (u_1, \cdots, u_n) \in (\mathbb{R}_+^*)^n, \quad \sum_{i=1}^n u_i < 1 \right\} \subset (0,1)^n. \tag{2.9}$$

Let us introduce the classical *entropy density* h (see for instance [BDFPS10], [Jue15a], [JZ14] and [LM13])

$$h : \begin{cases} \overline{\mathcal{D}} & \longrightarrow & \mathbb{R} \\ u := (u_i)_{1 \le i \le n} & \longmapsto & h(u) = \displaystyle\sum_{i=1}^n u_i \log u_i + (1 - \rho_u) \log(1 - \rho_u), \end{cases} \tag{2.10}$$

where $\rho_u := \sum_{i=1}^n u_i$. Some properties of h can be easily checked:

(P1) the function h belongs to $\mathcal{C}^0(\overline{\mathcal{D}}) \cap \mathcal{C}^2(\mathcal{D})$; consequently, h is bounded on $\overline{\mathcal{D}}$;

(P2) the function h is strictly convex on $\mathcal{D}$;

(P3) its derivative

$$Dh : \begin{cases} \mathcal{D} & \longrightarrow & \mathbb{R}^n \\ (u_i)_{1 \le i \le n} & \mapsto & \left(\log \frac{u_i}{1 - \rho_u} \right)_{1 \le i \le n}, \end{cases}$$

is invertible and its inverse is given by

$$(Dh)^{-1} : \begin{cases} \mathbb{R}^n & \longrightarrow & \mathcal{D} \\ (w_i)_{1 \le i \le n} & \mapsto & \frac{e^{w_i}}{1 + \sum_{j=1}^n e^{w_j}}. \end{cases}$$

In the following, we denote by D^2h the Hessian of h. The *entropy functional* $\mathcal{E}$ is defined by

$$\mathcal{E} : \left\{ \begin{array}{ccc} L^\infty(\Omega;\overline{\mathcal{D}}) & \longrightarrow & \mathbb{R} \\ u & \longmapsto & \mathcal{E}(u) := \int_\Omega h(u(x))\,dx. \end{array} \right. \tag{2.11}$$

Throughout the article, for all $u \in L^\infty(\Omega;\mathcal{D})$, we shall denote by $D\mathcal{E}(u)$ the measurable vector-valued function defined by

$$D\mathcal{E}(u) : \left\{ \begin{array}{ccc} \Omega & \to & \mathbb{R}^n \\ x & \mapsto & Dh(u(x)). \end{array} \right.$$

The system (2.6) can then be formally rewritten under the following gradient flow structure

$$\left\{ \begin{array}{ll} \partial_t u - \mathrm{div}_x\left(M(u)\nabla_x D\mathcal{E}(u)\right) = 0, & \text{for } (t,x) \in \mathbb{R}_+^* \times \Omega, \\ \left(M(u)\nabla_x D\mathcal{E}(u)\right)\cdot \mathbf{n} = 0, & \text{for } (t,x) \in \mathbb{R}_+^* \times \partial\Omega, \\ u(0,x) = u^0(x), & \text{for } x \in \Omega, \end{array} \right. \tag{2.12}$$

where $M : \overline{\mathcal{D}} \to \mathbb{R}^{n\times n}$ is the so-called *mobility matrix* of the system defined for all $u \in \mathcal{D}$ by

$$M(u) := A(u)(D^2h(u))^{-1}.$$

More precisely, it holds that for all $u \in \overline{\mathcal{D}}$, $M(u) = (M_{ij}(u))_{1\le i,j\le n}$ where for all $1 \le i \ne j \le n$,

$$M_{ii}(u) = K_{i0}(1-\rho_u)u_i + \sum_{1\le j\ne i\le n} K_{ij}u_iu_j \quad \text{and} \quad M_{ij}(u) = -K_{ij}u_iu_j. \tag{2.13}$$

2.2.2 Existence of global weak solutions by the boundedness by entropy technique

The formal gradient flow formulation of a system of cross-diffusion equations is a key point in the boundedness by entropy technique. In the example presented in Section 2.2.1, it implies in particular that $\mathcal{E}$ is a Lyapunov functional for the system (2.6) [BDFPS10, Jue15a]. However, the mobility matrix obtained for these systems is not a concave function of the densities, so that standard gradient flow theory arguments (such as the minimizing movement method) cannot be applied in this context [ZM15, DNS09, JKO98, LM13]. However, the existence of a global weak solution to (2.6) can still be proved. Let us recall here a simplified version of Theorem 2 of [Jue15a] which is adapted to our context.

Theorem 2.2 (Theorem 2 of [Jue15a])**.** *Let $\mathcal{D} \subset \mathbb{R}^n$ be the domain defined by (2.9). Let $A : u \in \overline{\mathcal{D}} \mapsto A(u) := (A_{ij}(u))_{1\le i,j\le n} \in \mathbb{R}^{n\times n}$ be a matrix-valued functional defined on $\overline{\mathcal{D}}$ satisfying $A \in \mathcal{C}^0(\overline{\mathcal{D}};\mathbb{R}^{n\times n})$ and the following assumptions:*

(H1) There exists a bounded from below convex function $h \in \mathcal{C}^2(\mathcal{D},\mathbb{R})$ such that its derivative $Dh : \mathcal{D} \to \mathbb{R}^n$ is invertible on $\mathbb{R}^n$;

(H2) There exists $\alpha > 0$, and for all $1 \le i \le n$, there exist $1 \ge m_i > 0$, such that for all $z = (z_1,\cdots,z_n)^T \in \mathbb{R}^n$ and $u = (u_1,\cdots,u_n)^T \in \mathcal{D}$,

$$z^T D^2h(u)A(u)z \ge \alpha \sum_{i=1}^n u_i^{2m_i-2}z_i^2.$$

Let $u^0 \in L^1(\Omega;\mathcal{D})$ so that $w^0 := Dh(u^0) \in L^\infty(\Omega;\mathbb{R}^n)$. Then, there exists a weak solution u with initial condition u^0 to

$$\begin{cases} \partial_t u = \mathrm{div}_x(A(u)\nabla_x u), & \text{for } (t,x) \in \mathbb{R}_+^* \times \Omega, \\ (A(u)\nabla_x u)\cdot \boldsymbol{n} = 0, & \text{for } (t,x) \in \mathbb{R}_+^* \times \partial\Omega, \end{cases} \tag{2.14}$$

such that for almost all $(t,x) \in \mathbb{R}_+^* \times \Omega$, $u(t,x) \in \overline{\mathcal{D}}$ with

$$u \in L^2_{\mathrm{loc}}(\mathbb{R}_+; H^1(\Omega,\mathbb{R}^n)) \text{ and } \partial_t u \in L^2_{\mathrm{loc}}(\mathbb{R}_+; (H^1(\Omega;\mathbb{R}^n))').$$

Lemma 2.3 states that the prototypical example presented in Section 2.2.1 falls into the framework of Theorem 2.2. The proof of the latter is given Section 2.4.1 for the sake of completeness, and relies on ideas introduced in [JZ14].

Lemma 2.3. *Let $\mathcal{D} \subset \mathbb{R}^n$ be the domain defined by (2.9) and $A : u \in \overline{\mathcal{D}} \mapsto A(u) := (A_{ij}(u))_{1\le i,j\le n} \in \mathbb{R}^{n\times n}$ be the matrix-valued function defined by (2.7). Then, $A \in \mathcal{C}^0(\overline{\mathcal{D}};\mathbb{R}^{n\times n})$ and satisfies assumptions (H1)-(H2) of Theorem 2.2, with h given by (2.10), $\alpha = \min_{1\le i\ne j\le n} K_{ij}$ and $m_i = \frac{1}{2}$ for all $1 \le i \le n$.*

The existence of global weak solutions to (2.6) is then a direct consequence of Theorem 2.2 and Lemma 2.3.

Let us point out that the uniqueness of solutions to general systems of the form (2.14) remains an open theoretical question, at least up to our knowledge. It can be obtained in some particular cases. When the diffusion matrix A is defined by (2.7) and when all the diffusion coefficients K_{ij} are identically equal to some constant $K > 0$, the uniqueness of the solution can be trivially obtained since the system boils down to a set of n decoupled heat equation for the evolution of the density of each species.

2.3 Case of non-zero flux boundary conditions and moving domain

In the sequel, we restrict the study to the case when $d = 1$. In this section, we propose a model for the description of a PVD process and present theoretical results whose proofs are postponed to Section 5.3. The global existence of a weak solution is proved. The long-time behaviour of such a solution is studied in the case of constant external fluxes. Lastly, under the assumption that the coefficients K_{ij} are chosen so that there is a unique solution to the system, we prove the existence of a solution to an optimization problem.

2.3.1 Presentation of the model

For the sake of simplicity, we assume that non-zero fluxes are only imposed on the right-hand side of the domain occupied by the solid. At some time $t > 0$, this domain is denoted by $\Omega_t := (0, e(t))$ where $e(t) > 0$ models the thickness of the layer. Initially, we assume that the domain Ω_0 occupied by the solid at time $t = 0$ is the interval $(0, e_0)$ for some initial thickness $e_0 > 0$.

The evolution of the thickness of the film $e(t)$ is determined by the external fluxes of the atomic species that are absorbed at its surface. More precisely, let us assume that

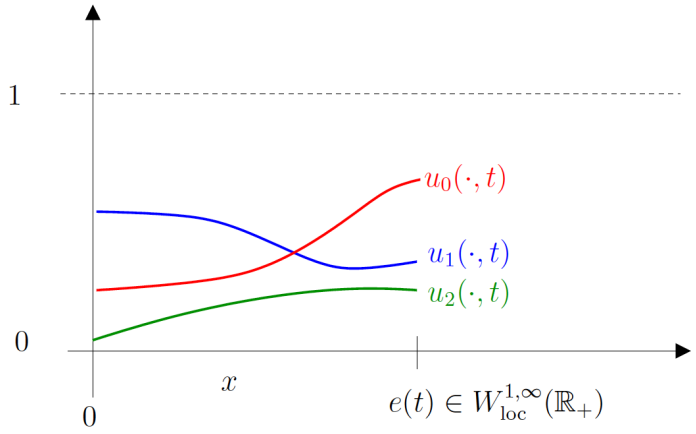


Figure 2.1 – Illustration of the composition of the film layer at time t in the case $n = 2$

there are $n+1$ different chemical species composing the solid layer and let $(\phi_0, \cdots, \phi_n)$ belong to $L^\infty_{\mathrm{loc}}(\mathbb{R}_+; \mathbb{R}_+^{n+1})$. For all $0 \leq i \leq n$, the function $\phi_i(t)$ represents the flux of the species i absorbed at the surface at time $t > 0$ and is assumed to be non-negative. In this one-dimensional model, the evolution of the thickness of the solid is assumed to be given by

$$e(t) := e_0 + \int_0^t \sum_{i=0}^{n} \phi_i(s)\, ds. \tag{2.15}$$

In the following, we will denote by $\varphi := (\phi_1, \cdots, \phi_n)^T$ (see Figure 2.1).

For all $t \geq 0$ and $0 \leq i \leq n$, the local concentration of species i at time t and point $x \in (0, e(t))$ is denoted by $u_i(t,x)$. The evolution of the vector $u := (u_1, \cdots, u_n)$ is given by the system of cross-diffusion equations

$$\partial_t u - \partial_x \left(A(u)\partial_x u\right) = 0, \text{ for } t \in \mathbb{R}_+^*,\ x \in (0, e(t)), \tag{2.16}$$

where $A : \overline{\mathcal{D}} \to \mathbb{R}^{n\times n}$ is a well-chosen diffusion matrix satisfying (H1)-(H2).

We consider that for every $t > 0$, the system satisfies the following conditions on the boundary $\partial\Omega_t$:

$$\left(A(u)\partial_x u\right)(t,0) = 0 \text{ and } \left(A(u)\partial_x u\right)(t,e(t)) + e'(t)u(t,e(t)) = \varphi(t). \tag{2.17}$$

An easy calculation shows that these boundary conditions, in addition to (2.15) and (2.16), ensure that, for all $0 \leq i \leq n$,

$$\frac{d}{dt}\left(\int_{\Omega_t} u_i(t,x)\, dx\right) = \phi_i(t).$$

Indeed, it holds that

$$\begin{aligned}\frac{d}{dt}\left(\int_{\Omega_t} u(t,x)\,dx\right) &= \int_0^{e(t)} \partial_t u(t,x)\,dx + e'(t)u(t,e(t)),\\ &= \int_0^{e(t)} \partial_x\left(A(u)\partial_x u\right) + e'(t)u(t,e(t)),\\ &= (A(u)\partial_x u)(t,e(t)) + e'(t)u(t,e(t)) - (A(u)\partial_x u)(t,0),\\ &= \varphi(t).\end{aligned}$$

The calculation for the 0^{th} species reads:

$$\begin{aligned}\frac{d}{dt}\left(\int_{\Omega_t} u_0(t,x)\,dx\right) &= \frac{d}{dt}\left(|\Omega_t| - \sum_{i=1}^n \int_{\Omega_t} u_i(t,x)\,dx\right)\\ &= e'(t) - \sum_{i=1}^n \frac{d}{dt}\left(\int_{\Omega_t} u_i(t,x)\,dx\right)\\ &= \sum_{i=0}^n \phi_i(t) - \sum_{i=1}^n \phi_i(t) = \phi_0(t).\end{aligned}$$

To sum up, the final system of interest reads:

$$\begin{cases} e(t) = e_0 + \int_0^t \sum_{i=0}^n \phi_i(s)\,ds, & \text{for } t\in\mathbb{R}_+^*,\\ \partial_t u - \partial_x\left(A(u)\partial_x u\right) = 0, & \text{for } t\in\mathbb{R}_+^*,\ x\in(0,e(t)),\\ \left(A(u)\partial_x u\right)(t,0) = 0, & \text{for } t\in\mathbb{R}_+^*,\\ \left(A(u)\partial_x u\right)(t,e(t)) + e'(t)u(t,e(t)) = \varphi(t), & \text{for } t\in\mathbb{R}_+^*,\\ u(0,x) = u^0(x), & \text{for } x\in(0,e_0), \end{cases} \tag{2.18}$$

where $u^0 \in L^1(0,e_0)$ is an initial condition satisfying $u^0(x)\in\mathcal{D}$ for almost all $x\in(0,e_0)$. We assume in addition that $w^0 := Dh(u^0)$ belongs to $L^\infty((0,e^0);\mathbb{R}^n)$.

Rescaled version of the model

We introduce here a rescaled version of system (2.18). For all $0\le i\le n$, $t\ge 0$ and $y\in(0,1)$, let us denote by $v_i(t,y) := u_i(t,e(t)y)$. It holds that

$$\partial_t v(t,y) = \partial_t u(t,e(t)y) + e'(t)y\partial_x u(t,e(t)y) \quad\text{and}\quad \partial_y v(t,y) = e(t)\partial_x u(t,e(t)y),$$

where $v := (v_1,\cdots,v_n)$. Thus, u is a solution of (2.18) if and only if v is a solution to the following system:

$$\begin{cases} e(t) = e_0 + \displaystyle\int_0^t \sum_{i=0}^n \phi_i(s)\,ds, & \text{for } t\in\mathbb{R}_+^*,\\ \partial_t v - \frac{1}{e(t)^2}\partial_y\left(A(v)\partial_y v\right) - \frac{e'(t)}{e(t)} y\partial_y v = 0, & \text{for } (t,y)\in\mathbb{R}_+^*\times(0,1),\\ \frac{1}{e(t)}(A(v)\partial_y v)(t,1) + e'(t)v(t,1) = \varphi(t), & \text{for } (t,y)\in\mathbb{R}_+^*\times(0,1),\\ \frac{1}{e(t)}(A(v)\partial_y v)(t,0) = 0, & \text{for } (t,y)\in\mathbb{R}_+^*\times(0,1)\\ v(0,y) = v^0(y), & \text{for } y\in(0,1), \end{cases} \tag{2.19}$$

where $v^0(y) := u^0(e_0 y)$.

Proving the existence of a global weak solution to (2.18) is equivalent to proving the existence of a global weak solution to (2.19).

Actually, it can be seen that the entropy of the system (2.19) satisfies a formal inequality at the continuous level which is at the heart of the proof of our existence result. Indeed, let us denote by

$$\mathcal{E}(t) := \int_0^1 h(v(t,y))\,dy,$$

where v is a solution to (2.19). Then, formal calculations yield that

$$\begin{aligned}
\frac{d\mathcal{E}}{dt}(t) &= \int_0^1 \partial_t v(t,y)\cdot Dh(v(t,y))\,dy \\
&= \frac{1}{e(t)^2}\int_0^1 \partial_y\left(A(v(t,y))\partial_y v(t,y)\right)\cdot Dh(v(t,y))\,dy \\
&+ \frac{e'(t)}{e(t)}\int_0^1 y\partial_y v(t,y)\cdot Dh(v(t,y))\,dy \\
&= -\frac{1}{e(t)^2}\int_0^1 \partial_y v(t,y)\cdot D^2h(v(t,y))A(v(t,y))\partial_y v(t,y)\,dy \\
&+ \frac{1}{e(t)^2}\left(A(v(t,1))\partial_y v(t,1)\right)\cdot Dh(v(t,1)) + \frac{e'(t)}{e(t)}\int_0^1 y\partial_y(h(v(t,y)))\,dy \\
&= -\frac{1}{e(t)^2}\int_0^1 \partial_y v(t,y)\cdot D^2h(v(t,y))A(v(t,y))\partial_y v(t,y)\,dy \\
&+ \frac{1}{e(t)}\left(\varphi(t) - e'(t)v(t,1)\right)\cdot Dh(v(t,1)) \\
&+ \frac{e'(t)}{e(t)}h(v(t,1)) - \frac{e'(t)}{e(t)}\int_0^1 h(v(t,y))\,dy.
\end{aligned}$$

Denoting by $\overline{f}(t) := \frac{\varphi(t)}{e'(t)}$, it holds that $\overline{f}(t)\in\overline{\mathcal{D}}$ for all $t>0$. Besides, using assumption (H2), we obtain that

$$-\int_0^1 \partial_y v(t,y)\cdot D^2h(v(t,y))A(v(t,y))\partial_y v(t,y)\,dy \le 0,$$

which yields that

$$\frac{d\mathcal{E}}{dt}(t) \le \frac{e'(t)}{e(t)}\left[h(v(t,1)+Dh(v(t,1))\cdot\left(\overline{f}(t)-v(t,1)\right) - \int_0^1 h(v(t,y))\,dy\right].$$

Using the convexity of h, we obtain that $h(v(t,1)+Dh(v(t,1))\cdot\left(\overline{f}(t)-v(t,1)\right) \le h(\overline{f}(t))$, so that

$$\frac{d\mathcal{E}}{dt}(t) \le \frac{e'(t)}{e(t)}\left[h(\overline{f}(t)) - \int_0^1 h(v(t,y))\,dy\right]. \tag{2.20}$$

Inequality (2.20) is not an entropy dissipation inequality in the sense that the quantity $\mathcal{E}(t)$ may increase with time. However, using the fact $e'\in L^\infty_{\mathrm{loc}}(\mathbb{R}_+;\mathbb{R}_+)$ and assumption (H3), it implies that the quantity $\mathcal{E}(t)$ cannot blow up in finite time, which is sufficient for our purpose.

2.3.2 Theoretical results

Global in time existence of weak solutions

Our first result deals with the global in time existence of bounded weak solutions to (2.19) (and thus to (2.18)).

Theorem 2.4. *Let $\mathcal{D} := \{(u_1, \cdots, u_n)^T \in (\mathbb{R}_+^*)^n,\ \sum_{i=1}^n u_i < 1\} \subset (0,1)^n$. Let $A : \overline{\mathcal{D}} \to \mathbb{R}^{n\times n}$ be a matrix-valued functional satisfying $A \in \mathcal{C}^0(\overline{\mathcal{D}}; \mathbb{R}^{n\times n})$ and assumptions (H1)-(H2) of Theorem 2.2 for some well-chosen entropy density $h : \overline{\mathcal{D}} \to \mathbb{R}$. We assume in addition that*

(H3) $h \in \mathcal{C}^0(\overline{\mathcal{D}})$.

Let $e_0 > 0$, $u^0 \in L^1((0,e_0);\mathcal{D})$ so that $w^0 := (Dh)^{-1}(u^0) \in L^\infty((0,e_0);\mathbb{R}^n)$ and $(\phi_0, \cdots, \phi_n) \in L^\infty_{\rm loc}(\mathbb{R}_+; \mathbb{R}_+^{n+1})$. Let us define for almost all $y \in (0,1)$, $v^0(y) := u^0(e_0 y)$ and $\varphi := (\phi_1, \cdots, \phi_n)^T$. Then, there exists a weak solution v with initial condition v^0 to (2.19) such that for almost all $(t,y) \in \mathbb{R}_+^* \times (0,1)$, $v(t,y) \in \overline{\mathcal{D}}$. Besides,

$$v \in L^2_{\rm loc}(\mathbb{R}_+; H^1((0,1);\mathbb{R}^n)) \text{ and } \partial_t v \in L^2_{\rm loc}(\mathbb{R}_+; (H^1((0,1);\mathbb{R}^n))').$$

In particular, $v \in \mathcal{C}^0(\mathbb{R}_+; L^2((0,1);\mathbb{R}^n))$.

Let us point out that the example described in Section 2.2.1 satisfies all the assumptions of Theorem 2.4 since the entropy density h defined by (2.10) belongs to $\mathcal{C}^0(\overline{\mathcal{D}})$. Let us also point here that the form of (2.19) is different from the system considered in [Jue15a] through i) the boundary conditions and ii) the existence of the drift term $\frac{e'(t)}{e(t)} y \partial_y v$.

The strategy of proof developped in [BDFPS10, Jue15a] is still adapted to our case though, because a discrete entropy inequality can still be obtained. The proof of Theorem 2.4 is given in full details in Section 2.4.2.

Long-time behaviour for constant fluxes

In the case when the fluxes are constant in time, we obtain long-time asymptotics for the functions v_i, provided that the entropy density h is given by (2.10). More precisely, the following result holds:

Proposition 2.5. *Under the assumptions of Theorem 2.4, let us make the following additional hypotheses:*

(T1) for all $0 \le i \le n$, there exists $\overline{\phi}_i > 0$ so that $\phi_i(t) = \overline{\phi}_i$, for all $t \in \mathbb{R}_+$;

(T2) for all $u \in \overline{D}$, the entropy density h can be chosen so that $h(u) = \sum_{i=1}^n u_i \log u_i + (1-\rho_u)\log(1-\rho_u)$, where $\rho_u = 1 - \sum_{1\le i\le n} u_i$.

For all $0 \le i \le n$, let us define $\overline{f}_i := \frac{\overline{\phi}_i}{\sum_{j=0}^n \overline{\phi}_j}$ and by $\overline{f} := (\overline{f}_i)_{1\le i\le n} \in \mathcal{D}$. Let us also denote by

$$\overline{h} : \left\{ \begin{array}{ccc} \overline{\mathcal{D}} & \mapsto & \mathbb{R} \\ u & \mapsto & h(u) - h(\overline{f}) - Dh(\overline{f})(u - \overline{f}) \end{array} \right.$$

the relative entropy associated to h and $\overline{f}$. Then, there exists a global weak solution v to (2.19) and a constant $C > 0$ such that

$$\int_0^1 \overline{h}\left(v(t,y)\right)\, dy \leq \frac{C}{t+1}, \tag{2.21}$$

and

$$\forall 1 \leq i \leq n, \quad \|v_i(t,\cdot) - \overline{f}_i\|_{L^1(0,1)} \leq \frac{C}{\sqrt{t+1}} \text{ and } \left\|\left(1-\rho_{v(t,\cdot)}\right) - \overline{f}_0\right\|_{L^1(0,1)} \leq \frac{C}{\sqrt{t+1}}. \tag{2.22}$$

The proof of Proposition 2.5 is given in Section 2.4.3. Numerical results presented in Section 5.4 illustrate the rate of convergence of the rescaled concentrations to constant profiles in $\mathcal{O}\left(\frac{1}{t}\right)$.

Let us comment here on assumption (T2). For the sake of simplicity, we chose to restrict ourselves to the case of logarithmic entropy density in Proposition 2.5. Actually, Proposition 2.5 can be easily generalized provided that the relative entropy density $\overline{h}$ satisfies a generalized Csizar-Kullback type inequality [AUT00].

The central ingredient of the proof is the following formal entropy inequality. In the case when h is given by (2.10), it can be easily seen that $\overline{h}$ is also a valid entropy density for the diffusion coefficient A in the sense that $\overline{h}$ also satisfies assumptions (H1)-(H2)-(H3). Thus, inequality (2.20) holds with $\overline{h}$ instead of h so that

$$\frac{d\overline{\mathcal{E}}}{dt}(t) \leq \frac{e'(t)}{e(t)}\left[\overline{h}(\overline{f}) - \int_0^1 \overline{h}(v(t,y))\, dy\right] = \frac{e'(t)}{e(t)}\left[\overline{h}(\overline{f}) - \overline{\mathcal{E}}(t)\right],$$

where for all $t > 0$, $\overline{\mathcal{E}}(t) := \int_0^1 \overline{h}(v(t,y))\, dy$. Denoting by $V := \sum_{i=0}^n \overline{\phi}_i$, it holds that $e'(t) = V$ and $e(t) = e_0 + Vt$ for all $t \geq 0$. Finally, using the fact that $\overline{h} \geq 0$ and that $\overline{h}(\overline{f}) = 0$, we obtain that

$$\left(\frac{e_0}{V} + t\right)\frac{d\overline{\mathcal{E}}}{dt}(t) + \overline{\mathcal{E}}(t) = \frac{d}{dt}\left(\left(\frac{e_0}{V} + t\right)\overline{\mathcal{E}}(t)\right) \leq 0.$$

This inequality implies that there exists a constant $C > 0$ such that for all $t \geq 0$,

$$\overline{\mathcal{E}}(t) \leq \frac{C}{t+1}.$$

The rates on the L^1 norm of the solutions are then obtained using the Csizàr-Kullback inequality.

Let us finally point out that the quantity $\int_0^1 h(v(t,y))\, dy = \frac{1}{e(t)}\int_0^{e(t)} h(u(t,x))\, dx$ can be seen as an average entropy. In particular, the result of Propositon 2.5 does not imply in general the convergence of $u(t,x)$ to a constant vector $L^1_{\text{loc}}(\mathbb{R}_+)$ for instance. Whether such a convergence may hold true remains an open question.

Optimization of the fluxes

As mentioned in the introduction, our main motivation for studying this system is the control of the gazeous fluxes injected during a PVD process. It is assumed here that the wafer remains in the hot chamber where the different atomic species are injected during a time $T > 0$. The cross-diffusion phenomena occur in the bulk of the thin film layer because of the high temperatures that are imposed during the process. Once the wafer is taken out of the chamber, the composition of the film is *freezed* and no diffusion phenomena take place anymore. The profiles of the local volumic fractions of the different chemical species in the film thus remain unchanged after the time T. It is of practical interest to adapt the fluxes through time so that these final concentration profiles are as close as possible to target functions chosen a priori.

Let $e_0 > 0$ be the initial thickness of the solid. In practice, the maximal value of the fluxes which can be injected is limited due to device constraints. Let $F > 0$ and let us then denote by $\Xi := \left\{\Phi \in L^\infty((0,T);\mathbb{R}_+^{n+1}), \quad \|\Phi\|_{L^\infty} \leq F\right\}$. For all $\Phi := (\phi_0, \cdots, \phi_n) \in \Xi$, we denote by $e_\Phi : t \in [0,T] \mapsto e_0 + \int_0^t \sum_{i=0}^n \phi_i(s)\, ds$ the time-dependent thickness of the film, and by v_Φ a solution to (2.19) associated with the external fluxes Φ.

Let us point out here the uniqueness of a solution to (2.18) (or (2.19)) remains an open problem in general. When the diffusion matrix A is defined by (2.7), the only case for which uniqueness of a global solution can be obtained is the trivial case where the cross-diffusion coefficients K_{ij} are identical to some constant $K > 0$ for all $0 \leq i \neq j \leq n$. Indeed, in this case, it can be seen that the system (2.19) can be written as a set of n independent advection-diffusion PDEs on each of the rescaled concentration profiles v_i $(1 \leq i \leq n)$. Thus, we will have to make some assumption on the cross-diffusion coefficients $(K_{ij})_{0\leq i\neq j\leq n}$ in the general case.

We make the following assumption on the diffusion matrix A:

(C1) For any $\Phi \in \Xi$, there exists a unique global weak solution v_Φ to system (2.19) so that for almost all $(t,y) \in \mathbb{R}_+^* \times (0,1)$, $v_\Phi(t,y) \in \overline{\mathcal{D}}$.

The goal of the optimization problem consists in the identification of optimal time-dependent non-negative functions $\Phi \in \Xi$ so that the final thickness of the film $e_\Phi(T)$ and the (rescaled) concentration profiles for the different chemical species $v_\Phi(T,\cdot)$ at the end of the fabrication process are as close as possible to desired targets denoted by $e_{\rm opt} > e_0$ and $v_{\rm opt} \in L^2((0,1);\overline{\mathcal{D}})$.

The real-valued functional $\mathcal{J} : \Xi \to \mathbb{R}$ defined by

$$\forall \Phi \in \Xi, \quad \mathcal{J}(\Phi) := |e_\Phi(T) - e_{\rm opt}|^2 + \|v_\Phi(T,\cdot) - v_{\rm opt}\|^2_{L^2(0,1)}, \tag{2.23}$$

is the cost function we consider here. More precisely, we have the following result, which is proved in Section 2.4.4.

Proposition 2.6. *Under the assumptions of Theorem 2.4, let us make the additional assumption (C1). Then, the functional $\mathcal{J}$ is well-defined and there exists a minimizer $\Phi^* \in \Xi$ to the minimization problem*

$$\Phi^* \in \underset{\Phi\in\Xi}{\rm argmin}\, \mathcal{J}(\Phi). \tag{2.24}$$

Of course, uniqueness of such a solution Φ^* is not expected in general.

2.4 Proofs

2.4.1 Proof of Lemma 2.3

Let us prove that the matrix-valued function A defined in (2.7) satisfies the assumptions of Theorem 2.2 with the entropy functional h given by (2.10).

As mentioned in Section 2.2.1, the entropy density h belongs to $\mathcal{C}^0(\overline{\mathcal{D}};\mathbb{R})\cap\mathcal{C}^2(\mathcal{D};\mathbb{R})$ (thus is bounded on $\overline{\mathcal{D}}$), is strictly convex on $\mathcal{D}$, and its derivative $Dh:\mathcal{D}\to\mathbb{R}^n$ is invertible. As a consequence, h satisfies assumption (H1) of Theorem 2.2.

Let us now prove that assumption (H2) of Theorem 2.2 is satisfied with $m_i=\frac{1}{2}$ for all $1\le i\le n$. To this aim, we follow the same strategy of proof as the one used in [JZ14]. Let us prove that there exists $\beta>0$ such that for all $u\in\mathcal{D}$,

$$H(u)A(u)\ge\beta\Lambda(u), \tag{2.25}$$

$$\text{where } H(u):=D^2h(u),\ \Lambda(u):=\operatorname{diag}\left(\left(\frac{1}{u_i}\right)_{1\le i\le n}\right) \text{ and } \beta:=\min_{0\le i\ne j\le n}K_{ij}.$$

This inequality implies (H2) with $\alpha=\beta$ and $m_i=\frac{1}{2}$ for all $1\le i\le n$.

Let $u\in D$. We have for all $1\le i,j\le n$,

$$H_{ii}(u)=\frac{1}{u_i}+\frac{1}{1-\rho_u}\text{ and }H_{ij}(u)=\frac{1}{1-\rho_u}\text{ if }i\ne j.$$

Introducing $P(u):=(P_{ij}(u))_{1\le i,j\le n}$, where for all $1\le i,j\le n$,

$$P_{ii}(u)=1-u_i\text{ and }P_{ij}(u)=-u_i\text{ if }i\ne j,$$

it holds that $H(u)P(u)=\Lambda(u)$. Thus, $H(u)A(u)-\beta\Lambda(u)=H(u)(A(u)-\beta P(u))$. It can be easily checked that $A(u)-\beta P(u)=\widetilde{A}(u)+\beta D(u)$, where $\widetilde{A}(u)$ has the same structure as $A(u)$ but with diffusion coefficients $K_{ij}-\beta$ instead of K_{ij}, and $D(u):=(D_{ij}(u))_{1\le i,j\le n}$ where $D_{ij}(u)=u_i$ for all $1\le i\le n$.

On the one hand, $H(u)D(u)=\frac{1}{1-\rho_u}Z$ where Z is the $n\times n$ matrix whose all coefficients are identically equal to 1. Since the matrix Z is a semi-definite positive matrix, so is $H(u)D(u)$.

On the other hand, since h is strictly convex on $\mathcal{D}$, $H(u)\widetilde{A}(u)$ is semi-definite positive if and only if $\widetilde{M}(u):=\widetilde{A}(u)H(u)^{-1}$ is semi-definite positive. Indeed, for all $z\in\mathbb{R}^n$, we have $z^TH(u)\widetilde{A}(u)z=(H(u)z)^T\left(\widetilde{A}(u)H(u)^{-1}\right)(H(u)z)$. It can be observed that $\widetilde{M}(u)=(\widetilde{M}_{ij}(u))_{1\le i,j\le n}$, where for all $1\le i,j\le n$,

$$\widetilde{M}_{ii}(u)=(K_{i0}-\beta)(1-\rho_u)u_i+\sum_{1\le j\ne i\le n}(K_{ij}-\beta)u_iu_j\text{ and }\widetilde{M}_{ij}(u)=-(K_{ij}-\beta)u_iu_j\text{ if }j\ne i.$$

For all $z=(z_1,\cdots,z_n)^T\in\mathbb{R}^n$, we have

$$\begin{aligned}z^T\widetilde{M}(u)z&=\sum_{i=1}^n(K_{i0}-\beta)(1-\rho_u)u_iz_i^2+\sum_{i=1}^n\sum_{1\le j\ne i\le n}(K_{ij}-\beta)u_iu_j(z_i^2-z_iz_j),\\&=\sum_{i=1}^n(K_{i0}-\beta)(1-\rho_u)u_iz_i^2+\sum_{1\le i\ne j\le n}(K_{ij}-\beta)u_iu_j\left(\frac{1}{2}z_i^2+\frac{1}{2}z_j^2-z_iz_j\right),\\&\ge 0.\end{aligned}$$

The matrix $\widetilde{M}(u)$ is indeed a semi-definite positive matrix. Hence we have proved inequality (2.25), which yields the desired result.

2.4.2 Proof of Theorem 2.4

For the sake of simplicity, we will prove the existence of a solution v on the finite time interval $[0,T]$ where $T>0$ is an arbitrary positive constant. Actually, the proof can be easily adapted to obtain the existence of a global solution for an infinite time horizon.

The proof follows similar lines as the proof of Theorem 2 of [Jue15a] and is divided in three main steps. Firstly, an approximate time-discrete problem is introduced for which uniform bounds are proved in a second step. Lastly, passing to the limit in this approximate problem using the obtained bounds enables to obtain the existence of a weak solution.

Step 1 : Approximate time-discrete problem

Let us first assume at this point that $\phi_0,\cdots,\phi_n$ belong to $\mathcal{C}^0([0,T])$.

Let $N\in\mathbb{N}$, $\tau=\frac{T}{N}$ and $\epsilon>0$. For all $k\in\mathbb{N}^*$ so that $k\tau\leq T$, let us denote by $e_k:=e(k\tau)$, $e'_k:=e'(k\tau)$ and $\varphi_k=(\phi_{1,k},\cdots,\phi_{n,k})^T:=\varphi(k\tau)$. Let us also define

$$f_k:=\begin{cases}\frac{\varphi_k}{e'_k} & \text{if } e'_k>0,\\ 0 & \text{otherwise,}\end{cases} \tag{2.26}$$

so that $f_k\in\overline{D}$ and $\varphi_k=e'_kf_k$.

By assumption, $w^0(y):=Dh(v^0(y))$ belongs to $L^\infty((0,1);\mathbb{R}^n)$. In the rest of the proof, for any $w\in\mathbb{R}^n$, we denote by $v(w):=(Dh)^{-1}(w)=(v_i(w))_{1\leq i\leq n}$ and by $B(w):=M(v(w))$.

Let us already mention at this point that the (formal) weak formulation of (2.19) reads as follows: for all $\psi\in L^2((0,T);H^1((0,1);\mathbb{R}^n))$,

$$\int_0^T\int_0^1\partial_tv\cdot\psi+\int_0^T\int_0^1\partial_y\frac{1}{e^2}\psi\cdot(A(v)\partial_yv)+\int_0^T\int_0^1\frac{e'}{e}(v\cdot\psi+yv\cdot\partial_y\psi)=\int_0^T\frac{1}{e}\varphi\cdot\psi(\cdot,1).$$

Let us first prove the following lemma.

Lemma 2.7. *Assume that $\phi_0,\cdots,\phi_n\in\mathcal{C}^0([0,T])$. Then, for all $k\in\mathbb{N}^*$ such that $k\tau\leq T$, there exists $w^k\in H^1((0,1);\mathbb{R}^n)$ solution of*

$$\frac{1}{\tau}\int_0^1\Big(v(w^k)-v(w^{k-1})\Big)\cdot\psi+\frac{1}{e_k^2}\int_0^1\partial_y\psi\cdot(B(w^k)\partial_yw^k)+\epsilon\int_0^1(\partial_yw^k\cdot\partial_y\psi+w^k\cdot\psi) \tag{2.27}$$

$$+\frac{e'_k}{e_k}\int_0^1(v(w^k)\cdot\psi+yv(w^k)\cdot\partial_y\psi)=\frac{1}{e_k}\varphi_k\cdot\psi(1),$$

for all $\psi \in H^1((0,1);\mathbb{R}^n)$. *Besides, the following discrete inequality holds for all* $k \in \mathbb{N}^*$ *such that* $k\tau \leq T$,

$$\begin{aligned}&\frac{1}{\tau}\int_0^1 h(v(w^k)) + \epsilon\int_0^1 \left(|\partial_y w^k|^2 + |w^k|^2\right) + \frac{1}{e_k^2}\int_0^1 \partial_y w^k \cdot (B(w^k)\partial_y w^k) \\ &\leq \frac{1}{\tau}\int_0^1 h(v(w^{k-1})) + \frac{e_k'}{e_k}\left(h(f_k) - \int_0^1 h(v(w^k))\right).\end{aligned} \tag{2.28}$$

The proof of this lemma is postponed until Section 2.4.2. Let us point out the following fact: from (2.28), we obtain

$$\begin{aligned}&\left(\frac{1}{\tau} + \frac{e_k'}{e_k}\right)\int_0^1 h(v(w^k)) + \epsilon\int_0^1 (|\partial_y w^k|^2 + |w^k|^2) + \frac{1}{e_k^2}\int_0^1 \partial_y w^k \cdot B(w^k)\partial_y w^k \\ &\leq \frac{1}{\tau}\int_0^1 h(v(w^{k-1})) + \frac{e_k'}{e_k}\|h\|_{L^\infty(\overline{\mathcal{D}})},\end{aligned} \tag{2.29}$$

which implies

$$\begin{aligned}&\frac{1}{\tau}\int_0^1 h(v(w^k)) + \epsilon\int_0^1 (|\partial_y w^k|^2 + |w^k|^2) + \frac{1}{e_k^2}\int_0^1 \partial_y w^k \cdot B(w^k)\partial_y w^k \\ &\leq \frac{1}{\tau}\int_0^1 h(v(w^{k-1})) + 2\frac{e_k'}{e_k}\|h\|_{L^\infty(\overline{\mathcal{D}})}.\end{aligned} \tag{2.30}$$

Step 2: Uniform bounds

For all $0 \leq i \leq n$, let $(\phi_{i,p})_{p\in\mathbb{N}}$ be a sequence of non-negative functions of $\mathcal{C}^0([0,T])$ which weakly-* converges to ϕ_i in $L^\infty(0,T)$ as p goes to infinity, and for all $p \in \mathbb{N}$,

$$\|\phi_{i,p}\|_{L^\infty(0,T)} \leq \|\phi_i\|_{L^\infty(0,T)}.$$

Let us define

$$\varphi_p := (\phi_{1,p}, \cdots, \phi_{n,p})^T, \quad \text{and } e_p(t) := e_0 + \int_0^t \sum_{i=0}^n \phi_{i,p}(s)\, ds.$$

It holds that $(e_p)_{p\in\mathbb{N}^*}$ strongly converges to e in $L^\infty(0,T)$. Indeed, let $\varepsilon > 0$. Since e is continuous on $[0,T]$, it is uniformly continuous, and there exists $\eta > 0$ so that for all $t, t' \in [0,T]$ satisfying $|t - t'| \leq \eta$, then $|e(t) - e(t')| \leq \varepsilon/2$. Let $M \in \mathbb{N}^*$ and $0 = s_0 < s_1 < \cdots < s_M = T$ so that for all $0 \leq j \leq M-1$, $|s_j - s_{j+1}| \leq \eta$. Then, it holds that

$$\max_{0\leq j\leq M} |e_p(s_j) - e(s_j)| \underset{p\to+\infty}{\longrightarrow} 0,$$

because of the weak-* convergence in $L^\infty[0,T]$ of $(\phi_{i,p})_{p\in\mathbb{N}^*}$ to ϕ_i for all $0 \leq i \leq n$.

Thus, there exists $p_0 \in \mathbb{N}^*$ large enough such that for all $p \geq p_0$, $\max_{0\leq j\leq M} |e_p(s_j) - e(s_j)| \leq \varepsilon/2$. Besides, the non-negativity of the functions ϕ_i and $\phi_{i,p}$ implies that e and e_p are non-decreasing functions, so that for all $0 \leq j \leq M-1$ and all $p \in \mathbb{N}^*$,

$$\forall s \in [s_j, s_{j+1}], \qquad e(s_j) \leq e(s) \leq e(s_{j+1}) \quad \text{and} \quad e_p(s_j) \leq e_p(s) \leq e_p(s_{j+1}).$$

As a consequence, for all $p \geq p_0$, all $0 \leq j \leq M-1$ and all $s \in [s_j, s_{j+1}]$,

$$\begin{aligned} |e(s) - e_p(s)| &\leq \max\left(|e(s_{j+1}) - e_p(s_j)|, |e_p(s_{j+1}) - e(s_j)|\right) \\ &\leq \max\left(|e(s_{j+1}) - e(s_j)| + |e(s_j) - e_p(s_j)|, |e_p(s_{j+1}) - e(s_{j+1})| + |e(s_{j+1}) - e(s_j)|\right) \\ &\leq \varepsilon. \end{aligned}$$

Hence, for all $p \geq p_0$, $\|e - e_p\|_{L^\infty(0,T)} \leq \varepsilon$, which yields the strong convergence of the sequence $(e_p)_{p\in\mathbb{N}^*}$ to e in $L^\infty(0,T)$.

For all $k \in \mathbb{N}^*$ such that $k\tau \leq T$, we denote by $w^{k,p}$ a solution to (2.27) associated to the fluxes $(\phi_{i,p})_{0\leq i\leq n}$. The time-discretized associated quantities are denoted (using obvious notation) by $\varphi_{k,p}$, $e_{k,p}$ and $e'_{k,p}$.

Let us define the piecewise constant in time functions $w^{(\epsilon,\tau,p)}(y,t)$, $v^{(\epsilon,\tau,p)}(y,t)$, $\sigma_\tau v^{(\epsilon,\tau,p)}(y,t)$, $e_{(\tau,p)}(t)$ and $e^d_{(\tau,p)}(t)$ as follows: for all $k \geq 1$ such that $k\tau \leq T$, $(k-1)\tau < t \leq k\tau$ and almost all $y \in (0,1)$,

$$w^{(\epsilon,\tau,p)}(y,t) := w^{k,p}(y), \quad v^{(\epsilon,\tau,p)}(y,t) := Dh(w^{k,p}(y)), \quad \sigma_\tau v^{(\epsilon,\tau,p)}(y,t) = Dh(w^{k-1,p}(y)),$$
$$e_{(\tau,p)}(t) = e_{k,p}, \quad e^d_{(\tau,p)}(t) := e'_{k,p}, \quad \varphi_{(\tau,p)} := \varphi_{k,p}.$$

Besides, let us set $w^{(\epsilon,\tau,p)}(0,\cdot) = Dh(v^0)$ and $v^{(\epsilon,\tau,p)}(0,\cdot) = v^0$. Let us also denote by $(v_1^{(\epsilon,\tau,p)}, \cdots, v_n^{(\epsilon,\tau,p)})$ the n components of $v^{(\epsilon,\tau,p)}$.

Then, the following system holds for all piecewise constant in time functions $\psi : (0,T) \to H^1((0,1);\mathbb{R}^n)$,

$$\begin{aligned} &\frac{1}{\tau}\int_0^T\int_0^1 \left(v^{(\epsilon,\tau,p)} - \sigma_\tau v^{(\epsilon,\tau,p)}\right)\cdot\psi\, dy\, dt + \int_0^T \frac{1}{e^2_{(\tau,p)}}\int_0^1 \partial_y\psi\cdot(B(w^{(\epsilon,\tau,p)})\partial_y w^{(\epsilon,\tau,p)})\, dy\, dt \\ &+ \epsilon\int_0^T\int_0^1 (\partial_y w^{(\epsilon,\tau,p)}\cdot\partial_y\psi + w^{(\epsilon,\tau,p)}\cdot\psi)\, dy\, dt + \int_0^T \frac{e^d_{(\tau,p)}}{e_{(\tau,p)}}\int_0^1 v(w^{(\epsilon,\tau,p)})\cdot\psi \\ &+ yv(w^{(\epsilon,\tau,p)})\cdot\partial_y\psi)\, dy\, dt \\ &= \int_0^T \frac{1}{e_{(\tau,p)}}\varphi_{(\tau,p)}\cdot\psi(1)\, dt. \end{aligned} \tag{2.31}$$

The set of piecewise constant functions in time $\psi : (0,T) \to H^1((0,1);\mathbb{R}^n)$ is dense in $L^2((0,T);H^1((0,1);\mathbb{R}^n))$, so that (2.31) also holds for any $\psi \in L^2((0,T);H^1((0,1);\mathbb{R}^n))$.

Using the fact that A satisfies assumption (H2) of Theorem 2.2 and the fact that $\partial_y w^{k,p} = D^2 h(v^{k,p}) \partial_y v^{k,p}$, we obtain for all $k \in \mathbb{N}^*$ such that $k\tau \leq T$,

$$\begin{aligned}
\int_0^1 \partial_y w^{k,p} \cdot (B(w^{k,p}) \partial_y w^{k,p}) &= \int_0^1 \partial_y v(w^{k,p}) \cdot \left[D^2 h(v(w^{k,p})) A(v(w^{k,p})) \partial_y v(w^{k,p}) \right] dy \\
&\geq \sum_{i=1}^n \int_0^1 \alpha \left| v_i(w^{k,p}) \right|^{2m_i - 2} |\partial_y v_i(w^{k,p})|^2 \, dy \\
&= \sum_{i=1}^n \int_0^1 |\partial_y G_i(v_i(w^{k,p}))|^2 \, dy \\
&= \int_0^1 |\partial_y G(v(w^{k,p}))|^2 \, dy,
\end{aligned}$$

where $G_i(s) := \frac{\sqrt{\alpha}}{m_i} |s|^{m_i}$ for all $s \in (0,1)$ and $G(z) = (G_i(z_i))_{1 \leq i \leq n}$ for all $z := (z_i)_{1 \leq i \leq n} \in (0,1)^n$. It follows from (2.30) that for all $k \in \mathbb{N}^*$ such that $k\tau \leq T$,

$$\begin{aligned}
&\int_0^1 h(v(w^{k,p})) + \tau \int_0^1 |\partial_y \widetilde{\alpha}(v(w^{k,p}))|^2 \\
&+ \epsilon\tau \int_0^1 \left(|\partial_y w^{k,p}|^2 + |w^{k,p}|^2 \right) \leq 2\tau \|h\|_{L^\infty(\overline{\mathcal{D}})} \frac{e'_{k,p}}{e_{k,p}} + \int_0^1 h(v(w^{k-1,p})).
\end{aligned}$$

Summing these inequalities yields, for $k \in \mathbb{N}^*$ so that $k\tau \leq T$,

$$\begin{aligned}
&\int_0^1 h(v(w^{k,p})) + \tau \sum_{j=1}^k \int_0^1 |\partial_y G(v(w^{j,p}))|^2 + \epsilon\tau \sum_{j=1}^k \int_0^1 (|\partial_y w^{j,p}|^2 + |w^{j,p}|^2) \qquad (2.32) \\
&\leq 2\tau \|h\|_{L^\infty(\overline{\mathcal{D}})} \sum_{j=1}^k \frac{e'_{j,p}}{e_{j,p}} + \int_0^1 h(v^0), \\
&\leq 2\|h\|_{L^\infty(\overline{\mathcal{D}})} \frac{1}{e_0} \sum_{j=1}^k \tau e'_{j,p} + \int_0^1 h(v^0), \\
&\leq 2\|h\|_{L^\infty(\overline{\mathcal{D}})} \frac{(n+1)\|\Phi\|_{L^\infty(0,T)}}{e_0} T + \int_0^1 h(v^0).
\end{aligned}$$

In the sequel, C will denote an arbitrary constant, which may change along the calculations, but remains independent on ϵ, τ, p and Φ. We are deliberately keeping here the explicit dependence of the constants on $\|\Phi\|_{L^\infty(0,T)}$ in view of the proof of Proposition 2.6. It then holds that

$$\|e^d_{(\tau,p)}\|_{L^\infty(0,T)} \leq C\|\Phi\|_{L^\infty(0,T)} \text{ and } 0 < e_0 \leq \|e_{(\tau,p)}\|_{L^\infty(0,T)} \leq C\|\Phi\|_{L^\infty(0,T)}.$$

We also obtain from (2.32) and the fact that $\|G_i\|_{L^\infty(0,1)} \leq \frac{\sqrt{\alpha}}{m_i}$ for all $1 \leq i \leq n$ that

$$\|G(v^{(\epsilon,\tau,p)})\|_{L^2((0,T);H^1(0,1)^n)} \leq C\left(1 + \|\Phi\|_{L^\infty(0,T)}\right) \qquad (2.33)$$

and

$$\sqrt{\epsilon}\|w^{(\epsilon,\tau,p)}\|_{L^2((0,T);H^1(0,1)^n)} \leq C\left(1+\|\Phi\|_{L^\infty(0,T)}\right). \tag{2.34}$$

Since for all $1 \leq i \leq n$, $m_i \leq 1$, this implies that

$$\begin{aligned}\|\partial_y v_i^{(\epsilon,\tau,p)}\|_{L^2((0,T);L^2(0,1))} &= \left\| \frac{\left|v_i^{(\epsilon,\tau,p)}\right|^{1-m_i}}{m_i} \partial_y \left(|v_i^{(\epsilon,\tau,p)}|^{m_i}\right) \right\|_{L^2((0,T);L^2(0,1))} \\ &= \left\| \frac{\left|v_i^{(\epsilon,\tau,p)}\right|^{1-m_i}}{\sqrt{\alpha}} \partial_y G_i(v_i^{(\epsilon,\tau,p)}) \right\|_{L^2((0,T);L^2(0,1))} \\ &\leq C\|\partial_y G_i(v_i^{(\epsilon,\tau,p)})\|_{L^2((0,T);L^2(0,1))} \leq C\left(1+\|\Phi\|_{L^\infty(0,T)}\right).\end{aligned} \tag{2.35}$$

Besides,

$$\begin{aligned}\left\|A(v^{(\epsilon,\tau,p)}\partial_y v^{(\epsilon,\tau,p)}\right\|^2_{L^2((0,T);L^2(0,1)^n)} &\leq \left\|A(v^{(\epsilon,\tau,p)})\right\|^2_{L^\infty((0,T);L^\infty(0,1)^{n\times n})} \left\|\partial_y v^{(\epsilon,\tau,p)}\right\|^2_{L^2((0,T);L^2(0,1)^n)} \\ &\leq C\left(1+\|\Phi\|_{L^\infty(0,T)}\right),\end{aligned} \tag{2.36}$$

using the fact that $A \in \mathcal{C}^0(\overline{\mathcal{D}};\mathbb{R}^{n\times n})$.

This yields that for all $\psi \in L^2((0,T);H^1((0,1);\mathbb{R}^n))$,

$$\begin{aligned}\frac{1}{\tau}\left|\int_\tau^T\int_0^1 (v^{(\epsilon,\tau,p)}-\sigma_\tau v^{(\epsilon,\tau,p)})\cdot\psi\,dy\,dt\right| &\leq \frac{1}{e_0^2}\|A(v^{(\epsilon,\tau,p)}\partial_y v^{(\epsilon,\tau,p)}\|_{L^2((0,T);L^2(0,1)^n)}\|\partial_y\psi\|_{L^2((0,T);L^2(0,1)^n)} \\ &\quad + \epsilon\|w^{(\epsilon,\tau,p)}\|_{L^2((0,T);H^1(0,1)^n)}\|\psi\|_{L^2((0,T);H^1(0,1)^n)} \\ &\quad + 2\frac{\|e^d_{(\tau,p)}\|_{L^\infty(0,T)}}{e_0}\|v^{(\epsilon,\tau,p)}\|_{L^2((0,T);H^1(0,1)^n)}\|\psi\|_{L^2((0,T);H^1(0,1)^n)} \\ &\quad + \frac{1}{e_0}\|\Phi\|_{L^\infty(0,T)}\|\psi\|_{L^2((0,T);H^1(0,1)^n)}, \\ &\leq C\left(1+\|\Phi\|_{L^\infty(0,T)}\right)\|\psi\|_{L^2((0,T);H^1(0,1)^n)}.\end{aligned}$$

This last inequality shows that

$$\frac{1}{\tau}\|v^{(\epsilon,\tau,p)}-\sigma_\tau v^{(\epsilon,\tau,p)}\|_{L^2((\tau,T);(H^1(0,1)^n)')} \leq C\left(1+\|\Phi\|_{L^\infty(0,T)}\right). \tag{2.37}$$

Step 3: The limit $p\to+\infty$ and $\epsilon,\tau\to 0$

For all $p\in\mathbb{N}^*$, the functions e_p' and e_p are continuous on $[0,T]$, and hence are uniformly continuous. As a consequence, there exists $\tau_p>0$ small enough so that for any $t,t'\in[0,T]$ satisfying $|t-t'|\leq\tau_p$, then $|e_p'(t)-e_p'(t')|\leq\frac{1}{p}$ and $|e_p(t)-e_p(t')|\leq\frac{1}{p}$. This implies in particular that

$$\|e^d_{(\tau_p,p)}-e_p'\|_{L^\infty(0,T)}\leq\frac{1}{p} \quad \text{and } \|e_{(\tau_p,p)}-e_p\|_{L^\infty(0,T)}\leq\frac{1}{p}.$$

These inequalities, together with the fact that $(e'_p)_{p\in\mathbb{N}^*}$ weakly-* converges to e' in $L^\infty(0,T)$ (respectively that $(e_p)_{p\in\mathbb{N}^*}$ strongly converges to e in $L^\infty(0,T)$), imply that the sequence $\left(e^d_{(\tau_p,p)}\right)_{p\in\mathbb{N}^*}$ (respectively $\left(e_{(\tau_p,p)}\right)_{p\in\mathbb{N}^*}$) also weakly-* converges to e' in $L^\infty(0,T)$ (respectively strongly converges to e in $L^\infty(0,T)$).

In the following, we consider such a subsequence $(\tau_p)_{p\in\mathbb{N}^*}$. The uniform estimates (2.37) and (2.35) allow us to apply the Aubin lemma in the version of Theorem 1 of [DJ12]. Up to extracting a subsequence which is not relabeled, there exists $v = (v_i)_{1\le i\le n} \in H^1((0,T);(H^1((0,1);\mathbb{R}^n))') \cap L^2((0,T);H^1((0,1);\mathbb{R}^n))$ so that as p goes to infinity and ϵ goes to 0,

$$v^{(\epsilon,\tau_p,p)} \underset{p\to+\infty,\epsilon\to0}{\longrightarrow} v, \quad \begin{cases} \text{strongly in } L^2((0,T);L^2((0,1);\mathbb{R}^n)), \\ \text{weakly in } L^2((0,T);H^1((0,1);\mathbb{R}^n)), \\ \text{and a.e. in } (0,T)\times(0,1), \end{cases}$$

$$\frac{1}{\tau_p}\left(v^{(\epsilon,\tau_p,p)} - \sigma_{\tau_p} v^{(\epsilon,\tau_p,p)}\right) \underset{p\to+\infty,\epsilon\to0}{\rightharpoonup} \partial_t v \text{ weakly in } L^2((0,T);(H^1((0,1);\mathbb{R}^n))').$$

Because of the boundedness of $v^{(\epsilon,\tau_p,p)}$ in $L^\infty((0,T);L^\infty((0,1);\mathbb{R}^n))$, the convergence even holds strongly in $L^q((0,T);L^q((0,1);\mathbb{R}^n))$ for any $q < +\infty$, which is a consequence of the dominated convergence theorem. The latter theorem, together with $A \in \mathcal{C}^0(\overline{\mathcal{D}};\mathbb{R}^{n\times n})$ implies also that the convergence $A(v^{(\epsilon,\tau_p,p)}) \longrightarrow A(v)$ holds strongly in $L^q((0,T);L^q((0,1);\mathbb{R}^{n\times n}))$. Moreover, using (2.36) and (2.34), up to extracting another subsequence, there exists $V \in L^2((0,T);L^2((0,1);\mathbb{R}^n))$ so that

$$\begin{aligned} A(v^{(\epsilon,\tau_p,p)})\partial_y v^{(\epsilon,\tau_p,p)} &\rightharpoonup V \text{ weakly in } L^2((0,T);L^2((0,1);\mathbb{R}^n)), \\ \epsilon w^{(\epsilon,\tau_p,p)} &\longrightarrow 0 \text{ strongly in } L^2((0,T);H^1((0,1);\mathbb{R}^n)). \end{aligned}$$

The strong convergence of $A(v^{(\epsilon,\tau_p,p)})$ in $L^q((0,T);L^q((0,1);\mathbb{R}^n))$ and the weak convergence of $\partial_y v^{(\epsilon,\tau_p,p)}$ in $L^2((0,T);L^2((0,1);\mathbb{R}^n))$ implies necessarily that $V = A(v)\partial_y v$.

We are now in position to pass to the limit $\epsilon \to 0$ and $p \to +\infty$ in (2.31) with $\tau = \tau_p$ and $\psi \in L^2((0,T);H^1((0,1);\mathbb{R}^n))$. Let us recall that $\left(e_{(\tau_p,p)}\right)_{p\in\mathbb{N}^*}$ (respectively $\left(e^d_{(\tau_p,p)}\right)_{p\in\mathbb{N}^*}$) converges strongly (respectively weakly-*) to e (respectively e') in $L^\infty(0,T)$. We obtain that v is a solution to

$$\begin{aligned} &\int_0^T\int_0^1 \partial_t v\cdot\psi\,dy\,dt + \int_0^T \frac{1}{e(t)^2}\int_0^1 \partial_y\psi\cdot(A(v)\partial_y v)\,dy\,dt \\ &+ \int_0^T \frac{e'(t)}{e(t)}\int_0^1 (v\cdot\psi + yv\cdot\partial_y\psi)\,dy\,dt = \int_0^T \frac{1}{e(t)}\varphi\cdot\psi(1)\,dt, \end{aligned} \tag{2.38}$$

yielding the result.

Proof of Lemma 2.7

Proof of Lemma 2.7. We prove Lemma 2.7 by induction using the Leray-Schauder fixed-point theorem. Let $z \in L^\infty((0,1);\mathbb{R}^n)$ and $\delta \in [0,1]$. We consider the following linear

problem: find $w \in H^1((0,1);\mathbb{R}^n)$ solution of

$$\forall \psi \in H^1((0,1);\mathbb{R}^n), \quad a_z(w,\psi) = l_{\delta,z}(\psi), \tag{2.39}$$

where

$$a_z(w,\psi) := \frac{1}{e_k^2}\int_0^1 \partial_y\psi \cdot B(z)\partial_y w + \epsilon \int_0^1 (\partial_y w \cdot \partial_y \psi + w \cdot \psi)$$

and

$$l_{\delta,z}(\psi) := -\frac{\delta}{\tau}\int_0^1 (v(z) - v(w^{k-1})) \cdot \psi + \frac{\delta}{e_k}\varphi_k \cdot \psi(1) - \delta\frac{e_k'}{e_k}\int_0^1 (v(z)\cdot\psi + yv(z)\cdot\partial_y\psi).$$

As a consequence of (H2), the matrix $B(z)$ is positive semi-definite for any $z \in \mathbb{R}^n$. Thus, the bilinear form a_z is coercive and continuous on $H^1((0,1);\mathbb{R}^n)$, and it holds that

$$\forall \psi \in H^1((0,1);\mathbb{R}^n),\ a_z(\psi,\psi) \geq \epsilon \|\psi\|^2_{H^1(0,1)}. \tag{2.40}$$

Since $v(z) \in L^\infty((0,1);\mathbb{R}^n)$ and $\|v(z)\|_{L^\infty(0,1)} \leq 1$, the linear form $l_{\delta,z}$ is continuous. From the Agmon inequality, there exists $C > 0$ independent of $\Phi := (\phi_0, \cdots, \phi_n)$, ϵ or τ such that for all $\psi \in H^1((0,1);\mathbb{R}^n)$,

$$|l_{\delta,z}(\psi)| \leq \left(\frac{2}{\tau} + C\,\|\Phi\|_{L^\infty(0,T)}\right)\|\psi\|_{H^1(0,1)}, \tag{2.41}$$

where $\|\Phi\|_{L^\infty(0,T)} = \max\limits_{i=0,\cdots,n} \|\phi_i\|_{L^\infty(0,T)}$. It immediately follows from the Lax-Milgram theorem that there exists a unique solution $w \in H^1((0,1);\mathbb{R}^n)$ to (2.39).

We define the operator $S : [0,1]\times L^\infty((0,1);\mathbb{R}^n) \to L^\infty((0,1);\mathbb{R}^n)$ as follows. For all $\delta \in [0,1]$ and $\chi \in L^\infty((0,1);\mathbb{R}^n)$, $S(\delta,\chi)$ is the unique solution $w \in H^1((0,1);\mathbb{R}^n) \hookrightarrow L^\infty((0,1);\mathbb{R}^n)$ of (2.39). We are going to prove that there exists a fixed-point $w^k \in H^1((0,1);\mathbb{R}^n)$ of the equation $S(1,w^k) = w^k$ using the Leray-Schauder fixed-point theorem (Theorem 2.8 in the Appendix). This will end the proof of Lemma 2.7 since such a fixed-point w^k is a solution of (2.27).

Let us check that all the assumptions of Theorem 2.8 are satisfied:

(A1) For all $\chi \in L^\infty((0,1);\mathbb{R}^n)$, $S(0,\chi) = 0$;

(A2) Let us prove that S is a compact map. To this aim, let us first prove that it is continuous. Let $(\delta_n)_{n\in\mathbb{N}}$ and $(\chi_n)_{n\in\mathbb{N}}$ be sequences in $[0,1]$ and $L^\infty((0,1);\mathbb{R}^n)$ respectively, $\delta \in [0,1]$ and $\chi \in L^\infty((0,1);\mathbb{R}^n)$ so that $\delta_n \underset{n\to+\infty}{\longrightarrow} \delta$ and $\chi_n \underset{n\to+\infty}{\longrightarrow} \chi$ strongly in $L^\infty((0,1);\mathbb{R}^n)$. For all $n \in \mathbb{N}$, let $w_n := S(\delta_n,\chi_n)$. From assumption (H1) and the global inversion theorem, $h : \mathcal{D} \to \mathbb{R}^n$ is a $\mathcal{C}^2$-diffeomorphism. Thus, together with the fact that $A \in \mathcal{C}^0(\overline{\mathcal{D}};\mathbb{R}^{n\times n})$, it holds that the applications $z \in \mathbb{R}^n \mapsto v(z) = (Dh)^{-1}(z)$ and $z \in \mathbb{R}^n \mapsto B(z) = A(v(z))D^2h((Dh)^{-1}(z)) = A(v(z)D\left(Dh^{-1}\right)(z)$ are continuous. Hence, $v(\chi_n) \underset{n\to+\infty}{\longrightarrow} v(\chi)$ and $B(\chi_n) \underset{n\to+\infty}{\longrightarrow} B(\chi)$ strongly in $L^\infty((0,1);\mathbb{R}^n)$ and $L^\infty((0,1);\mathbb{R}^{n\times n})$ respectively.

Besides, the uniform coercivity and continuity estimates (2.40) and (2.41) imply that $(w_n)_{n\in\mathbb{N}}$ is a bounded sequence in $H^1((0,1);\mathbb{R}^n)$. Thus, up to the extraction of a subsequence which is not relabeled, $(w_n)_{n\in\mathbb{N}}$ weakly converges

to some w in $H^1((0,1);\mathbb{R}^n)$. Passing to the limit $n \to +\infty$ in (2.39) implies that $w = S(\delta, \chi)$. The uniqueness of the limit yields that the whole sequence $(w_n)_{n\in\mathbb{N}}$ weakly converges to $S(\delta,\chi)$ in $H^1((0,1);\mathbb{R}^n)$. The convergence thus holds strongly in $L^\infty((0,1);\mathbb{R}^n)$ because of the compact embedding $H^1((0,1);\mathbb{R}^n) \hookrightarrow L^\infty((0,1);\mathbb{R}^n)$. This proves the continuity of the map S and its compactness follows again from the compact embedding $H^1((0,1);\mathbb{R}^n) \hookrightarrow L^\infty((0,1);\mathbb{R}^n)$.

(A3) Let $\delta \in [0,1]$ and $w \in L^\infty((0,1);\mathbb{R}^n)$ so that $S(\delta, w) = w$. It holds that (taking $\psi = w$ as a test function in (2.39) with $\chi = w$),

$$\frac{1}{e_k^2}\int_0^1 \partial_y w \cdot (B(w)\partial_y w) + \epsilon \int_0^1 (|\partial_y w|^2 + |w|^2) = \tag{2.42}$$

$$-\frac{\delta}{\tau}\int_0^1 (v(w) - v(w^{k-1}))\cdot w + \frac{\delta}{e_k}\varphi_k \cdot w(1) - \delta\frac{e_k'}{e_k}\int_0^1 (v(w)\cdot w + yv(w)\cdot\partial_y w). \tag{2.43}$$

Let us consider separately the different terms appearing in (2.43). First, by convexity of h, and using the fact that $w = Dh(v(w))$, it holds that

$$\frac{\delta}{\tau}\int_0^1 (v(w)-v(w^{k-1}))\cdot w = \frac{\delta}{\tau}\int_0^1 (v(w)-v(w^{k-1}))\cdot Dh(v(w)) \geq \frac{\delta}{\tau}\int_0^1 (h(v(w))-h(v(w^{k-1}))). \tag{2.44}$$

Besides, using an integration by parts,

$$\begin{aligned}
\delta\frac{e_k'}{e_k}\int_0^1 (v(w)\cdot w + yv(w)\cdot\partial_y w) &= \delta\frac{e_k'}{e_k}\left(v(w)(1)\cdot w(1) - \int_0^1 yw\cdot\partial_y v(w)\right),\\
&= \delta\frac{e_k'}{e_k}\left(v(w)(1)\cdot Dh(v(w)(1)) - \int_0^1 yDh(v(w))\cdot\partial_y v(w)\right),\\
&= \delta\frac{e_k'}{e_k}\left(v(w)(1)\cdot Dh(v(w)(1)) - \int_0^1 y\partial_y(h(v(w)))\right),\\
&= \delta\frac{e_k'}{e_k}\left(v(w)(1)\cdot Dh(v(w)(1)) - h(v(w)(1)) + \int_0^1 h(v(w))\right).
\end{aligned} \tag{2.45}$$

Using (2.26), we obtain

$$\frac{\delta}{e_k}\varphi_k\cdot w(1) = \delta\frac{e_k'}{e_k}f_k\cdot Dh(v(w)(1)). \tag{2.46}$$

Finally, using (2.43), (2.44), (2.45) and (2.46), and again the convexity of h, we obtain

$$\begin{aligned}
&\frac{\delta}{\tau}\int_0^1 h(v(w)) + \epsilon\int_0^1 (|\partial_y w|^2 + |w|^2) + \frac{1}{e_k^2}\int_0^1 \partial_y w\cdot(B(w)\partial_y w) \qquad (2.47)\\
&\leq \frac{\delta}{\tau}\int_0^1 h(v(w^{k-1})) + \delta\frac{e_k'}{e_k}\left((f_k - v(w)(1))\cdot Dh(v(w)(1)) + h(v(w)(1)) - \int_0^1 h(v(w))\right)\\
&= \frac{\delta}{\tau}\int_0^1 h(v(w^{k-1})) + \frac{e_k'}{e_k}\left(h(f_k) - \int_0^1 h(v(w))\right).
\end{aligned}$$

This inequality implies that

$$\epsilon\|w\|^2_{H^1((0,1);\mathbb{R}^n)} \leq \left(\frac{2}{\tau} + C\|\Phi\|_{L^\infty(0,T)}\right)\|h\|_{L^\infty(\overline{D})},$$

for some constant $C > 0$ independent of ϵ, τ of Φ.

All the assumptions of the Leray-Schauder fixed-point theorem are thus satisfied. This yields the existence of a fixed-point solution $w^k \in H^1((0,1);\mathbb{R}^n)$ to $S(1, w^k) = w^k$. Besides, using (2.47) with $\delta = 1$, we have the discrete entropy inequality (2.28). □

2.4.3 Proof of Proposition 2.5

Let us define by $V := \sum_{i=0}^n \overline{\phi}_i \in \mathbb{R}_+^*$, $\overline{\varphi} := (\overline{\phi}_1, \cdots, \overline{\phi}_n)^T$ and $\overline{f} := \frac{\overline{\varphi}}{V}$. From (T1), the vector $\overline{f} := \left(\overline{f}_i\right)_{1\leq i\leq n}$ obviously belongs to the set $\mathcal{D}$.

If h defined by (2.10) is an entropy density for which A satisfies assumptions (H1)-(H2)-(H3), then A satisfies the same assumptions with the entropy density

$$\overline{h} : \left\{ \begin{array}{ccc} \mathcal{D} & \to & \mathbb{R} \\ u & \mapsto & h(u) - h(\overline{f}) - Dh(\overline{f})(u - \overline{f}). \end{array} \right.$$

Indeed, for all $u \in \mathcal{D}$, $D\overline{h}(u) = Dh(u) + \overline{g}$, where $\overline{g} := Dh(\overline{f})$ is a constant vector in $\mathbb{R}^n$ and $D^2\overline{h}(u) = D^2h(u)$. Moreover, the entropy density $\overline{h}$ has the following interesting property: $\overline{f}$ is a minimizer of $\overline{h}$ on $\overline{\mathcal{D}}$ so that $\overline{h}(u) \geq \overline{h}(\overline{f}) = 0$ for all $u \in \overline{\mathcal{D}}$. In the rest of the proof, for all $w \in \mathbb{R}^n$, we will denote by $\overline{v}(w) = (\overline{v}_i(w))_{1\leq i\leq n} := (D\overline{h})^{-1}(w) = Dh^{-1}(w - \overline{g})$.

Let $\left(\overline{w}^{\epsilon,k}\right)_{k\in\mathbb{N}}$ be a sequence of solutions to the regularized time-discrete problems (2.27) defined in Lemma 2.7 with the constant fluxes $(\overline{\phi}_0, \cdots, \overline{\phi}_n)$ and the entropy density $\overline{h}$. The entropy inequality (2.28) then reads

$$\begin{aligned} &\frac{1}{\tau}\int_0^1 \overline{h}(\overline{v}(\overline{w}^{\epsilon,k})) + \epsilon\int_0^1 (|\partial_y\overline{w}^{\epsilon,k}|^2 + |\overline{w}^{\epsilon,k}|^2) + \frac{1}{e_k^2}\int_0^1 \partial_y\overline{w}^{\epsilon,k}\cdot B(\overline{w}^{\epsilon,k})\partial_y\overline{w}^{\epsilon,k} \qquad (2.48)\\ &\leq \frac{1}{\tau}\int_0^1 \overline{h}(\overline{v}(\overline{w}^{\epsilon,k-1})) + \frac{e_k'}{e_k}\left(\overline{h}(\overline{f}) - \int_0^1 \overline{h}(\overline{v}(\overline{w}^{\epsilon,k}))\right). \end{aligned}$$

In our particular case, for all $k \in \mathbb{N}$, $e_k' = V$, $e_k = e_0 + Vk\tau$ and $\overline{h}(\overline{f}) = 0$, so that we obtain

$$\frac{e_0 + V(k+1)\tau}{\tau}\int_0^1 \overline{h}(\overline{v}(\overline{w}^{\epsilon,k})) - \frac{e_0 + Vk\tau}{\tau}\int_0^1 \overline{h}(\overline{v}(\overline{w}^{\epsilon,k-1})) \leq 0.$$

This implies that for all $k \in \mathbb{N}$ and $\epsilon > 0$,

$$(e_0 + V(k+1)\tau)\int_0^1 \overline{h}(\overline{v}(\overline{w}^{\epsilon,k})) \leq (e_0 + V\tau)\int_0^1 \overline{h}(\overline{v}(w^0)). \qquad (2.49)$$

Let us denote by $\overline{w}^{(\epsilon,\tau)} : \mathbb{R}_+^* \to H^1((0,1);\mathbb{R}^n)$ the piecewise constant in time function defined by

$$\text{for a.a. } y \in (0,1), \quad \overline{w}^{(\epsilon,\tau)}(t,y) = \overline{w}^{\epsilon,k}(y) \text{ if } (k-1)\tau < t \leq k\tau.$$

Let $T > 0$ and $\xi \in L^1(0,T)$ such that $\xi \geq 0$ a.e. in $(0,T)$. Inequality (2.49) and Fubini's theorem for integrable functions implies that

$$\int_0^T \int_0^1 \left[(e_0 + V(k+1)\tau)\overline{h}(\overline{v}(\overline{w}^{(\epsilon,\tau)})) - (e_0 + V\tau)\overline{h}(\overline{v}(\overline{w}^0)) \right] \xi(t)\, dy\, dt \leq 0.$$

From the proof of Theorem 2.4, we know that up to the extraction of a subsequence which is not relabeled, $\left(\overline{v}(\overline{w}^{(\epsilon,\tau)})\right)_{\epsilon,\tau>0}$ converges strongly in $L^2_{\text{loc}}(\mathbb{R}_+^*; L^2((0,1);\mathbb{R}^n))$ and a.e. in $\mathbb{R}_+^* \times (0,1)$ as ϵ and τ go to zero to a global weak solution v to (2.19). Using Lebesgue dominated convergence theorem, and passing to the limit $\epsilon, \tau \to 0$ in the above inequality yields

$$\int_0^T \int_0^1 \left[(e_0 + Vt)\overline{h}(v) - e_0\overline{h}(\overline{v}(w^0)) \right] \xi(t)\, dy\, dt \leq 0,$$

which implies that there exists $C > 0$ such that for almost all $t > 0$,

$$(e_0 + Vt) \int_0^1 \overline{h}(v) \leq C, \tag{2.50}$$

which yields inequality (2.21). In the rest of the proof, C will denote an arbitrary positive constant independent on the time $t > 0$.

Furthermore, since $v \in H^1((0,T);\ (H^1((0,1);\ \mathbb{R}^n))') \cap L^2((0,T); H^1((0,1);\mathbb{R}^n))$, it holds that $v \in \mathcal{C}^0((0,T); L^2((0,1);\mathbb{R}^n))$ from [LM12], and the Lebesgue dominated convergence theorem implies that $t \in \mathbb{R}_+^* \mapsto \int_0^1 \overline{h}(v(t,y))\, dy$ is a continuous function. Inequality (2.50) then holds for all $t > 0$.

For all $0 \leq i \leq n$, let us denote by $\overline{v_i}(t) := \int_0^1 v_i(t,y)\, dy$. By convention, we define $v_0(t,y) := 1 - \rho_{v(t,y)}$ and $\overline{f}_0 := 1 - \rho_{\overline{f}}$. It can be checked from the weak formulation of (2.27) that

$$\int_0^1 \overline{v}_i\left(\overline{w}^{\epsilon,k}\right) = \frac{k\overline{\phi}_i\tau + e_0 \int_0^1 v_i^0}{e_0 + V(k+1)\tau}.$$

Passing to the limit $\epsilon, \tau \to 0$ using the Lebesgue dominated convergence theorem, we obtain that for almost all $t > 0$,

$$\overline{v}_i(t) = \frac{e_0 \int_0^1 v_i^0(y)\, dy + t\overline{\phi}_i}{e_0 + Vt},$$

so that $|\overline{v}_i(t) - \overline{f}_i| \leq \dfrac{C}{e_0 + Vt}$. The continuity of $\overline{v}_i$ implies that this equality holds for all $t > 0$.

The Csizàr-Kullback inequality states that for all $t > 0$,

$$\begin{aligned} \|v_i(t,\cdot) - \overline{v}_i(t)\|^2_{L^1(0,1)} &\leq 2 \int_0^1 v_i(t,y) \log \frac{v_i(t,y)}{\overline{v}_i(t)}\, dy \\ &= 2 \int_0^1 v_i(t,y) \log \frac{v_i(t,y)}{\overline{f}_i}\, dy + 2 \int_0^1 v_i(t,y) \log \frac{\overline{f}_i}{\overline{v}_i(t)}\, dy. \end{aligned}$$

Thus,

$$\begin{aligned}\sum_{i=0}^{n}\left\|v_i(t,\cdot)-\overline{f}_i\right\|_{L^1(0,1)} &\leq \sum_{i=0}^{n}\|v_i(t,\cdot)-\overline{v}_i(t)\|_{L^1(0,1)}+|\overline{f}_i-\overline{v}_i(t)|\\ &\leq \sqrt{2\int_0^1 \overline{h}(v)}+\sum_{i=0}^{n}\left[\sqrt{2\left|\log\frac{\overline{v}_i(t)}{\overline{f}_i}\right|}+|\overline{f}_i-\overline{v}_i(t)|\right]\\ &\leq \sqrt{\frac{C}{e_0+Vt}}.\end{aligned}$$

Hence inequality (2.22) and the result.

2.4.4 Proof of Proposition 2.6

Let $(\Phi^m)_{m\in\mathbb{N}}\subset\Xi$ be a minimizing sequence for $\mathcal{J}$ i.e such that

$$\lim_{m\to+\infty}\mathcal{J}(\Phi^m)=\inf_{\Phi\in\Xi}\mathcal{J}(\Phi).$$

By definition of the set Ξ, the sequence $(\Phi^m)_{m\in\mathbb{N}}$ is bounded in $L^\infty(0,T)$. Thus, up to a non relabeled extraction, it weakly-* converges to some limit $\Phi^*\in\Xi$ in $L^\infty(0,T)$. As a consequence, $\left(\frac{d}{dt}e_{\Phi^m}\right)_{m\in\mathbb{N}}$ (respectively $(e_{\Phi^m})_{m\in\mathbb{N}}$) converges weakly-* (respectively strongly) in $L^\infty(0,T)$ to $\frac{d}{dt}e_{\Phi^*}$ (respectively e_{Φ^*}).

For each $m\in\mathbb{N}$, let v_{Φ^m} be the unique global weak solution to (2.19) associated to the fluxes Φ^m. Its uniqueness is a consequence of assumption (C1). From the bounds obtained in the proof of Theorem 2.4 and the boundedness of $(\Phi^m)_{m\in\mathbb{N}}$ in $L^\infty(0,T)$, it holds that the sequences $\|\partial_t v_{\Phi^m}\|_{L^2((0,T);(H^1(0,1))')}$, $\|A(v_{\Phi^m})\partial_y v_{\Phi^m}\|_{L^2((0,T);L^2(0,1))}$ and $\|\partial_y v_{\Phi^m}\|_{L^2((0,T);L^2(0,1))}$ are also uniformly bounded in m.

Thus, up to the extraction of a subsequence which is not relabeled, using the compact injection of $L^2((0,T);H^1((0,1);\mathbb{R}^n))\cap H^1((0,T);(H^1((0,1);\mathbb{R}^n))')$ into $\mathcal{C}((0,T);L^2((0,1);\mathbb{R}^n))$ (see [LM12]), there exists $v_*\in L^2((0,T);H^1((0,1);\mathbb{R}^n))\cap H^1((0,T);(H^1((0,1);\mathbb{R}^n))')$ and $V_*\in L^2((0,T);L^2((0,1);\mathbb{R}^n))$ so that

$$\begin{aligned}v_{\Phi^m}&\rightharpoonup v_* \text{ weakly in } L^2((0,T);H^1((0,1);\mathbb{R}^n))\cap H^1((0,T);(H^1((0,1);\mathbb{R}^n))'),\\ v_{\Phi^m}&\longrightarrow v_* \text{ strongly in } \mathcal{C}((0,T);L^2((0,1);\mathbb{R}^n)) \text{ and a.e. in } (0,T)\times(0,1),\\ A(v_{\Phi^m})\partial_y v_{\Phi^m}&\rightharpoonup V_* \text{ weakly in } L^2((0,T);L^2((0,1);\mathbb{R}^n)).\end{aligned}$$

Using similar arguments as in the proof of Theorem 2.4, we also obtain that V_* is necessarily equal to $A(v_*)\partial_y v_*$. Passing to the limit $m\to+\infty$, we obtain that for all $\psi\in L^2((0,T);H^1((0,1);\mathbb{R}^n))$,

$$\begin{aligned}&\int_0^T\int_0^1 \partial_t v_*\cdot\psi\,dt\,dy+\int_0^T\int_0^1\frac{1}{e_{\Phi^*}(t)^2}\partial_y\psi\cdot(A(v_*)\partial_y v_*)\,dt\,dy\\ &+\int_0^T\frac{\frac{d}{dt}e_{\Phi^*}(t)}{e_{\Phi^*}(t)}\int_0^1(v_*\cdot\psi+yv_*\cdot\partial_y\psi)\,dt\,dy=\int_0^T\frac{1}{e_{\Phi^*}(t)}\varphi_*(t)\cdot\psi(1)\,dt.\end{aligned}$$

Assumption (C1) yields $v_* = v_{\Phi^*}$. The above convergence results then imply that

$$\mathcal{J}\left(\Phi^m\right) \underset{m\to+\infty}{\longrightarrow} \mathcal{J}(\Phi^*),$$

and hence Φ^* is a minimizer of problem (2.24). Hence the result.

2.5 Numerical tests

In this section, we present some numerical tests illustrating the results of Section 2.3 on the prototypical example of Section 2.2.1. In Section 2.5.1, we present the numerical scheme used in our simulations to compute an approximation of a solution of (2.19). In Section 2.5.2 and Section 2.5.3, some numerical tests which illustrate Proposition 2.5 and Proposition 2.6 are detailed.

2.5.1 Discretization scheme

In view of the optimization problem (2.24) we are aiming at, it appears that a fully implicit unconditionally stable scheme is needed to allow the use of reasonably large time steps.

We present here the numerical scheme used for the discretization of (2.19), for the particular model presented in Section 2.2.1. We do not provide a rigorous numerical analysis for this scheme here.

Let $M \in \mathbb{N}^*$ and $\Delta t := \frac{T}{M}$. We define for all $0 \leq m \leq M$, $t_m := m\Delta t$. The discrete external fluxes are characterized for every $0 \leq i \leq n$ by vectors $\widehat{\phi}_i := \left(\widehat{\phi}_i^m\right)_{1\leq m\leq M} \in \mathbb{R}_+^M$, where $\widehat{\phi}_i^m = \int_{t_{m-1}}^{t_m} \phi_i(s)\, ds$. For every $1 \leq m \leq M$, the thickness of the thin film and it derivative at time t_m are approximated respectively by

$$e_m := e_0 + \sum_{p=1}^{m}\sum_{i=0}^{n} \widehat{\phi}_i^p \Delta t \approx e(t_m), \qquad \text{and } e_m^d := \sum_{i=0}^{n} \widehat{\phi}_i^m \approx e'(t_m).$$

In addition, let $Q \in \mathbb{N}^*$ and $\Delta y := \frac{1}{Q}$ and $y_q := (q-0.5)\Delta y$. For all $0 \leq i \leq n$, $1 \leq q \leq Q$ and $0 \leq m \leq M$, we denote by $v_i^{m,q}$ the finite difference approximation of v_i at time t_m and point $y_q \in (0,1)$. Here again, we use the convention that $v_0 = 1 - \rho_v$.

We use a centered second-order finite difference scheme for the diffusive part of the equation, and a first-order upwind scheme for the advection part, together with a fully implicit time scheme. Assuming that the approximation $\left(v_i^{m-1,q}\right)_{0\leq i\leq n, 1\leq q\leq Q}$ is known, one computes $(\widetilde{v}_i^{m,q})_{0\leq i\leq n,\ 1\leq q\leq Q}$ as solutions of the following sets of equations.

For all $0 \leq i \leq n$ and $2 \leq q \leq Q-1$,

$$\frac{\left(\widetilde{v}_i^{m,q} - v_i^{m-1,q}\right)}{\Delta t} = \frac{e_m^d}{e_m} y_q \left(\frac{\widetilde{v}_i^{m,q+1} - \widetilde{v}_i^{m,q}}{\Delta y}\right) + \sum_{0 \leq j \neq i \leq n} \frac{K_{ij}}{e_m^2} \widetilde{v}_j^{m,q} \left(\frac{\widetilde{v}_i^{m,q+1} + \widetilde{v}_i^{m,q-1} - 2\widetilde{v}_j^{m,q}}{2\Delta y^2}\right) - \sum_{0 \leq j \neq i \leq n} \frac{K_{ij}}{e_m^2} \widetilde{v}_i^{m,q} \left(\frac{\widetilde{v}_j^{m,q+1} + \widetilde{v}_j^{m,q-1} - 2\widetilde{v}_j^{m,q}}{2\Delta y^2}\right) \tag{2.51}$$

together with boundary conditions which reads for all $0 \leq i \leq n$,

$$\sum_{0 \leq j \neq i \leq n} \frac{K_{ij}}{e_m} \left[\widetilde{v}_j^{m,1} \left(\frac{\widetilde{v}_i^{m,2} - \widetilde{v}_i^{m,1}}{\Delta y}\right) - \widetilde{v}_i^{m,1} \left(\frac{\widetilde{v}_j^{m,2} - \widetilde{v}_j^{m,1}}{\Delta y}\right)\right] = 0, \tag{2.52}$$

$$\sum_{0 \leq j \neq i \leq n} \frac{K_{ij}}{e_m} \left[\widetilde{v}_j^{m,Q} \left(\frac{\widetilde{v}_i^{m,Q-1} - \widetilde{v}_i^{m,Q}}{\Delta y}\right) - \widetilde{v}_i^{m,Q} \left(\frac{\widetilde{v}_j^{m,Q-1} - \widetilde{v}_j^{m,Q}}{\Delta y}\right)\right] = -e_m^d \widetilde{v}_i^{m,Q} + \widehat{\phi}_i^m. \tag{2.53}$$

The nonlinear system of equations (2.51)-(3.6)-(2.53), whose unknowns are $(\widetilde{v}_i^{m,q})_{0 \leq i \leq n, 1 \leq q \leq Q}$ is solved using Newton iterations with initial guess $\left(v_i^{m-1,q}\right)_{0 \leq i \leq n, 1 \leq q \leq Q}$. The obtained solution does not satisfy in general the desired non-negativeness and volumic constraints. This is the reason why an additional projection step is performed. For all $0 \leq i \leq n$ and $1 \leq q \leq Q$, we define

$$v_i^{m,q} := \frac{[\widetilde{v}_i^{m,q}]_+}{\sum_{j=0}^n [\widetilde{v}_j^{m,q}]_+},$$

so that

$$v_i^{m,q} \geq 0 \quad \text{and} \quad \sum_{j=0}^n v_j^{m,q} = 1.$$

We numerically observe that this scheme is unconditionally stable with respect to the choice of discretization parameters Δt and Δy.

A standard practice in the production of thin film CIGS (Copper, Indium, Gallium, Selenium) solar cells by means of PVD process is to consider piecewise-constant external fluxes. In the following numerical tests, we consider time-dependent functions of the form

$$\phi_{\mathrm{i}}(t) = \begin{cases} \alpha_1^i & 0 < t \leq \tau_1^i, \\ \alpha_2^i & \tau_1^i < t \leq \tau_2^i, \\ \alpha_3^i & \tau_2^i < t \leq T, \end{cases} \tag{2.54}$$

where $0 < \tau_1^i < \tau_2^i < T$ and $(\alpha_1^i, \alpha_2^i, \alpha_3^i) \in (\mathbb{R}_+)^3$ are non-negative constants for all $0 \leq i \leq n$. Besides, we consider initial condition of the form

$$v_i^0(y) = \frac{w_i(y)}{\sum_{j=0}^n w_j(y)} \quad \forall 0 \leq i \leq n, \tag{2.55}$$

where $w_i : [0,1] \to \mathbb{R}_+$ are functions which will be precised below. In the whole section, system (2.19) is simulated with four species (i.e. $n = 3$).

In Figure 2.2 are plotted the results obtained for the simulation of (2.19) with the following parameters :

- $T = 200$, $M = 200$, $Q = 100$, $\Delta t = 1$, $\Delta y = 0.01$, $e_0 = 1$.
- Cross-diffusion coefficients K_{ij}

	$j=0$	$j=1$	$j=2$	$j=3$
$i=0$	0	0.1141	0.0776	0.0905
$i=1$	0.1141	0	0.0646	0.0905
$i=2$	0.0776	0.0646	0	0.0905
$i=3$	0.0905	0.0905	0.0905	0

- External fluxes of the form (2.54) with $\tau_1^i = 66$ and $\tau_2^i = 132$ for every $0 \leq i \leq n$ and with

	$i=0$	$i=1$	$i=2$	$i=3$
α_1^i	0.9	2	0.2	0.7
α_2^i	1.4	1.5	1.2	0.3
α_3^i	0.9	2	0.2	0.7

- Initial concentrations v_i^0 of the form (2.55) with $w_0(y) = y$, $w_1(y) = 2y$, $w_2(y) = \sqrt{y}$ and $w_3(y) = 0$.

The profile of the external fluxes is plotted in Figure 2.2-(a). In Figure 2.2-(b) and Figure 2.2-(c) are given respectively the the initial and the final concentrations of the four species.

2.5.2 Long-time behaviour results

In this section is given a numerical illustration of Proposition 2.5. We consider time-dependent functions of the form

$$\phi_i(t) = \beta_i, \quad \forall 0 \leq t \leq T. \tag{2.56}$$

where $(\beta_i)_{0 \leq i \leq n} \in (\mathbb{R}_+^*)^{n+1}$. In Figure 2.3 are plotted the results obtained for the the simulation of (2.19) with the following parameters :

- $T = 2000$, $M = 2000$, $Q = 100$, $\Delta t = 1$, $\Delta y = 0.01$, $e_0 = 1$.
- Cross-diffusion coefficients K_{ij}

	$j=0$	$j=1$	$j=2$	$j=3$
$i=0$	0	0.1141	0.0776	0.0905
$i=1$	0.1141	0	0.0646	0.0905
$i=2$	0.0776	0.0646	0	0.0905
$i=3$	0.0905	0.0905	0.0905	0

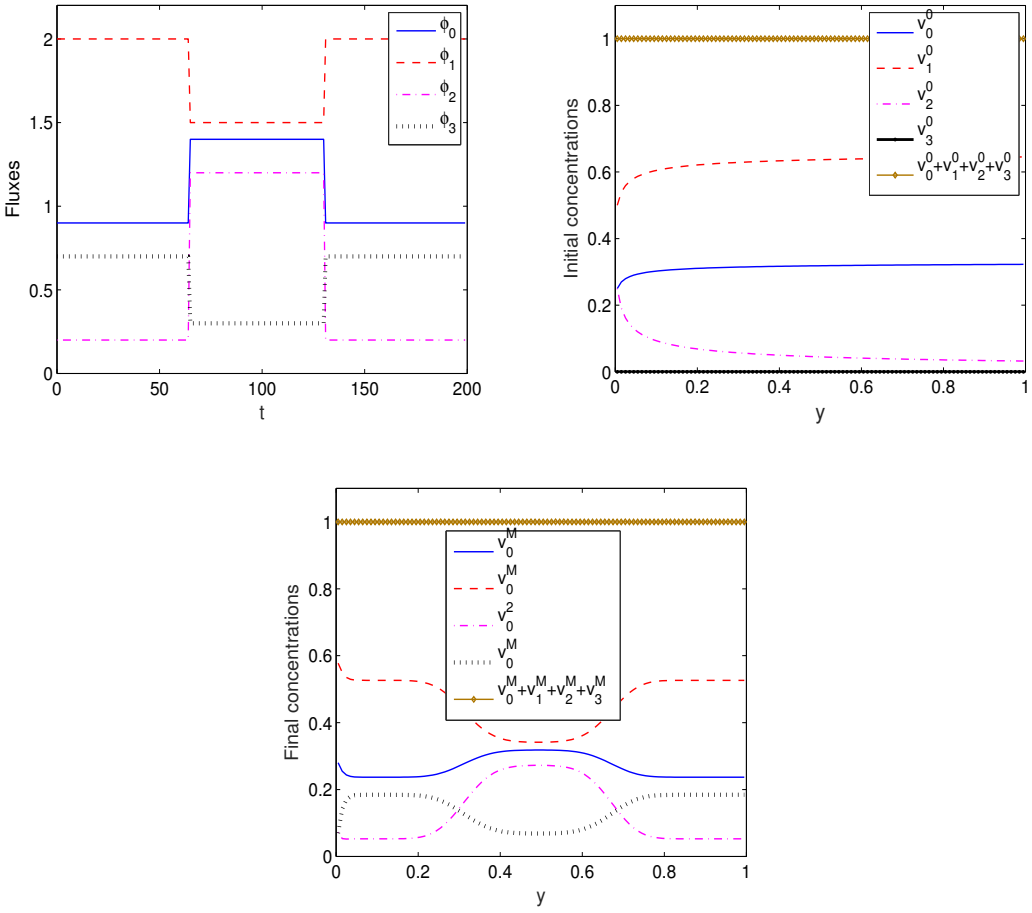


Figure 2.2 – Simulation of (2.19).

- External fluxes of the form (2.56) with

	i=0	i=1	i=2	i=3
β^i	0.9	0.8	1.7	0.5

- Initial concentrations v_i^0 of the form (2.55) with

$$w_0(y) = \exp\left(-\frac{(y-0.5)^2}{0.04}\right), \quad w_1(y) = y^2, \quad w_2(y) = 1 - w_0(y), \quad w_3(y) = |\sin(\pi y)|.$$

For all $0 \leq i \leq n$, let $\bar{v}_i := \beta_i / \sum_{j=0}^n \beta_j$. We consider the time-dependent quantity

$$\gamma(t) = \frac{1}{\overline{h}(v(t,\cdot))}$$

where the relative entropy $\overline{h}$ is defined in (2.21). We also consider the quantities

$$\eta_i(t) = \frac{1}{\|v_i(t,\cdot) - \bar{v}_i\|^2_{L^1(0,1)}}$$

and

$$\eta(t) = \frac{1}{\sum_{i=0}^{n} \|v_i(t,\cdot) - \bar{v}_i\|^2_{L^1(0,1)}}$$

In Figure 2.3-(a) and 2.3-(b) are plotted respectively the initial and the final concentration profiles.

The evolution of $(\eta_i(t))_{0\leq i\leq n}$ (respectively $\eta(t)$ and $\gamma(t)$) with respect to t is shown in Figure 2.3-(c) (respectively 2.3-(d) and 2.3-(e)). We numerically observe that these quantities are affine functions of t in the asymptotic regime which illustrates the theoretical result of Proposition 2.5.

2.5.3 Optimization of the fluxes

The optimization problem (2.24) is solved in practice using an adjoint formulation associated to the discretization scheme described in Section 2.5.1. We refer the reader to Chapter 3 for more details and comparisons between our model and experimental results obtained in the context of thin film CIGS solar cell fabrication. To illustrate Proposition 2.6, we proceed as follows: first, we perform a simulation of (2.19) with external fluxes Φ_{sim} for a duration T to obtain a final thickness $e_{\Phi_{\text{sim}}}(T)$ and final concentrations $v_{\Phi_{\text{sim}}}(T,\cdot)$, then, we solve the minimization problem (2.24) to obtain optimal fluxes Φ^* where the target concentrations are

$$v_{\text{opt}}(y) = v_{\Phi_{\text{sim}}}(T,y) \quad \forall\ y \in (0,1)$$

and the target thickness is

$$e_{\text{opt}} = e_{\Phi_{\text{sim}}}(T).$$

Lastly, we perform another simulation of (2.19) with the obtained optimal fluxes Φ^* and compare the final concentrations v_{Φ^*} and the final thickness e_{Φ^*} to the target concentrations v_{opt} and the target thickness e_{opt}.

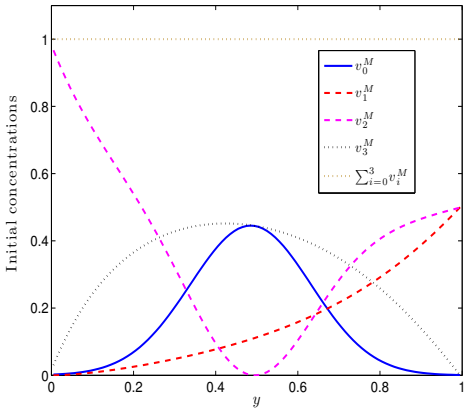
In Figures 2.4-(a), 2.5-(a), 2.6-(a) and 2.7-(a) are plotted the final concentration profiles $v_{\Phi_{\text{sim}}}(T,\cdot)$ resulting from the simulation of (2.19) with the following parameters :

- $T = 120$, $M = 120$, $Q = 100$, $\Delta t = 1$, $\Delta y = 0.01$, $e_0 = 1$.
- Cross-diffusion coefficients K_{ij}

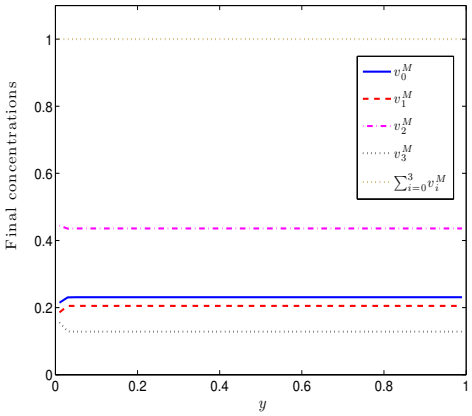
	$j=0$	$j=1$	$j=2$	$j=3$
$i=0$	0	0.1141	0.0776	0.0905
$i=1$	0.1141	0	0.0646	0.0905
$i=2$	0.0776	0.0646	0	0.0905
$i=3$	0.0905	0.0905	0.0905	0

- External fluxes Φ_{sim} of the form (2.54) with

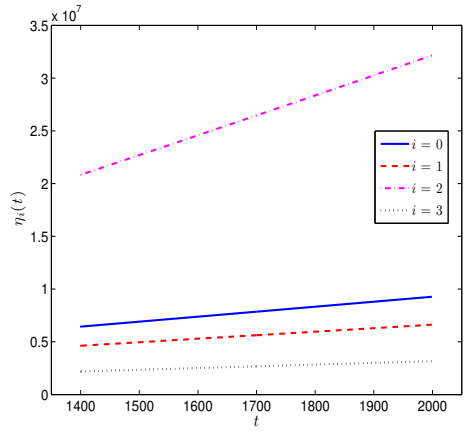
	$i=0$	$i=1$	$i=2$	$i=3$
α_1^i	0.9	2	0.2	0.7
α_2^i	1.4	1.5	1.2	0.3
α_3^i	0.9	2	0.2	0.7



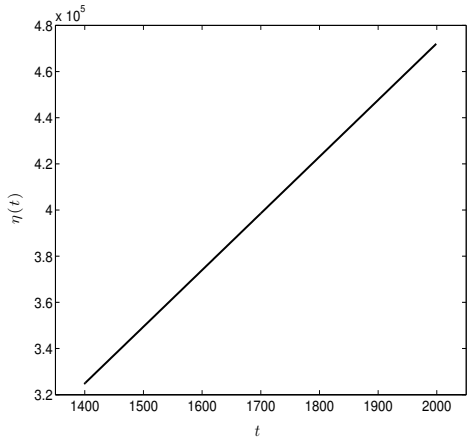
(a)



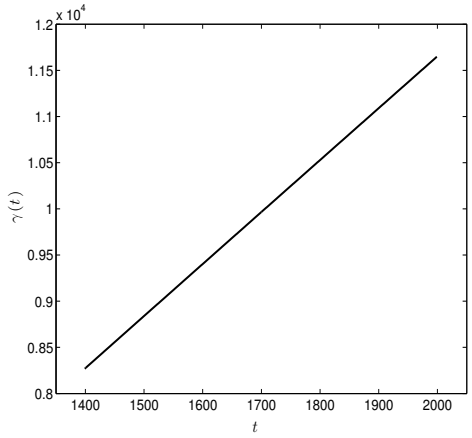
(b)



(c)



(d)



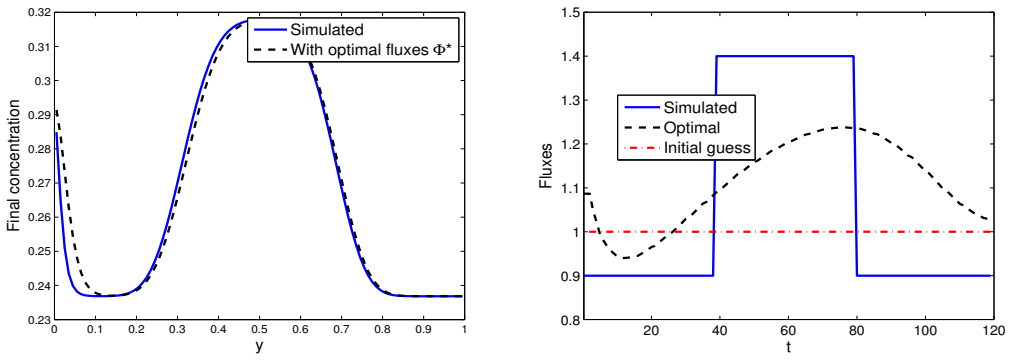
(e)

Figure 2.3 – Long-time behavior in the case of non negative constant external fluxes.

- Initial concentrations v_i^0 of the form (2.55) with $w_0(y) = y$, $w_1(y) = 2y$, $w_2(y) = \sqrt{y}$ and $w_3(y) = 0$.

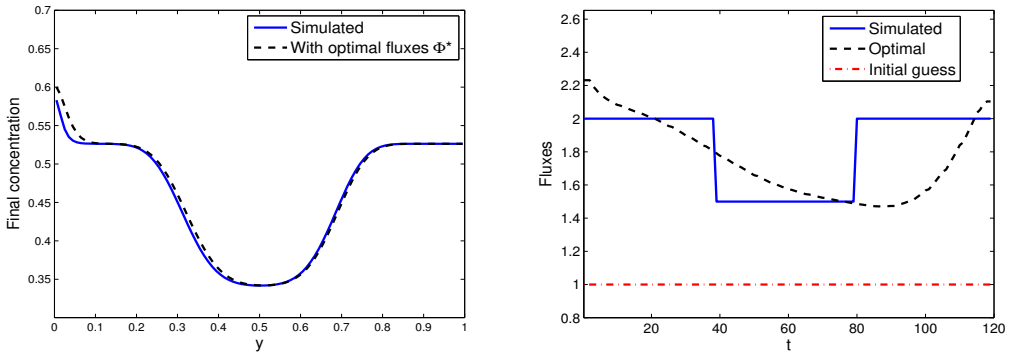
We use a quasi-Newton iterative gradient algorithm for the resolution of the minimization problem. At each iteration of the algorithm, the approximate hessian is updated by means of a BFGS procedure and the optimal step size is the solution of a line search subproblem. More details on the numerical optimization algorithms can be found in [JBS06]. The initial guess Φ^0 is taken of the form (2.56) where $\beta^i = 1$ for all $0 \leq i \leq n$.

The algorithm is run until one of the following stopping criterion is reached : either $(\mathcal{J}(\Phi) \leq \varepsilon)$ or $(\|\nabla_\Phi \mathcal{J}(\Phi)\|_{L^2} \leq \nu)$ with $\varepsilon = 10^{-5}$ and $\nu = 10^{-5}$.



(a) (b)

Figure 2.4 – Reconstruction of the final concentration of the species $i = 0$.

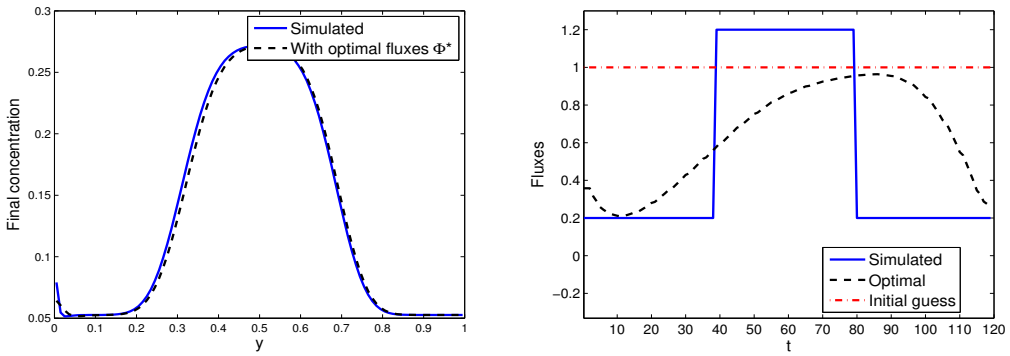


(a) (b)

Figure 2.5 – Reconstruction of the final concentration of the species $i = 1$.

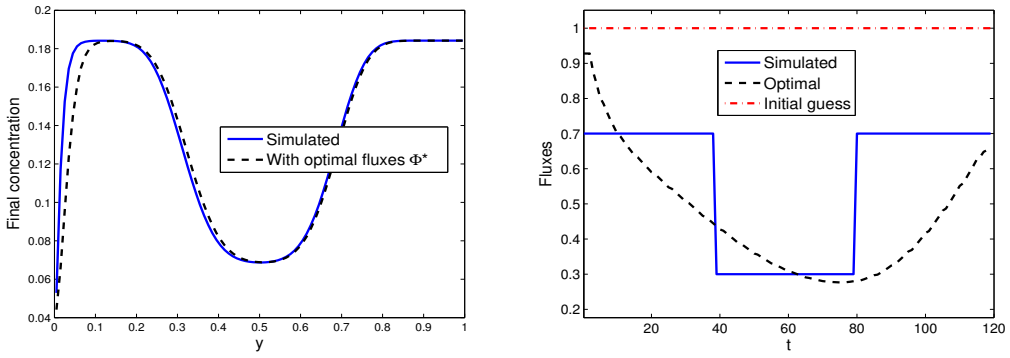
In Figure 2.8-(a) we plot the evolution of the value of the cost $\mathcal{J}(\Phi)$ with respect to the number of iterations.

We numerically observe that all the concentrations are well reconstructed and that the value of the optimal thickness $e_{\Phi^*} = 483.4022$ is close to the target thickness $e_{\Phi_{\text{sim}}} = 483.4$. Unlike the external fluxes Φ_{sim}, the optimal fluxesΦ^* are not piecewise constant. These tests show that the uniqueness of a solution to the optimization problem (2.24) can not be expected in general.



(a) (b)

Figure 2.6 – Reconstruction of the final concentration of the species $i = 2$.



(a) (b)

Figure 2.7 – Reconstruction of the final concentration of the species $i = 3$.

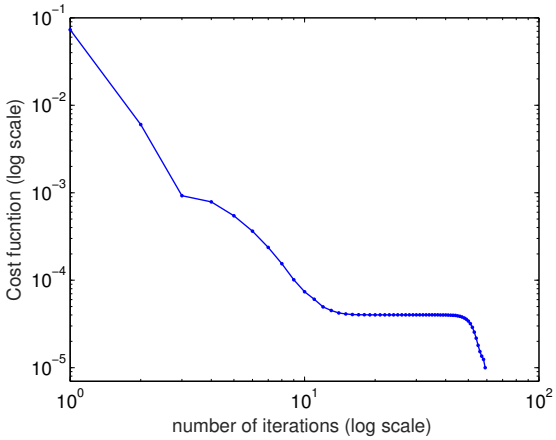


Figure 2.8 – Convergence of the BFGS gradient descent algorithm for the minimization problem (2.24).

2.6 Conclusion

In this work, we propose and analyze a one-dimensional model for the description of a PVD process. The evolution of the local concentrations of the different chemical species in the bulk of the growing layer is described via a system of cross-diffusion equations similar to the ones studied in [BDFPS10, Jue15a]. The growth of the thickness of the layer is related to the external fluxes of atoms that are absorbed at the surface of the film.

The existence of a global weak solution to the final system using the boundedness by entropy method under assumptions on the diffusion matrix of the system close to those needed in [Jue15a] is established. In addition, the entropy density h is required to be continuous (hence bounded) on the set $\overline{\mathcal{D}} = \left\{u = (u_i)_{1\leq i\leq n} \in \mathbb{R}_+^n,\ \sum_{i=1}^n u_i \leq 1\right\}$.

We prove the existence of a solution to an optimization problem under the assumption that there exists a unique global weak solution to the obtained system, whatever the value of the external fluxes.

Lastly, in the case when the entropy density is defined by $h(u) = \sum_{i=1}^n u_i \log u_i + (1-\rho_u)\log(1-\rho_u)$, we prove in addition that, when the external fluxes are constant and positive, the local concentrations converge in the long time to a constant profile at a rate which scales like $\mathcal{O}\left(\frac{1}{t}\right)$.

A discretization scheme, which is observed to be unconditionnaly stable, is introduced for the discretization of (2.19). This scheme enables to preserve constraints (2.5) at the discretized level.

We see this work as a preliminary step before tackling related problems in higher dimension, including surfacic diffusion effects. Besides, the proof of assumption (C1) remains an open question in general at least to our knowledge. Lastly, a nice theoretical question which is not tackled in this paper, but will be the object of future research, is the characterization of the set of reachable concentration profiles.

Acknowledgements

We are very grateful to Eric Cancès and Tony Lelièvre for very helpful advice and discussions. We would like to thank Jean-François Guillemoles, Marie Jubeault, Torben Klinkert and Sana Laribi from IRDEP, who introduced us to the problem of modeling PVD processes for the fabrication of thin film solar cells. The EDF company is acknowledged for funding. We would also like to thank Daniel Matthes for very helpful discussions on the theoretical part and the reviewers whose comments helped in very significantly improving the quality of the paper.

2.7 Appendix

2.7.1 Formal derivation of the diffusion model (2.3)

We present in this section a simplified formal derivation of the cross-diffusion model (2.3) from a one-dimensional microscopic lattice hopping model with size exclusion, in the same spirit than the one proposed in [BDFPS10].

We consider here a solid occupying the whole space $\mathbb{R}$ and discretize the domain using a uniform grid of step size $\Delta x > 0$. At any time $t \in [0, T]$, we denote by $u_i^{k,t}$ the number of atoms of type i ($0 \leq i \leq n$) in the k^{th} interval $[k\Delta x, (k+1)\Delta x)$ ($k \in \mathbb{Z}$). Let $\Delta t > 0$ denote a small enough time step. We assume that during the time interval Δt, an atom i located in the k^{th} interval can exchange its position with an atom of type j ($j \neq i$) located in one of the two neighbouring intervals with probability $p_{ij} = p_{ji} > 0$. In average, we obtain the following evolution equation for $u_i^{k,t}$:

$$\begin{aligned} u_i^{k,t+\Delta t} - u_i^{k,t} &= \sum_{0 \leq j \neq i \leq n} p_{ij} \left(u_i^{k+1,t} u_j^{k,t} + u_i^{k-1,t} u_j^{k,t} - u_i^{k,t} u_j^{k+1,t} - u_i^{k,t} u_j^{k-1,t} \right) \\ &= \sum_{0 \leq j \neq i \leq n} p_{ij} \left[u_j^{k,t} \left(u_i^{k+1,t} + u_i^{k-1,t} - 2u_i^{k,t} \right) - u_i^{k,t} \left(u_j^{k+1,t} + u_j^{k-1,t} - 2u_j^{k,t} \right) \right]. \end{aligned}$$

This yields that

$$\frac{u_i^{k,t+\Delta t} - u_i^{k,t}}{\Delta t} = \frac{2\Delta x^2}{\Delta t} \sum_{0 \leq j \neq i \leq n} p_{ij} \left[u_j^{k,t} \frac{u_i^{k+1,t} + u_i^{k-1,t} - 2u_i^{k,t}}{2\Delta x^2} - u_i^{k,t} \frac{u_j^{k+1,t} + u_j^{k-1,t} - 2u_j^{k,t}}{2\Delta x^2} \right].$$

Choosing Δt and Δx so that these quantities satisfy a classical diffusion scaling $\frac{2\Delta x^2}{\Delta t} = \alpha > 0$, denoting by $K_{ij} := \alpha p_{ij}$ and letting the time step and grid size go to 0, we formally obtain the following equation for the evolution of u_i on the continuous level:

$$\partial_t u_i = \sum_{0 \leq j \neq i \leq n} K_{ij} \left(u_j \Delta_x u_i - u_i \Delta_x u_j \right),$$

which is identical to the system of equations (2.3) introduced in the first section. Of course, this formal argument can be easily extended to any arbitrary dimension.

2.7.2 Leray-Schauder fixed-point theorem

Theorem 2.8 (Leray-Schauder fixed-point theorem)**.** *Let B be a Banach space and $S : B \times [0, 1] \to B$ be a continuous map such that*

(A1) $S(x, 0) = 0$ for each $x \in B$;

(A2) S is a compact map;

(A3) there exists a constant $M > 0$ such that for each pair $(x, \sigma) \in B \times [0, 1]$ which satisfies $x = S(x, \sigma)$, we have $\|x\| < M$.

Then, there exists a fixed-point $y \in B$ satisfying $y = S(y, 1)$.

CHAPTER 3

SIMULATION OF CIGS LAYER PRODUCTION PROCESS

In this chapter, we report some results of our collaboration work with the IRDEP lab.

Abstract. The one-dimensional model proposed and theoretically analyzed in Chapter 2 is extended to take into account the evolution of the temperature during the production process of CIGS thin film layer. An Arrhenius law is introduced in order to take into account the temperature dependence of the cross-diffusion coefficients and additional surface absorption rates are introduced in order to have a more realistic simulation of the co-evaporation process. Lastly, an inverse problem is proposed for the calibration of the values of the diffusion coefficients and the absorption rates from experimental measures. The numerical method used to solve the inverse problem is described and some numerical results are presented.

Contents

3.1 Présentation du modèle

Rappelons d'abord de manière succincte, le procédé de déposition de la couche CIGS par le procédé *Physical Vapor Deposition* [Mat10]. Une couche de molybdène est d'abord déposée sur un substrat de verre. Le "wafer" obtenu est ensuite introduit dans un four plasma dans lequel sont injectées sous forme gazeuse les différentes entités atomiques qui vont former la couche de CIGS (à savoir le Cuivre, l'Indium, le Gallium et le Sélénium). Voir la Figure 3.1.

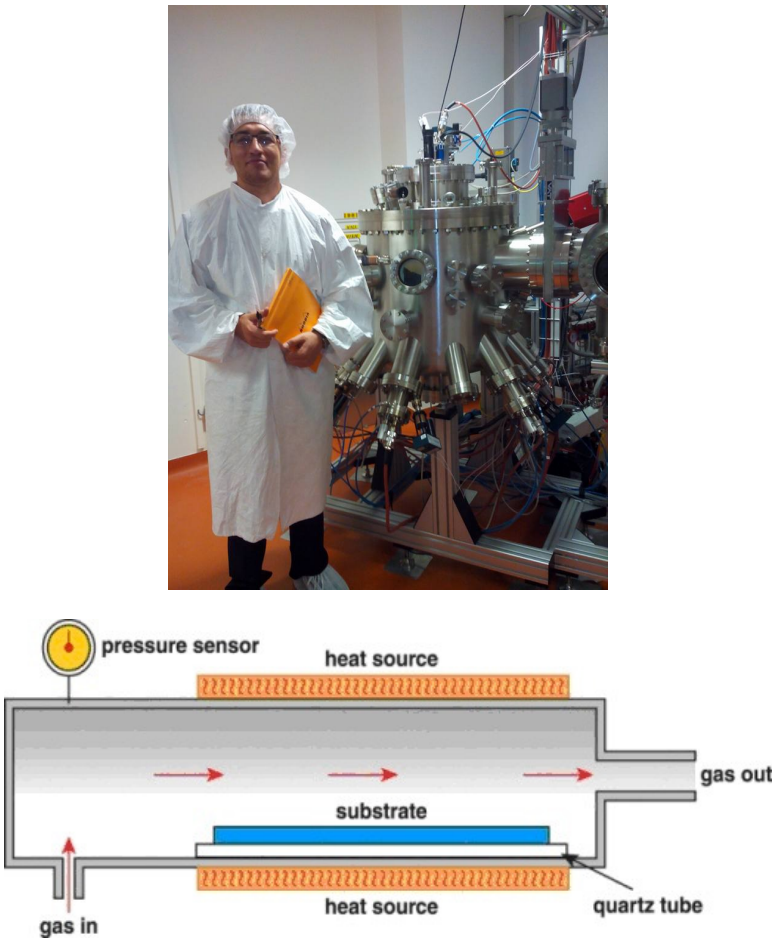


Figure 3.1 – Le procédé de co-évaporation pour la fabrication de cellules solaires à couches minces de type CIGS. En haut, une vraie image du four plasma (photos prise lors d'une visite à l'IRDEP en juin 2015). En bas, un schéma explicatif du procédé. Source : *https://www.azonano.com*

Le film mince de CIGS croît au fur et à mesure que les atomes des différentes espèces chimiques injectées se déposent sur le substrat. De plus, comme la température de l'échantillon est maintenue au cours du procédé à un niveau très élevé, les atomes des différentes espèces chimiques diffusent à l'intérieur du film ainsi formé. Les deux phénomènes suivants doivent donc être pris en compte : la diffusion inter-espèce due à la température élevée du système et la croissance du film due à la déposition des atomes au cours du temps. Notre modèle proposé et analysé théoriquement dans le Chapitre 2 permet de prendre en compte ces deux phénomènes mais sous les hypothèses simplificatrices suivantes : la diffusion inter-espèce est indépendante de la température et toutes les espèces sont absorbées par le film de la même manière. Ces deux hypothèses ne permettent pas de reproduire fidèlement les résultats des expériences. Rappelons, pour mémoire, les équations de notre modèle avant de présenter les extensions qui permettront de l'améliorer et ainsi reproduire plus fidèlement les résultats expérimentaux obtenus par l'IRDEP.

On note $T > 0$ la durée totale du procédé de fabrication et $\mathcal{A}$ l'ensemble des différentes espèces chimiques mises en présence lors de la croissance du film, qui sont dans

notre cas : le cuivre (Cu), l'indium (In), le gallium (Ga), le sélénium (Si) et le molybdène (Mo). Les atomes de sélénium sont situés sur un sous sous-réseau cristallin et ne diffusent pas avec les autre espèces chimiques. Aussi, on peut considérer seulement l'ensemble $\mathcal{A} := \{Cu,\ In,\ Ga,\ Mo\}$. Pour tout $t \in (0,T)$, nous notons $e(t)$ l'épaisseur

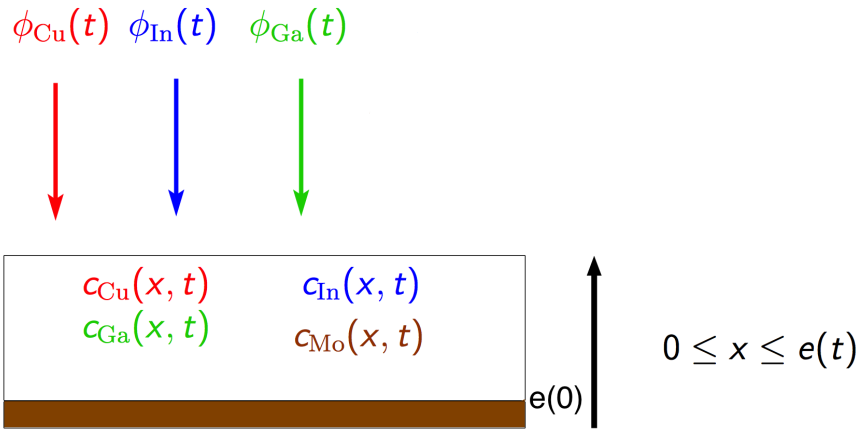


Figure 3.2 – Schéma simplifié du procédé PVD pour la fabrication de la couche CIGS.

du film mince à l'instant t. De plus, pour tout $x \in (0, e(t))$ et tout $A \in \mathcal{A}$, nous notons $c_A(x,t)$ la concentration en l'espèce A à l'instant t et à la profondeur x de l'échantillon.

On suppose qu'à l'instant $t = 0$ (correspondant au début du procédé), le film occupe un domaine $(0, e_0)$ avec $e_0 > 0$ (typiquement $e_0 = 1\mu m$ représente la couche de molybdène déposée en début de fabrication), et que les profils de concentration à l'instant initial sont connus. Pour tout $x \in (0, e_0)$ et $A \in \mathcal{A}$, $c_A^0(x)$ désigne la concentration à l'instant $t = 0$ en l'entité A localement au point $x \in (0, e_0)$. On suppose de plus que ces profils initiaux vérifient les conditions suivantes:

$$\forall x \in (0,e_0), \quad c_A(0,x) = c_A^0(x) \quad \text{avec } c_A^0(x) \ge 0 \quad \text{et} \quad \sum_{A\in\mathcal{A}} c_A^0(x) = 1.$$

Pour tout $t \in (0,T)$, et pour tout $A \in \mathcal{A}$, on note $\phi_A(t)$ la valeur du flux de l'espèce A imposé à l'instant t lors de la croissance du film. On suppose que $\phi_A : (0,T) \to \mathbb{R}_+$ est à valeurs positives. Enfin, notons K_{AB} le coefficient de diffusion des atomes de l'espèce A avec les atomes de l'espèce B pour chaque couple d'espèce $(A,B) \in \mathcal{A}^2$. L'ensemble des équations qui régissent la dynamique de notre système, en fonction des flux, de l'épaisseur initiale et des profils de concentrations est donnée par

$$\begin{cases} e(t) = e_0 + \int_0^t \Sigma_{A\in\mathcal{A}} \phi_A(s)\, ds, & t \in (0,T), \\ \partial_t c_A = \mathrm{div}_x J_A(t,x,c), & (t,x) \in (0,T)\times(0,e(t)), \\ J_A(t,0,c) = 0, & t \in (0,T), \\ J_A(t,e(t),c) + e'(t) c_A(t,e(t)) = \phi_A(t), & t \in (0,T), \\ c_A(0,x) = c_A^0(x), & x \in (0,e_0), \end{cases} \tag{3.1}$$

où le flux $J_A : (0,T)\times(0,e(t)) \mapsto \mathbb{R}$ est défini par

$$J_A(t,x,c) = \sum_{B\in\mathcal{A},\ B\neq A} K_{AB}\left(c_B \nabla_x c_A - c_A \nabla_x c_B\right) \tag{3.2}$$

Introduisons maintenant les trois extensions que nous ajoutons au modèle (3.1) afin qu'il soit plus descriptif.

1. La température : Notons par $\theta : (0,T) \mapsto \mathbb{R}_+^*$ le profil de température à l'intérieur du film au cours du temps, que l'on suppose homogène pour simplifier.

2. Les coefficients de diffusion : Considérons que la dépendance des coefficients de diffusion inter-espèces en la températeure est donnée par une loi d'Arrhénius. Plus précisément, la valeur du coefficient K_{AB} entre l'espèce A et l'espèce B en fonction de la température $\theta(t)$ de l'échantillon est supposée être de la forme

$$K_{AB}(\theta(t)) = D_{AB} \exp\left(-\frac{E_{AB}}{\kappa\theta(t)}\right)$$

où κ est la constante de Boltzmann[1], D_{AB} et E_{AB} sont des constantes positives à déterminer pour chaque couple $(A,B) \in \mathcal{A}^2$.

3. Les taux d'absorption : Lors du procédé d'évaporation, les taux d'absorption des différentes espèces chimiques par la surface le film en formation ne sont pas forcément égaux (les atomes de cuivre sont par exemple mieux absorbés que ceux de l'indium ou du gallium). On introduit donc pour chaque espèce $A \in \mathcal{A}$ un paramètre $\lambda_A \in \mathbb{R}_+$ qui modélise le taux d'absorption des atomes de type A à la surface du film en cours de formation. Ceux-ci, pour simplifier, sont supposés être indépendats de la tempéréture de l'échantillon.

Le nouveau système d'équations prenant en comptes ces extensions s'écrit alors:

$$\begin{cases} e(t) = e_0 + \int_0^t \Sigma_{A\in\mathcal{A}} \lambda_A \phi_A(s)\, ds, & t \in (0,T), \\ \partial_t c_A = \mathrm{div}_x J_A(t,x,c), & (t,x) \in (0,T)\times(0,e(t)), \\ J_A(t,0,c) = 0, & t \in (0,T), \\ J_A(t,e(t)c) + e'(t)c_A(e(t),t) = \lambda_A \phi_A(t), & t \in (0,T), \\ c_A(0,x) = c_A^0(x), & x \in (0,e_0) \end{cases} \tag{3.3}$$

avec le flux J_A donné par

$$J_A(t,x,c) = \sum_{B\in\mathcal{A},\ B\neq A} K_{AB}(\theta(t))\,(c_B \nabla_x c_A - c_A \nabla_x c_B) \tag{3.4}$$

Bien que les résultats produits par ce modèle simple soient satisfaisants, il est important de signaler ici quelques limitations de ce modèle.

D'une part, les effets de géométrie 3D (rugosité de surface, inhomogénéités longitudinales,...) ne sont pas pas pris en compte dans ce modèle 1D. D'autre part, la formation et la propagation de défauts dans la structure cristalline du film ne sont pas prises en compte. Toutefois, l'évolution des défauts de types lacunes ou impuretés peut être décrite par ce modèle en considérant ces défauts comme des espèces chimiques supplémentaires. Enfin, lors du dépôt des atomes dans le four à haute température, il se trouve que des réactions chimiques peuvent avoir lieu entre les différentes espèces comme par exemple la réaction $CuInSe_2 \leftrightharpoons Cu + In + 2Se$. Ces éventuelles réactions chimiques sont également ignorées.

[1] $\kappa = 1,38064852 \times 10^{-23} \mathrm{m^2 kgs^{-2} K^{-1}}$.

3.2 Discrétisation du système d'EDP

Cette section est dédiée au schéma numérique utilisée pour lé résolution du système (3.3). Le schéma présenté ici a déjà été partiellement décrit dans la Section 2.5.1 du chapitre précédent. Cela dit, pour des raisons de clarté et dans le but de fixer les notations qu'on utilisera dans la suite, nous allons rappeler brièvement la dérivation du schéma. A l'aide du changement de variables $y = \frac{x}{e(t)}$, les équations (3.3) sont reformulées afin de se ramener à un domaine de référence $(0,1)$ qui ne dépend plus du temps. Le système suivant d'inconnues $u_A : (0,T) \times (0,1) \ni (t,y) \mapsto u_A(t,y) \in \mathbb{R}$ pour $A \in \mathcal{A}$ est ainsi obtenu

$$\begin{cases} e(t) = e_0 + \int_0^t \Sigma_{A\in\mathcal{A}} \lambda_A \phi_A(s)\, ds, & t \in (0,T), \\ \partial_t u_A = \frac{e'(t)}{e(t)} y \partial_y u_A + \frac{1}{e^2(t)} \mathrm{div}_y J_A(t,y,u), & (t,y) \in (0,T)\times(0,1), \\ \frac{1}{e(t)} J_A(t,0,u) = 0, & t \in (0,T), \\ \frac{1}{e(t)} J_A(t,1,u) + e'(t) u_A(t,1) = \lambda_A \phi_A(t), & t \in (0,T), \\ u_A(0,y) = u_A^0(y), & y \in (0,1), \end{cases} \tag{3.5}$$

Afin de résoudre numériquement le système (3.5)-(3.4), on introduit les grilles de discrétisation uniformes en temps et en espace suivantes $\{0 = t_0, t_1, \cdots, t_N = T\}$ et $\{0 = y_0, y_1, \cdots, y_I = 1\}$ où $t_n := n\Delta t$ pour $1 \le n \le N$ et $y_i := i\Delta y$ pour $1 \le i \le I$ avec $\delta t = \frac{T}{N}$ et $\delta y = \frac{1}{I}$ pour des valeurs $N, I \in \mathbb{N}^*$ choisies. Le profil de température ainsi que les flux sont discrétisés comme suit

$$\Phi_A^n \approx \int_{t_{n-1}}^{t_n} \phi_A(s) ds, \quad \theta^n \approx \int_{t_{n-1}}^{t_n} \theta(s) ds, \quad \forall A \in \mathcal{A}, \forall 1 \le n \le N.$$

Notons $e_0 = e^0$ l'épaisseur initiale de l'échantillon. Pour tout $1 \le n \le N$, l'épaisseur de la couche ainsi que sa dérivée sont approchées respectivement par

$$e_n = e_0 + \sum_{k=1}^{n} \sum_{A\in\mathcal{A}} \lambda_A \Phi_A^k \Delta t \approx e(t_n), \quad e_n^d = \sum_{A\in\mathcal{A}} \lambda_A \Phi_A^n \approx e'(t_n).$$

L'approximation par des différences finies de la solution continue u_A à l'instant $t_n \in (0,T)$ au point $y_i \in (0,1)$ sera notée $u_A^{i,n}$. On approche le terme de diffusion par une différence finie d'ordre deux et le terme d'advection par une différence finie d'ordre un décentrée en les points $y_1 \le y_i \le y_{I-1}$ et centrée en le point y_I. On utilise un schéma d'Euler implicite pour la discrétisation en temps. Supposons que l'approximation $(u_A^{i,n-1})_{A\in\mathcal{A}}$ soit connue, alors $(\tilde{u}_A^{i,n})_{A\in\mathcal{A}}$ est obtenue comme solution du système suivant : Pour tout $A \in \mathcal{A}$ et $2 \le i \le I-1$,

$$\begin{aligned} \frac{\left(\widetilde{u}_A^{i,n} - u_A^{i,n-1}\right)}{\Delta t} &= \frac{e_n^d}{e_n} y_i \left(\frac{\widetilde{u}_A^{i+1,n} - \widetilde{u}_A^{i,n}}{\Delta y} \right) \\ &+ \sum_{B\neq A} \frac{K_{AB}(\theta^n)}{e_n^2} \widetilde{u}_B^{i,n} \left(\frac{\widetilde{u}_A^{i+1,n} + \widetilde{u}_A^{i-1,n} - 2\widetilde{u}_A^{i,n}}{2\Delta y^2} \right) \\ &- \sum_{B\neq A} \frac{K_{AB}(\theta^n)}{e_n^2} \widetilde{u}_A^{i,n} \left(\frac{\widetilde{u}_B^{i+1,n} + \widetilde{u}_B^{i-1,n} - 2\widetilde{u}_B^{i,n}}{2\Delta y^2} \right) \end{aligned}$$

avec des conditions au bords qui s'écrivent pour tout $A \in \mathcal{A}$ comme

$$\sum_{B\neq A} \frac{K_{AB}(\theta^n)}{e_n}\left(\widetilde{u}_B^{1,n}\left(\frac{\widetilde{u}_A^{2,n}-\widetilde{u}_A^{1,n}}{\Delta y}\right)-\widetilde{u}_A^{1,n}\left(\frac{\widetilde{u}_B^{2,n}-\widetilde{u}_B^{1,n}}{\Delta y}\right)\right)=0 \tag{3.6}$$

$$\sum_{B\neq A} \frac{K_{AB}(\theta^n)}{e_n}\left(\widetilde{u}_B^{I,n}\left(\frac{\widetilde{u}_A^{I-1,n}-\widetilde{u}_A^{I,n}}{\Delta y}\right)-\widetilde{u}_A^{I,n}\left(\frac{\widetilde{u}_B^{I-1,n}-\widetilde{u}_B^{I,n}}{\Delta y}\right)\right)=-e_n^d\widetilde{u}_A^{I,n}+\lambda_A\phi_A^n. \tag{3.7}$$

L'ensemble de ces équations peut s'écrire de manière équivalente sous forme matricielle: pour tout $A \in \mathcal{A}$ et tout $1 \leq n \leq N$:

$$\frac{\left(\widetilde{u}_A^n-u_A^{n-1}\right)}{\Delta t}=\frac{e_n^d}{e_n}Mu_A^n+P_A^n+\sum_{B\neq A}\frac{K_{AB}(\theta^n)}{e_n^2}\left(\widetilde{u}_B^n\odot D\widetilde{u}_A^n-\widetilde{u}_A^n\odot D\widetilde{u}_B^n\right) \tag{3.8}$$

où pour chaque $1 \leq n \leq N$ et chaque $A \in \mathcal{A}$, on pose $\tilde{u}_A^n = (\tilde{u}_A^{1,n}, \cdots, \tilde{u}_A^{I,n})$ et où les matrices $P_A^n \in \mathbb{R}^I$, $M \in \mathbb{R}^{I\times I}$ et $D \in \mathbb{R}^{I\times I}$ sont données par

$$\forall 1 \leq i \leq I, \quad (P_A^n)_i = \delta_{iI}\frac{\phi_A^n}{2\Delta y e_n}$$

$$\forall 1 \leq j \leq I; 1 \leq i \leq I-1, \quad (M)_{i,j} = \delta_{i,j}\frac{-y_i}{\Delta y}+\delta_{i+1,j}\frac{y_{i+1}}{\Delta y},$$

$$(M)_{I,I-1} = \frac{-y_i}{\Delta y},$$

$$(M)_{I,I} = \frac{y_i}{\Delta y}-\frac{y_i}{\Delta y}$$

et

$$\forall 1 \leq i,j \leq I-1, \quad (D)_{i,j} = \delta_{i,j}\frac{-2}{2\Delta y^2}+\delta_{i+1,j}\frac{1}{2\Delta y^2}+\delta_{i,j+1}\frac{1}{2\Delta y^2},$$

$$(D)_{1,1} = (D)_{I,I} = \frac{-1}{2\Delta y^2},$$

$$(D)_{1,2} = (D)_{I-1,I} = \frac{1}{2\Delta y^2}.$$

Le produit de Hadamard noté $\odot$ étant défini pour toutes matrices $A, B \in \mathbb{R}^{m\times n}$ comme suit

$$\forall 1 \leq i \leq m \ , \ 1 \leq j \leq n, \quad (A\odot B)_{i,j} := (A)_{i,j}(B)_{i,j}.$$

Le système non linéaire (3.8) d'inconnus $(\widetilde{u}_A^n)_{1\leq i\leq I}$, pour $A \in \mathcal{A}$ est en pratique résolu par une méthode itérative de Newton prenant u_A^{n-1} comme point initial. Cependant, les solutions obtenues ne satisfont en général pas les contraintes de positivité et de renormalisation souhaitées. Aussi, on applique une étape de projection pour garantir l'obtention de concentration comprises entre 0 et 1, dont la somme vaut 1. Pour tout $A \in \mathcal{A}$ et $1 \leq i \leq I$, on pose

$$u_A^{i,n} = \frac{f(\widetilde{u}_A^{i,n})}{\sum_{B\in\mathcal{A}} f(\widetilde{u}_B^{i,n})}, \quad \text{où} \quad f : \left\{\begin{array}{ccc} \mathbb{R} & \to & [0,1] \\ x & \mapsto & f(x) = \max(0, \min(1,x)). \end{array}\right.$$

Cette dernière étape du schéma numérique permet d'assurer que pour tout $1 \leq i \leq I$, pour tout $1 \leq n \leq N$ et pout tout $A \in \mathcal{A}$

$$0 \leq u_A^{i,n} \leq 1 \text{ et } \sum_{A\in\mathcal{A}} u_A^{i,n} = 1.$$

3.3 Post-traitement des données expérimentales

Dans cette étude nous avons considéré un ensemble de M profils expérimentaux fournis par l'équipe de l'IRDEP. A chacune des M expériences, une cellule photovoltaïque de type CIGS a été fabriquée selon le procédé de co-évaporation. Les concentrations finales des différentes entités chimiques ont été ensuite mesurées pour chaque cellule.

Pour chaque expérience $1 \leq m \leq M$, on récupère l'épaisseur finale de la cellule $e_m(T_m)$ et les valeurs des concentrations en des points spécifiques sur une grille uniforme $e_1 = y_1 < y_2 < y_3 < ... < y_I = e_m(T_m)$ où $e_1 = 85$nm. Les mesures ne peuvent pas être faites sur toute l'épaisseur du film en raison de contraintes techniques.

L'évaporation des atomes s'effectue en plusieurs étapes : d'abord, l'Indium et le Gallium sont évaporés sous une température T_1 pendant un temps t_1 ensuite le Cuivre sous une température T_2 pendant un temps t_2, et après un temps de repos $\Delta\tau$, l'Indium et le Gallium sont évaporés de nouveau sous la température $T_3 = T_2$ et pendant un temps t_3. Le solide est finalement retiré du four et laissé (refroidir) à la température ambiante pendant un temps t_4. Voir le schéma à la Figure 3.3.

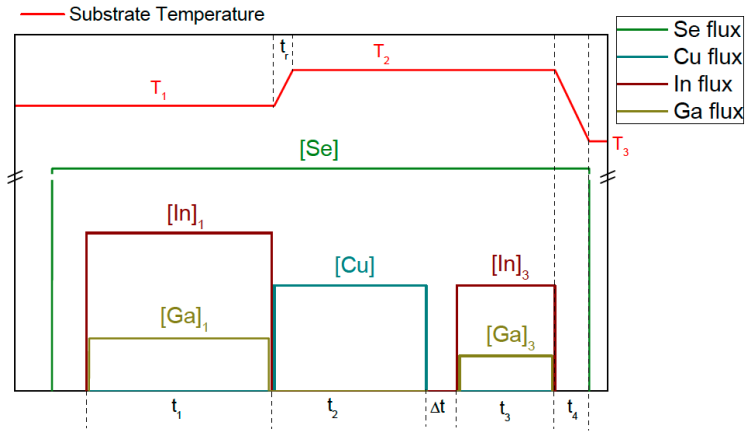


Figure 3.3 – Profils de température et des flux utilisés dans le production des couches minces de CIGS pour cette étude. Le protocole expérimental est détaillé dans la thèse [Kli15]

Pour chaque expérience $1 \leq m \leq M$, les durées $\{t_1^m, t_2^m, t_3^m, t_4^m, \Delta\tau^m\}$ et les températures $\{\theta_1^m; \theta_2^m; \theta_3^m\}$ de chaque régime ainsi que les valeurs $\{\phi_{In,m}, \phi_{Ga,m}, \phi_{Cu,m}\}$ des flux injectés au cours du temps sont récupérés. Comme les mesures de concentrations sont faites seulement à partir de l'épaisseur e_1 et que le sélénium n'est pas considéré dans le modèle, on renormalise les flux de sorte que la relation de conservation de masse suivante soit vérifiée

$$\sum_{n=1}^{N} \frac{T_m}{N}(\phi_{Ga,m} + \phi_{In,m} + \phi_{Cu,m}) = e_m(T_m) - e_m(0) \tag{3.9}$$

où $T_m := (t_1^m + t_2^m + t_3^m + t_4 + \Delta\tau^m)/N$ est le temps total de l'expérience. L'épaisseur initiale $e_0 = 1\mu$m est la même pour chaque expérience car le substrat est recouvert initialement d'une couche de 1μm de Molybdène pur.

3.4 Calibration du modèle

Nous expliquons dans cette section, comment les valeurs optimales pour les pré-facteurs D_{AB}, les énergies d'activation E_{AB} et les taux d'absorption λ_A sont identifées à partir de données expérimentales. Il s'agit de résoudre un problème inverse de calibration de paramètres.

3.4.1 Formulation du problème inverse.

Pour chacune des expériences $1 \leq m \leq M$, on note $(z^i_{A,m})_{1\leq i\leq I} =: z_{A,m} \in \mathbb{R}^I$ le profil mesuré de concentration finale de l'espèce A. On note T_m le temps total de l'expérience et $e_m(T_m)$ l'épaisseur finale du film produit.

Pour alléger les notations, on introduit l'ensemble $\Omega := \mathbb{R}^{2|\mathcal{A}|} \times \mathbb{R}^{2|\mathcal{A}|} \times [0,1]^{|\mathcal{A}|}$. Considérons la fonction $\mathcal{J} : \Omega \ni (D,E,\lambda) \mapsto \mathcal{J}(D,E,\lambda) = \sum_{m=1}^M \mathcal{J}_m((D,E,\lambda)) \in \mathbb{R}^+$ avec pour chaque $1 \leq m \leq M$,

$$\mathcal{J}_m(D,E,\lambda) := \sum_{A\in\mathcal{A}} \sum_{i=1}^{I} \Delta y \left(u_A^{i,N} - z^i_{A,m}\right)^2$$

où $u^N_{A,m} := (u^{i,N}_{A,m})_{1\leq i\leq I}$ est la solution au temps final t_N du système (3.3) par le schéma numérique décrit dans la Section 3.2 avec les valeurs (D,E,λ). L'objectif est donc de résoudre le problème d'optimisation sous contraintes suivant

$$(D^*,E^*,\lambda^*) \in \operatorname{argmin}_{(D,E,\lambda)\in\Omega} \mathcal{J}(D,E,\lambda), \tag{3.10}$$

$$1 \leq m \leq M, \quad e_0 + \sum_{A\in\mathcal{A}} \sum_{k=1}^{N} \lambda_A \phi^k_{A,m} = e_m(T_m). \tag{3.11}$$

Ce problème est en pratique résolu par une méthode de gradient itérative adaptée aux problèmes d'optimisation sous contraintes: *l'optimisation quadratique successive* (SQP)[NW06, GMSW17].

Remarque. La même expression de la fonction de coût est utilisée pour trouver le profil de température optimal ainsi que les profils de flux optimaux permettant d'atteindre des concentrations finales cibles. De plus, les arguments théoriques présentés dans le Chapitre 2 pour prouver l'existence de solutions au problème d'optimisation des flux sont applicables au problème d'optimisation du profil de température.

3.4.2 Calcul du gradient par approche duale

Dans cette section, nous présentons le calcul du gradient de la fonctionnelle de coût par rapport aux différentes variables $\eta \in \{(D_{AB}, E_{AB}, \lambda_A\} \in \Omega$. L'indice m sera omis dans la suite afin d'alléger les notations.

D'abord, une simple application des règles de dérivations donne

$$\partial_\eta \mathcal{J}(\eta) = 2 \sum_{A\in\mathcal{A}} \Delta y \langle u^N_A - z_A, \partial_\eta u^N_A \rangle$$

où la notation $\langle z_A^N, z_A \rangle$ désigne le produit scalaire usuel $\sum_{i=1}^I u_A^{i,N} z_A^i$ entre les vecteur $u_A^N \in \mathbb{R}^I$ et $z_A \in \mathbb{R}^I$. Le terme $\partial_\eta u_A^n$ est calculé par approche duale comme suit :

Etape 1: dynamique de $\partial_\eta u_A^n$. En dérivant les termes de l'équation (3.8) par rapport à la variable η, on obtient la dynamique suivante vérifiée par $\partial_\eta \widetilde{u}_A^n$ pour tout $1 \leq n \leq N$.

$$
\begin{aligned}
\partial_\eta \widetilde{u}_A^n = & \partial_\eta u_A^{n-1} \\
& + \Delta t \partial_\eta P_A^n \\
& + \Delta t \partial_\eta \left(\frac{e_n^d}{e_n} \right) M \widetilde{u}_A^n + \Delta t \left(\frac{e_n^d}{e_n} \right) M \partial_\eta \widetilde{u}_A^n \\
& + \Delta t \sum_{B \in \mathcal{A}, B \neq A} \partial_\eta \left(\frac{K_{AB}(\theta^n)}{e_n^2} \right) (\widetilde{u}_B^n \odot D\widetilde{u}_A^n - \widetilde{u}_A^n \odot D\widetilde{u}_B^n) \\
& + \Delta t \sum_{B \in \mathcal{A}, B \neq A} \left(\frac{K_{AB}(\theta^n)}{e_n^2} \right) (\partial_\eta \widetilde{u}_B^n \odot D\widetilde{u}_A^n) \\
& + \Delta t \sum_{B \in \mathcal{A}, B \neq A} \left(\frac{K_{AB}(\theta^n)}{e_n^2} \right) (\widetilde{u}_B^n \odot D\partial_\eta \widetilde{u}_A^n) \\
& + \Delta t \sum_{B \in \mathcal{A}, B \neq A} \left(\frac{K_{AB}(\theta^n)}{e_n^2} \right) (-\widetilde{u}_A^n \odot D\partial_\eta \widetilde{u}_B^n) \\
& + \Delta t \sum_{B \in \mathcal{A}, B \neq A} \left(\frac{K_{AB}(\theta^n)}{e_n^2} \right) (-\partial_\eta \widetilde{u}_A^n \odot D\widetilde{u}_B^n) \, .
\end{aligned}
$$

Reformulé autrement,

$$
\partial_\eta \widetilde{u}_A^n = \partial_\eta u_A^{n-1} + H_A^n + \sum_{B \in \mathcal{A}} G_{AB}^n \partial_\eta \widetilde{u}_B^n \tag{3.12}
$$

où pour tout $A, B \in \mathcal{A}$, les matrices $H_A \in \mathbb{R}^I$ et $G_{AB} \in \mathbb{R}^{I \times I}$ sont données par

$$
H_A^n = \partial_\eta P_A^n + \Delta t \partial_\eta \left(\frac{e_n^d}{e_n} \right) M \widetilde{u}_A^n + \Delta t \sum_{B \in \mathcal{A}, B \neq A} \partial_\eta \left(\frac{K_{AB}(\theta^n)}{e_n^2} \right) (\widetilde{u}_B^n \odot D\widetilde{u}_A^n - \widetilde{u}_A^n \odot D\widetilde{u}_B^n) \, .
$$

et

$$
G_{AB}^n = \begin{cases} \Delta t \left(\frac{e_n^d}{e_n} M \right) + \Delta t \sum_{B' \in \mathcal{A}, B' \neq A} \left(\frac{K_{AB'}(\theta^n)}{e_n^2} \right) (\Psi_{B'}^n - \mathrm{diag}(D\widetilde{u}_{B'}^n)) \, , & \text{si } A = B, \\ \Delta t \left(\frac{K_{AB}(\theta^n)}{e_n^2} \right) (\mathrm{diag}(D\widetilde{u}_A^n) - \Psi_A^n) \, , & \text{si } A \neq B \end{cases}
$$

avec les matrices $(\Psi_A)_{A \in \mathcal{A}} \in \mathbb{R}^{I \times I}$ définies pour tous $1 \leq i, j \leq I$ par $(\Psi_A^n)_{i,j} = D_{i,j} \widetilde{u}_A^{i,n}$.

Notons maintenant par $\widetilde{U}^n := (\widetilde{u}_A^n)_{A \in \mathcal{A}} \in \mathbb{R}^{I \times |\mathcal{A}|}$ et par $H^n := (H_A^n)_{A \in \mathcal{A}} \in \mathbb{R}^{I \times |\mathcal{A}|}$ et définissons par blocs la matrice $O^n \in \mathbb{R}^{(I \times |\mathcal{A}|) \times (I \times |\mathcal{A}|)}$ comme suit

$$
O_{AB}^n = \begin{cases} I - G_{AA}^n, & \text{si } A = B, \\ -G_{AB}^n, & \text{si } A \neq B. \end{cases}
$$

La dynamique (3.12) s'écrit alors de manière équivalente comme

$$\partial_\eta \widetilde{U}^n = (O^n)^{-1}\partial_\eta U^{n-1} + (O^n)^{-1}H^n(\widetilde{u}).$$

Rajoutons maintenant l'étape de projection. Pour ce faire on introduit la matrice $W^n \in \mathbb{R}^{(I\times|\mathcal{A}|)\times(I\times|\mathcal{A}|)}$ de sorte à obrenir la relation

$$W^n \partial_\eta \widetilde{U}^n = \partial_\eta U^n.$$

Pour tout $1 \leq n \leq N$, la matrice $W^n \in \mathbb{R}^{(I\times|\mathcal{A}|)\times(I\times|\mathcal{A}|)}$ est définie par blocs :

$$W^n := \begin{pmatrix} \Gamma^n_{AA} & \Gamma^n_{AB} & \Gamma^n_{AC} & \cdots \\ \Gamma^n_{BA} & \Gamma^n_{BB} & \Gamma^n_{BC} & \cdots \\ \Gamma^n_{CA} & \Gamma^n_{CB} & \Gamma^n_{CC} & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{pmatrix}.$$

Finalement, la dynamique de $\partial_\eta U^n$ s'écrit

$$\begin{aligned} \frac{\partial_\eta U^n - \partial_\eta U^{n-1}}{\Delta t} &= \frac{(I - W^n(O^n)^{-1})}{\Delta t}\partial_\eta U^{n-1} + \frac{W^n(O^n)^{-1}H^n}{\Delta t} \\ &=: L^n \partial_\eta U^{n-1} + X^n. \end{aligned}$$

Etape 2: problème dual. On cherche la solution $Q^n \in \mathbb{R}^{I\times|\mathcal{A}|}$ pour $1 \leq n \leq N$ au problème suivant

$$Q^N_A = 2\sum_{A\in\mathcal{A}} \Delta y(u^N_A - z_A) \tag{3.13}$$

$$Q^{n-1} = Q^n + \Delta t(L^n)^T Q^n. \tag{3.14}$$

Les equations (3.13) et (3.14) définissent un problème adjoint (car il fait intervenir les solutions u^n_A du problème direct) et backward en temps car la solution à chaque instant $n-1$ est donnée en fonction de la solution à l'instant n. Il faut donc d'abord résoudre le problème direct pour obtenir les solutions u^n_A, ensuite résoudre le problème dual par récurrence inversée.

Etape 3: l'expression du gradient. On a maintenant tous les ingrédients nécessaires pour évaluer le gradient $\nabla_\eta \mathcal{J}$ à l'aide du calcul suivant :

$$\begin{aligned} \sum_{n=1}^N \langle \frac{-Q^n + Q^{n-1}}{\Delta t}, \partial_\eta U^{n-1}\rangle &= \sum_{n=1}^N \langle (L^n)^T Q^n, \partial_\eta U^{n-1}\rangle \\ &= \sum_{n=1}^N \langle Q^n, L^n \partial_\eta U^{n-1}\rangle \\ &= \sum_{n=1}^N \langle Q^n, \frac{\partial_\eta U^n - \partial_\eta U^{n-1}}{\Delta t}\rangle - \sum_{n=1}^N \langle Q^n, X^n\rangle. \end{aligned}$$

Ainsi

$$\begin{aligned}\sum_{n=1}^{N}\langle Q^n, X^n\rangle &= \sum_{n=1}^{N}\langle Q^n, \frac{\partial_\eta U^n - \partial_\eta U^{n-1}}{\Delta t}\rangle - \sum_{n=1}^{N}\langle \frac{-Q^n + Q^{n-1}}{\Delta t}, \partial_\eta U^{n-1}\rangle \\ &= \frac{1}{\Delta t}\sum_{n=1}^{N}\langle Q^n, \partial_\eta U^n\rangle - \langle Q^n, \partial_\eta U^{n-1}\rangle + \langle Q^n, \partial_\eta U^{n-1}\rangle - \langle Q^{n-1}, \partial_\eta U^{n-1}\rangle \\ &= \frac{1}{\Delta t}\langle Q^N, \partial_\eta U^N\rangle - \langle Q^0, \partial_\eta U^0\rangle \\ &= \frac{2}{\Delta t}\sum_{A\in\mathcal{A}} \Delta y \langle (u_A^N - z_A), \partial_\eta u_A^N\rangle.\end{aligned}$$

En conclusion, on a

$$\partial_\eta \mathcal{J}(\eta) = \Delta t \sum_{n=1}^{N}\langle Q^n, X^n\rangle.$$

3.4.3 Résultats numériques

Dans cette section nous présentons quelques résultats numériques obtenus en utilisant le logiciel de calcul scientifique MATLAB.

Les valeurs optimales obtenues pour les pré-facteurs D^*_{AB}, les énergies d'activation E^*_{AB} ainsi que les taux d'absorption λ^*_A sont données respectivement dans les Tables 3.2, 3.1 et 3.3. Les paramètres numériques utilisés sont les suivants :

- Les données : $M = 12$ (calibration sur 12 profils expérimentaux).
- Point initial de la minimisation : pour tout $A, B \in \{\mathrm{Cu}, \mathrm{In}, \mathrm{Ga}, \mathrm{Mo}\}$,

$$\begin{aligned} D^0_{AB} &= 10^{-3} \times (1 - \delta_{AB})\ \mu\mathrm{m}^2\mathrm{min}^{-1}, \\ E^0_{AB} &= 10^{-1} \times (1 - \delta_{AB})\ \mathrm{eV}, \\ \lambda^0_A &= 1. \end{aligned}$$

- Taille des grilles en temps et en espace : $\Delta t = 0.5$ et $I = 50$.

Pour illustrer le résulat de la calibration, nous prenons une des M expériences dont les profils de flux et de température au cours du temps sont tracés sur la Figure 3.4. Les concentrations finales des espèces chimiques mesurées expérimentalement ainsi que les concentrations finales obtenue comme solution du système (3.3) en utilisant les valuers optimales des Tables 3.2, 3.1, 3.3 sont tracées sur la Figure 3.5.

	Cu	In	Ga	Mo
Cu	0	1.03	0.99	1.11
In	1.03	0	1.01	1.00
Ga	1.11	1.01	0	1.00
Mo	1.11	1.00	1.00	0

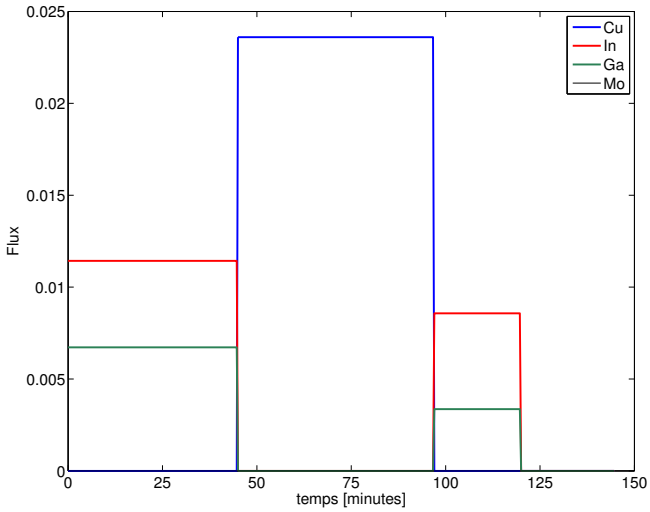
Table 3.1 – Valeurs optimales des énergies d'activation $E^*[10^{-1}\mathrm{eV}]$

	Cu	In	Ga	Mo
Cu	0	8.243	2.837	0.012
In	8.243	0	0.016	0.010
Ga	2.837	0.016	0	0.010
Mo	0.012	0.010	0.010	0

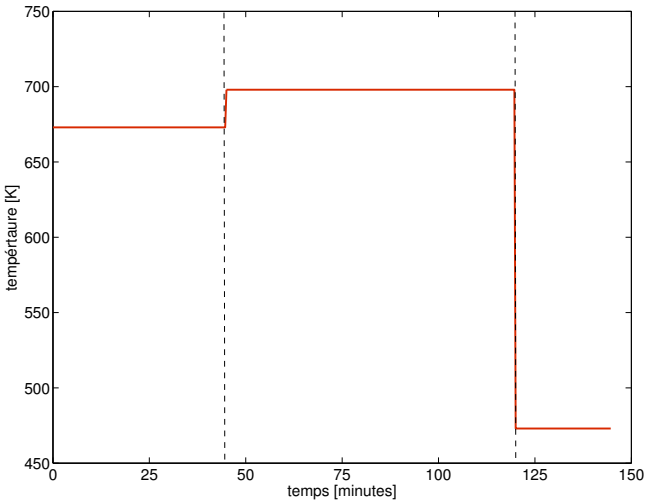
Table 3.2 – Valeurs optimales des préfacteurs $D^*[10^{-2}\mathrm{cm}^2\mathrm{s}^{-1}]$

Cu	In	Ga	Mo
1.20	0.44	0.90	0

Table 3.3 – Valeurs optimales des paramètres $(\lambda_A)_{A\in\mathcal{A}}$.



(a) Flux



(b) Température

Figure 3.4 – Profils de flux et de température associés à l'une des M expériences utilisées pour la calibration.

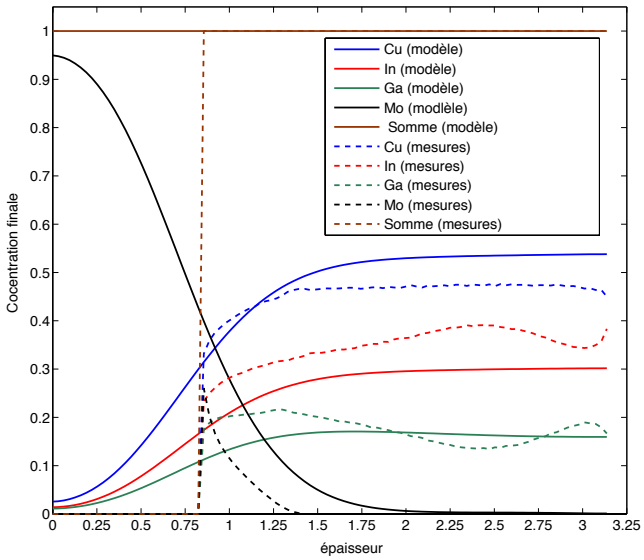


Figure 3.5 – Comparaison entre les concentrations issues des mesures expérimetales de l'une des M expériences et les concentrations finales associées au flux de la Figure 3.4 obtenues comme solution du système (3.3) en utilisant les valeurs optimales des Tables 3.2, 3.1 et 3.3.

Part II

Electronic Structure

CHAPTER 4

ELECTRONIC STRUCTURE OF PERFECT CRYSTALS

Abstract. Many electrical and optical properties of crystalline materials can be explained in terms of their **electronic structures**. The aim of this chapter is to present a concise overview of the standard mathematical tools used in electronic structure calculations.

Contents

A perfect crystal is a solid material composed of an infinite number of nuclei that are periodically arranged in space (see the schematic representation in Figure 4.1). Let $d \in \mathbb{N}^*$ denote the space dimension of the crystal and let $\mathcal{R}$ denote the associated periodic lattice on $\mathbb{R}^d$. In quantum mean-field models, the electronic structure of the considered crystal is characterized by the spectral properties of a periodic Schrödinger operator (called the Hamiltonian of the crystal) of the form

$$A = -\Delta + V$$

acting on $L^2(\mathbb{R}^d;\mathbb{C})$ where V is a real-valued $\mathcal{R}-$periodic potential. In Density Functional Theory (DFT), the potential V is obtained as a solution of some nonlinear self-consistent equation [CLBM06].

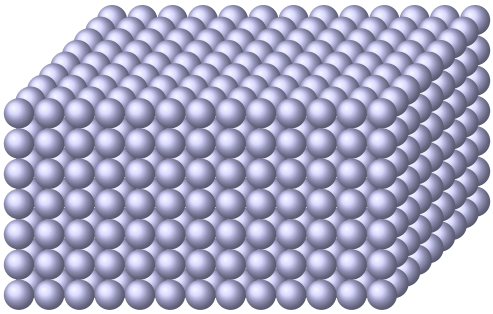


Figure 4.1 – A schematic representation of a perfect crystal in 3D.

Under some assumptions on the potential V, the operator A is self-adjoint and its spectral decomposition can be characterized using the so-called Bloch-Floquet transform presented hereafter. This chapter is organized as follows: Section 4.1 is dedicated to classical results concerning direct integrals of Hilbert spaces, Bloch-Floquet theory and the spectral decomposition of periodic Schrödinger operators. Inverse spectral problems are discussed in Section 4.2 together with a summary of our contributions to these problems. Section 4.3 is devoted to the presentation of classical results on Wannier functions along with a brief summary of our contribution.

4.1 Spectral Properties of Periodic Schrödinger Operators

4.1.1 Direct Integrals of Hilbert Spaces

The notion of direct integrals of Hilbert spaces was first introduced in 1949 by John Von Neumann in his paper on ring operators [Neu49] and has later become a key-tool in the Bloch-Floquet theory. Only the main definitions and results are gathered in this section. The reader is referred to [RS78b, Pan, Wil70] for a deeper analysis. Let $d \in \mathbb{N}^*$ denote the space dimension and consider a Borelian set $\mathcal{M} \subset \mathbb{R}^d$.

Definition 4.1 (Direct integral [RS78b] §III.16)**.** *Let $(\mathcal{H}_q)_{q\in\mathcal{M}}$ be a family of separable Hilbert spaces. The vector space denoted by $\int_{\mathcal{M}}^{\oplus} \mathcal{H}_q \mathrm{dq}$ and defined as follows*

$$\int_{\mathcal{M}}^{\oplus} \mathcal{H}_q \mathrm{d}q := \left\{ \psi = (\psi_q)_{q\in\mathcal{M}} \quad \middle| \quad \forall q \in \mathcal{M}, \quad \psi_q \in \mathcal{H}_q, \quad \int_{\mathcal{M}} \|\psi_q\|^2_{\mathcal{H}_q} \mathrm{dq} < \infty \right\} \tag{4.1}$$

*is called the **direct integral** of the spaces $(\mathcal{H}_q)_{q\in\mathcal{M}}$. A vector of this space is denoted by $\psi = \int_{\mathcal{M}}^{\oplus} \psi_q \mathrm{d}q$.*

Endowed with the inner product

$$\forall \phi, \psi \in \mathcal{H}, \quad \langle \phi | \psi \rangle := \int_{\mathcal{M}} \langle \phi_q | \psi_q \rangle_{\mathcal{H}_q} \mathrm{d}q,$$

the space $\int_{\mathcal{M}}^{\oplus} \mathcal{H}_q \mathrm{d}q$ is a Hilbert space.

Let us recall here some well-known notions about the decomposition of operators on Hilbert spaces that are isomorphic to direct integrals of fiber spaces.

Definition 4.2 (Direct decomposition of operators [RS78b] §III.16)**.** *Let $\mathcal{H}$ be a separable Hilbert space and let $\int_{\mathcal{M}}^{\oplus} \mathcal{H}_q \mathrm{d}q$ be the direct integral of the family $(\mathcal{H}_q)_{q\in\mathcal{M}}$. Consider an isometric isomorphism*[1] *$\mathcal{U} : \mathcal{H} \to \int_{\mathcal{M}}^{\oplus} \mathcal{H}_q \mathrm{d}q$. Then,*

1- *a bounded operator $T \in \mathcal{L}(\mathcal{H})$ is said to be decomposable by $\mathcal{U}$ if for almost all $q \in \mathcal{M}$ there exists a bounded operator $T_q \in \mathcal{L}(\mathcal{H}_q)$ and with $\operatorname*{essup}_{q\in\mathcal{M}} \|T_q\|_{\mathcal{L}(\mathcal{H}_q)} < \infty$ such that, for all $\phi \in \mathcal{H}$, $(\mathcal{U}(T\phi))_q = T_q(\mathcal{U}\phi)_q$.*

2- *a self-adjoint operator T is said to be decomposable by $\mathcal{U}$ if the (bounded) operator $(T - \mathrm{i})^{-1}$ is decomposed by $\mathcal{U}$.*

The notation $T = \mathcal{U}^{-1}\left(\int_{\mathcal{M}}^{\oplus} T_q \mathrm{d}q\right)\mathcal{U}$ is used when $T \in \mathcal{L}(\mathcal{H})$ is decomposable by $\mathcal{U}$ and the operators $(T_q)_{q\in\mathcal{M}}$ are sometimes called the *fibers* of T. Moreover, the following holds

$$\|T\|_{\mathcal{L}(\mathcal{H})} = \operatorname*{ess\,sup}_{q\in\mathcal{M}} \|T_q\|_{\mathcal{L}(\mathcal{H}_q)}.$$

The characterization given in Proposition 4.3 is used to decompose self-adjoint operators, which are of particular interest in our context.

Proposition 4.3 (Decomposition of self-adjoint operators [RS78b] §III.16)**.** *A self-adjoint operator T acting on the separable Hilbert space $\mathcal{H}$ with domain $\mathcal{D}(T)$ is decomposed by $\mathcal{U}$ if and only if for almost all $q \in \mathcal{M}$, there exists a self-adjoint operator T_q acting on $\mathcal{H}_q$ with domain $\mathcal{D}(T_q)$ such that:*

1. *the function $q \in \mathcal{M} \mapsto \|(T_q + i)^{-1}\|_{\mathcal{L}(\mathcal{H}_q)}$ is measurable;*

2. *$\mathcal{D}(T) = \left\{\phi \in \mathcal{H} \;\middle|\; (\mathcal{U}\phi)_q \in \mathcal{D}(T_q) \text{ a.e} \quad \text{and} \quad \int_{\mathcal{M}} \|T_q(\mathcal{U}\phi)_q\|^2_{\mathcal{H}_q} \mathrm{d}q < \infty\right\}$;*

3. *for all $\phi \in \mathcal{D}(T)$, $T\phi = \mathcal{U}^{-1}\left(\int_{\mathcal{M}} T_q(\mathcal{U}\phi)_q \mathrm{d}q\right)$.*

We conclude this section by a major result on the characterization of the spectrum of a self-adjoint decomposable operator using the spectra of its fibers. This characterization used in conjunction with the Bloch-Floquet decomposition discussed in Section 5.2.1, is an essential tool for the characterization of the spectrum of periodic Schrödinger operators.

Proposition 4.4 (Spectrum of decomposed self-adjoint operators [RS78b] §III.16)**.** *Let $T = \mathcal{U}^{-1}\left(\int_{\mathcal{M}}^{\oplus} T_q \mathrm{d}q\right)\mathcal{U}$ be a self-adjoint operator on $\mathcal{H}$ decomposed by $\mathcal{U}$. We denote by $\sigma(T)$ the spectrum of T and by $\sigma_p(T) \subset \sigma(T)$ its point spectrum. Then, the following statements hold*

$$\begin{aligned} \lambda \in \sigma(T) \quad &\Leftrightarrow \quad |\{q \in \mathcal{M},\ (\lambda - \varepsilon, \lambda + \varepsilon) \cap \sigma(T_q) \neq \emptyset\}| > 0, \quad \forall \varepsilon > 0, \\ \lambda \in \sigma_p(T) \quad &\Leftrightarrow \quad |\{q \in \mathcal{M},\ \lambda \in \sigma_p(T_q)\}| > 0. \end{aligned}$$

[1] We use the same terminology as in [RS78a]. See Chapter III, page 71.

4.1.2 Bloch-Floquet Transform

The Bloch-Floquet transform was first introduced by the mathematician Gaston Floquet [Flo83] for the study of differential equations with periodic coefficients and then by the physicist Felix Bloch [Blo29] in the context of electronic structure. This transform is strongly linked to the symmetry properties of the crystalline materials. Let us first recall some basic notions of crystallography. Let $(\mathbf{a}_1, \cdots, \mathbf{a}_d)$ be a basis of $\mathbb{R}^d$. The **Bravais lattice** $\mathcal{R}$ associated to the basis $(\mathbf{a}_1, \cdots, \mathbf{a}_d)$ is defined by

$$\mathcal{R} := \left\{ \mathbf{R} \in \mathbb{R}^d, \quad \mathbf{R} = \sum_{j=1}^{d} n_j \mathbf{a}_j \mid \quad n_j \in \mathbb{Z},\ 1 \leq j \leq d \right\}. \tag{4.2}$$

An admissible unit cell of $\mathcal{R}$ is given by

$$\Gamma := \left\{ \sum_{j=1}^{d} \alpha_j \mathbf{a}_j, \quad -1/2 \leq \alpha_j < 1/2, \quad 1 \leq j \leq d \right\}. \tag{4.3}$$

Let $(\mathbf{a}_1^*, \cdots, \mathbf{a}_d^*)$ be the dual basis associated to $(\mathbf{a}_1, \cdots, \mathbf{a}_d)$. These vectors are uniquely defined through the relations

$$\mathbf{a}_j^* \cdot \mathbf{a}_i = 2\pi\delta_{ij}, \quad \forall 1 \leq i, j \leq d.$$

The **dual lattice** $\mathcal{R}^*$ of $\mathcal{R}$ is then defined by

$$\mathcal{R}^* := \left\{ \mathbf{K} \in \mathbb{R}^d, \quad \mathbf{K} = \sum_{j=1}^{d} m_j \mathbf{a}_j^* \mid \quad m_j \in \mathbb{Z},\ 1 \leq j \leq d \right\}. \tag{4.4}$$

The **first Brillouin zone** Γ^* of the lattice $\mathcal{R}$ is defined as the Wigner-Seitz cell of the dual lattice $\mathcal{R}^*$, that is the set of points of $\mathbb{R}^d$ that are closer to the origin than to any other point of $\mathcal{R}^*$. More precisely,

$$\Gamma^* := \left\{ q \in \mathbb{R}^d, \quad |q| \leq |q - \mathbf{K}|, \quad \forall \mathbf{K} \in \mathcal{R}^* \right\}. \tag{4.5}$$

In the remaining sections, unless there is an ambiguity, we use the notation L^2 for the Hilbert space $L^2(\mathbb{R}^d; \mathbb{C})$. The Bloch-Floquet decomposition relies on the theory of direct integrals. There are two equivalent versions of the Bloch-Floquet decomposition (Theorem 4.5 and Theorem 4.6) that show how to decompose L^2 into a direct integral of two different families of fiber spaces. Before we state the two versions of the decomposition, let us introduce the definition of the involved spaces. Let $s \in \mathbb{N}^*$,

$$\begin{aligned} L^2_{\text{per}} &:= \left\{ u \in L^2_{\text{loc}} \mid u \text{ is } \mathcal{R}\text{-periodic} \right\}, \\ H^s_{\text{per}} &:= \left\{ u \in H^s_{\text{loc}} \mid u \text{ is } \mathcal{R}\text{-periodic} \right\}. \end{aligned} \tag{4.6}$$

It is classically known that the spaces L^2_{per} and H^s_{per} endowed respectively with the inner products

$$\begin{aligned} &\forall v, w \in L^2_{\text{per}}, \qquad \langle v, w \rangle_{L^2_{\text{per}}} = \textstyle\int_\Gamma \overline{v} w, \\ &\forall v, w \in H^s_{\text{per}}, \qquad \langle v, w \rangle_{H^s_{\text{per}}} = \displaystyle\sum_{\alpha \in \mathbb{N}^d, |\alpha| \leq s} \langle \partial^\alpha v, \partial^\alpha w \rangle_{L^2_{\text{per}}} \end{aligned} \tag{4.7}$$

are Hilbert spaces. Moreover, we define for every $q \in \Gamma^*$ the following spaces

$$\begin{array}{rcl} L^2_q & := \left\{ \psi \in L^2_{\mathrm{loc}} \quad | \quad \mathbb{R}^d \ni x \mapsto \psi(x) e^{-\mathrm{i}q\cdot x} \in L^2_{\mathrm{per}} \right\}, \\ H^s_q & := \left\{ \psi \in H^s_{\mathrm{loc}} \quad | \quad \mathbb{R}^d \ni x \mapsto \psi(x) e^{-\mathrm{i}q\cdot x} \in H^s_{\mathrm{per}} \right\}, \end{array} \tag{4.8}$$

which, respectively endowed with the following inner products,

$$\begin{aligned} \forall \psi, \phi \in L^2_q, \quad \langle \psi, \phi \rangle_{L^2_q} &:= \int_\Gamma \overline{\psi} \phi, \\ \forall \psi, \phi \in H^s_q, \quad \langle \psi, \phi \rangle_{H^s_q} &:= \sum_{\alpha \in \mathbb{N}^d, |\alpha| \leq s} \langle \partial^\alpha \psi, \partial^\alpha \phi \rangle_{L^2_q} \end{aligned}$$

are also Hilbert spaces. We are now in position to give the two versions of the Bloch-Floquet decomposition.

Theorem 4.5 (First Bloch-Floquet decomposition, [RS78b])**.** *We consider the direct integral space $\mathcal{H}$ given by*

$$\mathcal{H} = \frac{1}{|\Gamma^*|} \int_{\Gamma^*}^{\oplus} L^2_q dq.$$

Then,

1. *the linear map $\widetilde{\mathcal{B}}$ from $C^\infty_c(\mathbb{R}^d; \mathbb{C})$ to $\mathcal{H}$ defined for every $\phi \in \mathcal{C}^\infty_c(\mathbb{R}^d; \mathbb{C})$ by :*

$$\forall (q, x) \in \Gamma^* \times \Gamma, \quad (\widetilde{\mathcal{B}}\phi)_q(x) := \sum_{\mathbf{R} \in \mathcal{R}} \phi(x + \mathbf{R}) e^{-\mathrm{i}q\cdot\mathbf{R}} \tag{4.9}$$

   can be continuously extended to a unique isometric isomorphism from L^2 onto $\mathcal{H}$.

2. *the inverse of $\widetilde{\mathcal{B}}$ is given for all $\psi \in L^2$ by:*

$$(\widetilde{\mathcal{B}}^{-1}\psi)(x) = \frac{1}{|\Gamma^*|} \int_{\Gamma^*} \psi_q(x) dq, \quad \text{for a.e } x \in \Gamma. \tag{4.10}$$

Theorem 4.6 (Second Bloch-Floquet decomposition, [RS78b])**.** *We consider the direct integral $\mathcal{H}$ given by*

$$\mathcal{H} = \frac{1}{|\Gamma^*|} \int_{\Gamma^*}^{\oplus} L^2_{\mathrm{per}} dq.$$

Then,

1. *the linear map $\mathcal{B}$ from $C^\infty_c(\mathbb{R}^d; \mathbb{C})$ to $\mathcal{H}$ defined for every $\phi \in \mathcal{C}^\infty_c(\mathbb{R}^d; \mathbb{C})$ by :*

$$\forall (q, x) \in \Gamma^* \times \Gamma, \quad (\mathcal{B}\phi)_q(x) = (\widetilde{\mathcal{B}}\phi)_q(x) e^{-\mathrm{i}q\cdot x} = \sum_{\mathbf{R} \in \mathcal{R}} \phi(x + \mathbf{R}) e^{-\mathrm{i}q\cdot(\mathbf{R}+x)}$$

   can be continuously extended to a unique isometric isomorphism from L^2 onto $\mathcal{H}$.

2. *the inverse of $\mathcal{B}$ is given for all $\psi \in L^2$ by :*

$$(\mathcal{B}^{-1}\psi)(x) = \frac{1}{|\Gamma^*|} \int_{\Gamma^*} \psi_q(x) e^{\mathrm{i}q\cdot x} dq, \quad \text{for a.e } x \in \Gamma.$$

A proof of Theorem 4.5 is presented in [Pan]. Theorem 4.6 is a direct corollary of Theorem 4.5. However, the second decomposition is more advantageous in the context of periodic Schrödinger operators. Indeed it leads to a family of fiber operators with the same domain unlike the first decomposition where the domains depend on q.

Let us finally underline the close link between the symmetries of the lattice $\mathcal{R}$ and the Bloch-Floquet decomposition. To this aim, let us define for each lattice vector $\mathbf{R} \in \mathcal{R}$, a translation operator $\tau_{\mathbf{R}}$ defined by

$$\tau_{\mathbf{R}} : \left\{ \begin{array}{ccc} L^2 & \to & L^2 \\ \phi & \mapsto & \phi(\cdot + \mathbf{R}) \end{array} \right. . \tag{4.11}$$

Proposition 4.7 (Link between the symmetries of the lattice and Bloch-Floquet transform, [RS78b])**.** *Any linear bounded operator $T \in \mathcal{L}(L^2)$ which commutes with $\tau_{\mathbf{R}}$ for every $\mathbf{R} \in \mathcal{R}$ is decomposed by $\mathcal{B}$ and $\tilde{\mathcal{B}}$. Similarly, any self-adjoint operator acting on L^2 which commutes with $\tau_{\mathbf{R}}$ for every $\mathbf{R} \in \mathcal{R}$ is also decomposed by $\mathcal{B}$ and $\tilde{\mathcal{B}}$.*

4.1.3 Spectral Decomposition of Periodic Schrödinger Operators

Let V be a real-valued periodic potential belonging to the space L^p_{per} with $p = 2$ if $d \leq 3$ and $p > d/2$ if $d \geq 4$. Then, the periodic Schrödinger operator $A = -\Delta + V$ on L^2 is selfadjoint with domain $H^2(\mathbb{R}^d, \mathbb{C})$ and is bounded from below. Furthermore, the following properties are satisfied (proofs can be found in [RS78b, Pan]):

1. The operator A is decomposed by $\mathcal{B}$ and by $\widetilde{\mathcal{B}}$.

2. Furthermore,

$$A = \mathcal{B}^{-1} \left(\int_{\Gamma^*}^{\oplus} A_q \frac{1}{|\Gamma^*|} dq \right) \mathcal{B} \quad \text{and} \quad A = \widetilde{\mathcal{B}}^{-1} \left(\int_{\Gamma^*}^{\oplus} \widetilde{A}_q \frac{1}{|\Gamma^*|} dq \right) \widetilde{\mathcal{B}}$$

   where for every $q \in \Gamma^*$,

   - the operator $\widetilde{A}_q$ acting on L^2_q with domain $\mathcal{D}(\widetilde{A}_q) = H^2_q$ is defined by

$$\forall \psi_q \in \mathcal{D}(\widetilde{A}_q), \quad \widetilde{A}_q \psi_q = -\Delta \psi_q + V \psi_q;$$

   - the operator A_q acting on L^2_{per} with domain $\mathcal{D}(A_q) = H^2_{\mathrm{per}}$ is defined by

$$\forall u_q \in \mathcal{D}(A_q), \quad A_q u_q = |-\mathrm{i}\nabla + q|^2 u_q + V u_q.$$

3. For each $q \in \Gamma^*$, both operators A_q and $\widetilde{A}_q$ are bounded from below, self-adjoint, have compact resolvent and are unitary equivalent. Therefore, they share the same spectrum. Thus, for each $q \in \Gamma^*$, there exists

   - a non-decreasing sequence $(\varepsilon^V_{q,n})_{n \in \mathbb{N}^*}$ going to $+\infty$;
   - an orthonormal basis $(\psi^V_{q,n})_{n \in \mathbb{N}^*}$ of L^2_q and an orthonormal basis $(u^V_{q,n})_{n \in \mathbb{N}^*}$ of L^2_{per},

   such that for all $n \in \mathbb{N}^*$,

$$\widetilde{A}_q \psi_{q,n} = \varepsilon^V_{q,n} \psi_{q,n}, \quad A_q u_{q,n} = \varepsilon^V_{q,n} u_{q,n}, \quad u_{q,n}(x) = \psi_{q,n}(x) e^{-\mathrm{i}q \cdot x}.$$

4. For all $n \in \mathbb{N}^*$, the function $q \mapsto \varepsilon_{q,n}^V$ can be extended to a continuous $\mathcal{R}^*$-periodic function on $\mathbb{R}^d$, so that

$$\sigma(A) = \bigcup_{n=1}^{\infty} [\min_{q\in\Gamma^*} \varepsilon_{q,n}^V, \max_{q\in\Gamma^*} \varepsilon_{q,n}^V].$$

5. The spectrum of A is absolutely continuous real spectrum

$$\sigma(A) = \sigma_{ac}(A), \quad \text{and} \quad \sigma_{sg}(A) = \sigma_p(A) = \emptyset.$$

The function $\mathbb{R}^d \ni q \mapsto \varepsilon_{q,n}^V$ is called the nth **energy band** associated to the potential V. The term **dispersion relation** is sometimes used to refer to the complete set of energy bands. Let us mention that the energy bands (in arbitrary dimension d) satisfy the following symmetry properties for every $n \in \mathbb{N}^*$, every $q \in \Gamma^*$ and every $\mathbf{K} \in \mathcal{R}^*$

$$\varepsilon_{q,n}^V = \varepsilon_{-q,n}^V \quad \text{and} \quad \varepsilon_{q+\mathbf{K},n}^V = \varepsilon_{q,n}^V. \tag{4.12}$$

For the sake of completeness, we present in the appendix one of the most popular methods for the numerical approximation of the energy bands, namely the plane-wave discretization method.

4.2 Inverse Spectral Problems

Inverse spectral problems consist in recovering operators from their spectral characteristics. Such problems often appear in mathematics, physics and materials science [FY01].

4.2.1 Classical Inverse Problems

One of the most studied situations is the inverse Sturm-Liouville problem. The aim in this problem is to *recover* the potential function $V \in L^2(0,1)$ appearing in the one-dimensional elliptic eigenvalue problem

$$-u''(x) + V(x)u(x) = \varepsilon u(x), \quad \text{for} \quad 0 \le x \le 1,$$

with the impedance boundary conditions

$$u'(0) - lu(0) = 0, \quad \text{and} \quad u'(1) + Lu(1) = 0, \tag{4.13}$$

where l and L are given real numbers, from the knowledge of some spectral data. It is known that the complete spectrum $(\varepsilon_n^V)_{n\in\mathbb{N}^*}$ is not sufficient to reconstruct (uniquely) the potential V; additional data are needed [PT87]. Let us quote from the literature a few versions where existence and uniqueness of a solution to the inverse Sturm-Liouville problem were proved and refer the reader to [W.R92, FY01, PT87] for a more complete introduction to the subject. There are two cases where the (only) knowledge of $(\varepsilon_n^V)_{n\in\mathbb{N}^*}$ is sifficient to the unique reconstruction of V: i) when V is assumed to have the symmetry $V(x) = V(1-x)$ for $0 \le x \le 1$ and ii) when some a priori information on V is provided (for instance, when V is known on half of the interval $[0,1]$) [PT87, G.B46, H.H76]. Besides, the unique recovery is also possible if, in addition to

$(\varepsilon_n^V)_{n\in\mathbb{N}^*}$, a second complete spectrum $(\mu_n^V)_{n\in\mathbb{N}^*}$ is provided corresponding to different boundary conditons (the value of L is changed to $L' \neq L$) [G.B46]. Several numerical techniques were suggested to solve these questions [W.R92]. In most of the methods, the Sturm-Liouville operator is discretized using finite differences, fintie elements or Numerov's scheme [And04]. The inverse (continuous) problem is then transformed to an inverse (discrete) eigenvalue problem. Asymptotic correction terms are usually introduced to reduce the discretization error [GCH13].

The periodic framework also attracted mathematician's attention for decades. In this case, the aim is to recover the real-valued $\mathcal{R}$-periodic potential V appearing in the periodic Schrödinger operator $A = -\Delta + V$ from the knowledge of the dispersion relation (the set of energy bands). Several partial answers were proposed. One of the first contributions in the one-dimensional case is due to Borg [G.B46] where necessary and sufficient conditions were given on the dispersion relation for the potential V to be constant.

Let us introduce the complex *Bloch variety* $B(V)$ containing all points that can possibly be reached by analytic continuation of any energy band. In particular, the graph of any energy band $q \in \mathbb{R}^d \mapsto \varepsilon_{q,n}^V$ is a subset of $B(V)$. The generalization of Borg's result to arbitrary dimension gives raise to the following conjecture:

Conjecture 4.8 (Borg's conjecture, [AS78, Kuc16])**.** *The potential V is constant if and only if there exists an entire function $f : \mathbb{C}^d \to \mathbb{C}$ such that the Bloch variety $B(V)$ is the union of the graph of f and its translates under $\mathcal{R}^*$.*

One can think of the dispersion relation associated to the $2\pi\mathbb{Z}$-periodic potential $V \equiv 0$ which is given by $\mathbb{Z}-$translations of the graph of the function $\mathbb{R} \ni q \mapsto q^2$ (see Figure 4.2).

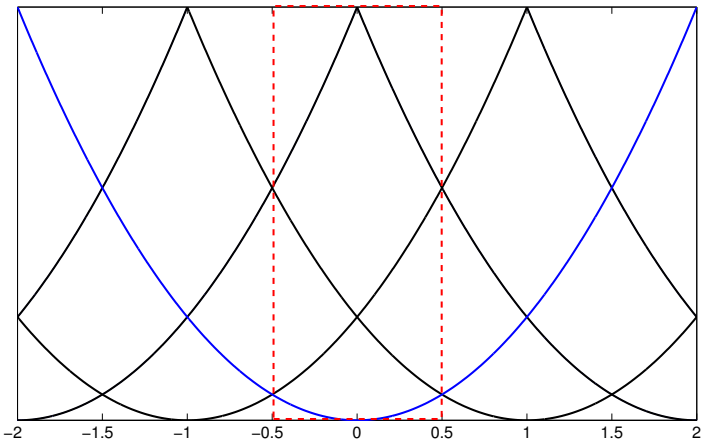


Figure 4.2 – Dispersion relation of the one-dimensional periodic Schrödinger operator $-d/dx^2 + V$ with the lattice $\mathcal{R} = 2\pi\mathbb{Z}$ and where $V \equiv 0$.

This conjecture was proved for $d = 2$ by Knörrer and Trubowitz [KT90] and to the best of my knowledge, the proof in higher dimension is still an open question.

Another point of view consists in the characterization of the **isospectral sets** which are the sets of potentials associated to the same energy bands. Two real-valued potentials V and W are said to be **Floquet isospectral** when

$$\sigma(A_q^V) = \sigma(A_q^W), \quad \forall q \in \Gamma^*,$$

where $\sigma(A_q^V)$ denotes the discrete spectrum of the Bloch operator A_q^V at $q \in \Gamma^*$ (see Section 4.1.3). The potentials V and W are said to be **isospectral** if $\sigma(A_0^V) = \sigma(A_0^W)$. Some particular cases can be immediately determined. It is clear for instance that any potential V is Floquet isospectral to all its translated versions $V(\cdot - \tau), \tau \in \mathbb{R}^d$. The question of interest in the general case is : what is the set of potentials that are isospectral to a given real-valued peridodic potential V ? Numerous results can be found in [PT87] for the one-dimensional setting. The multi-dimensional case was investigated in [Esk89, ERT84a, ERT84b]. The authors showed in particular that if two one-dimensional smooth periodic potentials $\mathbb{R} \ni x \mapsto \tilde{V}(x)$ and $\mathbb{R} \ni x \mapsto \tilde{W}(x)$ are isospectral and if $\delta \in \mathbb{R}^d$ is a vector such that $\delta \cdot \mathbf{R} \in \mathbb{Z}$ for every $\mathbf{R} \in \mathcal{R}$ then, the d-dimensional periodic potentials $\mathbb{R}^d \ni x \mapsto V(x) := \tilde{V}(x \cdot \mathbf{R})$ and $\mathbb{R}^d \ni x \mapsto W(x) := \tilde{W}(x \cdot \mathbf{R})$ are also isospectral. Another interesting result that we report from the review [Kuc16] is the following:

Theorem 4.9 (Floquet Isospectrality)**.** *Assume that the lattice $\mathcal{R}$ satisfies the following property for every $\mathbf{R}_1, \mathbf{R}_2 \in \mathcal{R}$:*

$$(|\mathbf{R}_1| = |\mathbf{R}_2|) \Rightarrow (\mathbf{R}_1 = \pm \mathbf{R}_2)\,.$$

Consider two $\mathcal{R}-$periodic potentials $V, W \in \mathcal{C}^\infty_{\mathrm{per}}$. If there exists $q_0 \in \Gamma^*$ such that $\cos(2\pi q_0 \cdot \mathbf{R}) \neq 0$ for all $\mathbf{R} \in \mathcal{R}$ and $\sigma(A_{q_0}^V) = \sigma(A_{q_0}^W)$, then the potentials V and W are Floquet isospectral.

This result points out that, when the potential is smooth, the spectrum of one Bloch operator at a particular q-point $q_0 \in \Gamma^*$ holds the complete information on the whole dispersion relation. Furthermore, the following (strong) property is conjectured to hold, which allows one to focus only on one open branch of a single energy band :

Conjecture 4.10 (Conjecture 5.17 [Kuc16])**.** *Let V be a real valued $\mathcal{R}$-periodic potential belonging to L^p_{per} with $p = 2$ if $d \leq 3$ and $p > d/2$ if $d \geq 4$. Then, for any level $n \in \mathbb{N}^*$ and any open set $\Omega \subset \mathbb{R}^d$, the branch $\Omega \ni q \mapsto \varepsilon_{q,n}$ of the n^{th} energy band determines uniquely the full energy band dispersion.*

This conjecture is proven in the one and two dimensional cases in [KT90]. To the best of my knowledge, the proof in higher dimension still remains open.

Let us finally mention the works by Veliev gathered in [Vel15] concerning smooth potentials. In the same spirit as [ERT84a, ERT84b], Veliev introduced a family of quantities (that he called spectral invariants) which can be constructed using the given energy bands. Then, explicit expressions relating these invariants to the Fourier coefficients of the unknown potential are provided. The class of potentials that can be recovered by Veliev's method is shown to be dense in H^s_{per} with $s \geq 6(3^d(d+1)^2) + d$. This density argument allows to conclude that: a smooth potential $V \in H^s_{\mathrm{per}}$ can be uniquely determined (up to translations $x \mapsto x + \tau$ with $\tau \in \mathbb{R}^d$ and inversions $x \mapsto -x$) from the knowledge of the whole set of its associated energy bands $\Gamma^* \ni q \to \varepsilon^V_{q,n}$ for $n \in \mathbb{N}^*$.

4.2.2 Contributions of the Thesis (Inverse Hill's problem)

The approaches presented above use the knowledge of the asymptotic behavior of the high-energy bands for the reconstruction of the potential, and therefore are unsuitable for practical purpose. Indeed, in practice only the low-energy bands of the crystal (more precisely, the conduction and the valence bands for physicists) are of interest. For applications, it is therefore interesting to know how to construct a potential such that only its lowest energy bands are close to some given target functions without additional information on the high energy bands. We adopt in this thesis a viewpoint which is different from the classical inverse problems presented above : we recast the problem as an *optimization* problem. More precisely, the following question is considered : *given a family of M functions $b_1, \cdots, b_M : \Gamma^* \to \mathbb{R}^d$, does there exist a real-valued $\mathcal{R}$-periodic potential V such that the associated first energy bands $\varepsilon_{q,1}^V, \cdots, \varepsilon_{q,M}^V$, are as close as possible (in some sense) to the target functions $b_1, \cdots, b_M$?*. In Chapter 5, we report the results of [BEG17] obtained with Virginie Ehrlacher (*Université Paris Est, CERMICS (ENPC), Inria, Paris, France*) and David Gontier (*Université Paris-Dauphine, CEREMADE, France*), where a theoretical answer to the above question is given for one target function $M = 1$ in the one-dimensional space and where an algorithm is proposed to answer the question numerically for an arbitrary number of target functions $M \in N^*$.

More precisely, in the case $d = 1$, we consider the space of non-negative $2\pi-$periodic regular Borel measures on $\mathbb{R}$ that we denote by $\mathcal{M}_{\text{per}}^+$. It holds in particular that $\mathcal{M}_{\text{per}}^+$ is compactly embedded in the set H_{per}^{-1} and that to each $\nu \in \mathcal{M}_{\text{per}}^+$ corresponds a unique real-valued potential $V_\nu \in H_{\text{per}}^{-1}$ defined by duality. Then, for a fixed constant $B \in \mathbb{R}$, we consider the set of B-bounded from below potentials

$$\mathcal{V}_B := \left\{ V \in H_{\text{per}}^{-1} | V \text{ is real-valued}, \quad \exists \nu \in \mathcal{M}_{\text{per}}^+, \quad V = V_\nu - B \right\}.$$

The following partial result is proved in Section 5.3.2 of Chapter 5:

Proposition 4.11. *Let $B \in \mathbb{R}$ and let $(V_n)_{n\in\mathbb{N}^*} \subset \mathcal{V}_B$. For all $n \in \mathbb{N}^*$, let $\nu_n \in \mathcal{M}_{\text{per}}^+$ such that $V_n := V_{\nu_n} - B$ and such that $\nu_n(\Gamma) \underset{n\to+\infty}{\longrightarrow} +\infty$. Assume that the sequence $\left(\varepsilon_{q=0,1}^{V_n}\right)_{n\in\mathbb{N}^*}$ is bounded. Then, up to a (non relabeled) subsequence, there exists $\varepsilon \geq \frac{1}{4} - B$ such that*

$$\max_{q\in[0,1/2]} \left| \varepsilon_{q,1}^{V_n} - \varepsilon \right| \underset{n\to\infty}{\longrightarrow} 0. \tag{4.14}$$

Conversely, for all $\varepsilon \geq \frac{1}{4} - B$, there is a sequence $(V_n)_{n\in\mathbb{N}^*} \subset \mathcal{V}_B$ such that (4.14) holds.

Roughly speaking, this result indicates that the first energy band associated to a unbounded sequence of potentials in $\mathcal{V}_B$ becomes flat.

Introduce now the set

$$\mathcal{T} := \left\{ b \in \mathcal{C}^0(\Gamma^*), \quad b \text{ is even and } b \text{ is increasing on } [0, 1/2] \right\} \tag{4.15}$$

of admissible target functions. Note that the first energy band of any real-valued periodic potential $V \in \mathcal{V}_B$ belongs to the set $\mathcal{T}$. For each $b \in \mathcal{T}$, we consider the functional $\mathcal{J}_b : \mathcal{V}_B \to \mathbb{R}$ defined by

$$\forall V \in \mathcal{V}_B, \quad \mathcal{J}_b(V) := \int_0^{1/2} |b(q) - \varepsilon_{q,1}^V|^2 \, dq. \tag{4.16}$$

The quantity $\mathcal{J}_b(V)$ measures the error (in the L^2-norm) between the first energy band $\varepsilon^V_{q,1}$ associated to the potential V and the target function $b \in \mathcal{T}$. Note that by virtue of the symmetry property $\varepsilon^V_{q,1} = \varepsilon^V_{-q,1}$, it is possible to consider the problem in half of the Brillouin zone only.

Our main theoretical result is the following

Theorem 4.12. *Let $b \in \mathcal{T}$, and denote by $b^* := \fint_{\Gamma^*} b(q)\,dq \in \mathbb{R}$. Then, for all $B > 1/4 - b^*$, there exists a solution $V_{b,B} \in \mathcal{V}_B$ to the minimization problem*

$$V_{b,B} \in \operatorname{argmin}_{V \in \mathcal{V}_B} \mathcal{J}_b(V). \tag{4.17}$$

The proof of this theorem is based on the Proposition 4.11 and is reported in Section 5.3.3 of Chapter 5.

From a numerical point of view, the problem is addressed with several target functions $b_1, \cdots, b_M$ defined on $[0, 1/2]$ that are assumed to be continuous and such that b_m is increasing when m is odd and decreasing when m is even. A uniform grid Γ^*_Q of size Q is considered on the interval $[0, 1/2]$ and the Bloch eigenvalues $(\varepsilon^{V,s}_{q,m})_{1 \le m \le M}$ are computed for every $q \in \Gamma^*_Q$ by the plane-wave method (see the appendix) in a Fourier space of dimension $2s + 1$ for some cutoff value $s \in \mathbb{N}^*$. Moreover, for $p \in \mathbb{N}^*$, a set Y_p of real-valued periodic potentials having $2p + 1$ Fourier coefficients in their Fourier series is introduced to approximate the search space. Eventually, the following discrete minimization problem is considered

$$V^{s,p} := \operatorname{argmin}_{V \in Y_p} \left(\frac{1}{Q} \sum_{q \in \Gamma^*_Q} \sum_{m=1}^{M} |b_m(q) - \varepsilon^{V,s}_{q,m}|^2 \right).$$

A standard gradient iterative procedure is first proposed to solve the problem. In this (naive) method, the numerical parameters p and s are chosen a priori at the beginning of the algorithm and kept fixed throughout the procedure. Although the method gives satisfactory numerical optimizers, it presents a major limitation: the computational time grows quickly with the values of the parameters p and s. In order to improve the efficiency of the numerical optimization procedure, an adaptive search algorithm is proposed. The idea of the adaptive approach is to start the optimization with small values of p and s and increase them (if necessary) during the optimization process. The adaptive search algorithm relies on the use of

i) an a posteriori error estimator for the approximation of the eigenvalues, which rules the choice of the discretization parameter s,

ii) a heuristic criterion used to determine the choice of the parameter p.

The numerical tests revealed that the adaptive approach is usually faster (in terms of the computational time) than the naive one even if it requires more iterations to converge. A detailed description of the algorithms along with several numerical tests are presented in Section 5.4 of Chapter 5.

Let us finally mention that the a posteriori error estimator used in the adaptive algorithm is based on a work done in collaboration with Damiano Lombardi (INIRA Paris, France) aiming to develop a certified and sharp a posteriori estimator for (more general) Hermitian eigenvalue problems. A preliminary version of this work can be found in [BL17].

4.3 Wannier functions

Wannier functions (WF) were introduced in 1937 by Gregory Wannier [Wan37] and have become a powerful tool in solid state physics. They are **localized**-in-space functions constructed from the eigenfunctions of the Bloch operators $(A_q)_{q\in\Gamma^*}$. Thus, Wannier functions can be seen as the solid-state equivalent of localized molecular orbitals. They provide intuition on the chemical bonding and play an essential role for several approximations such as the tight binding Hamiltonians [MSV03].

4.3.1 Theoretical Aspects

Let $\{\mathbb{R}^d \ni q \mapsto \varepsilon_{q,n}\}_{n\geq 1}$ be the energy bands associated to the periodic Shcrödinger operator $A = -\Delta + V$ on L^2, where V is an $\mathcal{R}$-periodic potential belonging to the space $V \in L^p_{\mathrm{per}}$ with $p = 2$ if $d \leq 3$ and $p > d/2$ if $d \geq 4$.

Definition 4.13 (Isolated bands)**.** *The periodic Schrödinger operator A is said to have a set of $N \geq 1$ bands isolated from the rest of the spectrum if there exist two continuous $\mathbb{R}$-valued $\mathcal{R}$-periodic functions $q \mapsto \mu_-(q)$ and $q \mapsto \mu_+(q)$ such that $\mu_-(q) < \mu_+(q)$, $\mu_\pm(q) \notin \sigma(A_q)$ and $\mathrm{tr}\big(\mathbb{1}_{[\mu_-(q),\mu_+(q)]}(A_q)\big) = N$ for all $q \in \mathbb{R}^d$.*

To lighten the notation, we denote by $\varepsilon_{q,1}, \cdots, \varepsilon_{q,N}$ the eigenvalues of A_q lying in the *energy window* $[\mu_-(q), \mu_+(q)]$ for each $q \in \mathbb{R}^d$. The reader should keep in mind that these bands do not necessarily correspond to the lowest N bands of the operator.

We recall from the Bloch Floquet theory (see Section 5.2.1) that, for each q-point $q \in \mathbb{R}^d$, there exists an orthonormal basis $(u_{q,n})_{n\in\mathbb{N}^*}$ of L^2_{per} such that

$$A_q u_{q,n} := |-\mathrm{i}\nabla + q|^2 u_{q,n} + V u_{q,n} = \varepsilon_{q,n} u_{q,n}, \quad \forall n \in \mathbb{N}^*.$$

Assume that the periodic Schrödinger operator A has an isolated set of $N \geq 1$ bands and consider a transformation U, which we refer to in the sequel as a *gauge transformation*, that associates a unitary matrix U^q to each q-point:

$$U : \left\{ \begin{array}{ccc} \mathbb{R}^d & \to & \mathcal{U}_N \\ q & \mapsto & U^q \end{array} \right. \tag{4.18}$$

where $\mathcal{U}_N \subset \mathbb{C}^{N\times N}$ denotes the space of complex-valued unitary matrices. For each $q \in \mathbb{R}^d$ and each $1 \leq n \leq N$, we introduce the following *generalized Bloch wave*

$$\widetilde{\psi}_{q,n} = \sum_{m=1}^{N} U^q_{mn} u_{q,n}. \tag{4.19}$$

We denote by $\mathcal{G} = \{U : \mathbb{R}^d \to \mathcal{U}_N\}$ the set of gauge transformations.

Definition 4.14 (Composite Wannier Functions)**.** *Composite Wannier functions $\{w_{\mathbf{R},n}\}_{\mathbf{R}\in\mathcal{R},1\leq n\leq N}$ associated to an isolated set of N bands are obtained by the following formula*

$$\forall x \in \mathbb{R}^d, \quad w_{0,n}(x) = \frac{1}{|\Gamma^*|} \int_{\Gamma^*} \widetilde{\psi}_{q,n}(x) \mathrm{e}^{\mathrm{i}q\cdot x} dq \tag{4.20}$$

and

$$\forall (x, \mathbf{R}) \in \mathbb{R}^d \times \mathcal{R}, \quad w_{\mathbf{R},n}(x) = w_{0,n}(x - \mathbf{R}).$$

The Wannier functions $\{w_{\mathbf{R},n}\}_{\mathbf{R}\in\mathcal{R},1\leq n\leq N}$ form a complete orthonormal basis of the subspace of L^2 associated to the isolated set of bands. Note that there is a *gauge freedom* in Definition 4.14 meaning that the Wannier functions are not uniquely determined. This originates form the arbitrary choice of the unitary transformation U in the definition of the generalized Bloch waves (4.19).

Let us now briefly discuss a few questions related to the construction of Wannier functions and give some relevant references.

The first question concerns the possibility of building Wannier functions that are exponentially decaying. As discussed in [Kun, MMY+12, MSV03], the localization of the Wannier functions is determined by the periodicity and the regularity of the generalized Bloch waves $\widetilde{\psi}_{q,n}$ as functions of $q \in \mathbb{R}^d$. It turns out, in dimension $d \geq 2$, that the existence or non-existence of exponentially localized Wannier functions is a topological characteristic of the bands [Kun]. The first answer to the question of existence of exponentially localized Wannier functions was given by Kohn [Koh73] for one-dimensional systems in the case of a single isolated band for a centrosymmetric potential. A proof of existence of exponentially localized Wannier functions in higher dimensions (also in the case of a single isolated band) is given in [DC64b, DC64a, Nen83]. The generalization to multiple bands is not straightforward. This has been investigated in [BPC+07] for two-dimensional and three-dimensional insulators (i.e, systems having a set of isolated bands). The authors first observed that the Chern numbers (see [BPC+07] for their rigorous definition) vanish for insulators with a real-valued potential. Then the following equivalence is proven : *exponentially decaying Wannier functions can be constructed if and only if all of the Chern numbers are zero*. Roughly speaking, the difference between single and multiple bands lies in the Abelian (commutative multiplication of numbers) or non-Abelian (non-commutative multiplication of matrices) character of the respective gauge transformations [MSV03, BPC+07].

The generalized Bloch waves are in general complex-valued functions. What about Wannier functions? A simple criterion to ensure existence of real-valued Wannier functions is given in [BPC+07]. The Wannier functions are real-valued if the gauge transformation U satisfies $[U^{-q}]^* = U^q$ for every real $q \in \mathbb{R}^d$. Moreover, it is conjectured in [MV97] that the Wannier functions of real Hamiltonians obtained by the spread-minimization method (presented in Section 4.3.2) are real-valued up to some general phase. To the best of my knowledge, the theoretical proof of this conjecture remains an open question.

Another interesting question is : is there any link between the symmetry of the crystal and the properties of the Wannier functions ? This question was first discussed by des Cloizeaux [DC63] from the view point of group theory. Basically, a Wannier function centered at some point $A \in \Gamma$ can be chosen to satisfy the symmetries of an irreducible representation of the point-group G_A (which is a subgroup of the total space group of the crystal) that leaves A invariant. There have been numerous other theoretical and numerical works considering symmetry-adapted Wannier functions [DC63, Koh73, VBC79, Krü87, SB94, SE05, PBMM02, CZWP06].

4.3.2 Numerical Construction

Let us now describe briefly the Marzari-Vanderbilt (MV) method proposed in [MV97] for the practical construction of maximally localized Wannier functions (MLWF). The MV procedure searches (iteratively) for a gauge transformation that leads to Wannier functions with minimal spreads around their centers. More precisely, assume that the periodic Schrödinger operator has an isolated set of N bands (in the sense of Definition 4.14) and consider the functional

$$\Omega : \begin{cases} \mathcal{G} & \to & \mathbb{R} \\ U & \mapsto & \Omega(U) = \sum_{n=1}^{N} \int_{\mathbb{R}^3} |x w_{0,n}(x)|^2 \, \mathrm{d}x - \left(\int_{\mathbb{R}^d} w_{0,n}(x) x \mathrm{d}x \right)^2 \end{cases} \tag{4.21}$$

which measures the quadratic spreads of the Wannier functions around their centers. Recall that the dependence on the transformation U is hidden in the definition of $w_{0,n}$ (see (4.19) and (4.20)). The functional Ω can be decomposed into a sum of a gauge-dependent part $\Omega_\sharp(U)$ and a gauge-independent part $\Omega_\perp$ where

$$\Omega_\sharp(U) = \sum_{n=1}^{N} \sum_{\mathbf{R} \in \mathcal{R}} \sum_{m=1}^{N} \left| \int_{\mathbb{R}^d} x \overline{w_{\mathbf{R},m}}(x) w_{0,n}(x) \mathrm{d}x \right|^2 .$$

Given a set of Bloch waves associated to the isolated energy bands, the aim in the MV algorithm is to find the choice of U that minimizes the value of $\Omega_\sharp(U)$. An expression for the gradient of Ω with respect to an infinitesimal variation δU of the gauge transformation is provided. We refer the reader to the original paper [MV97] and to [MSV03, Kun] for further details of the computation. Finally, the functional $\Omega_\sharp$ can be minimized by a sequence of gauge transformations obtained from an iterative gradient procedure. The MV algorithm is theoretically analyzed in [PP13] where it was proven that the minimizers of Ω do exist for $d \leq 3$. Moreover, under the assumption that the system has an isolated set of N bands, which together with the assumption that the potential is real-valued imply that the Chern numbers vanish, the exponential decay of the MV minimizers is shown in three different cases: i) $N = 1$ and $1 \leq d \leq 3$, ii) $N \geq 1$ and $1 \leq d \leq 2$, iii) $2 \leq N \leq 3$ and $d = 3$.

The MV algorithm has become a standard tool for Wannier functions construction since its implementation as Wannier90 computer program. However (like any local optimization approach) it suffers from two problems : how to determine a good initial guess ? and how to avoid non-global minima? Finding a good initial guess for the MV algorithm requires to find a continuous gauge transformation, which is a mathematically non-trivial task. The authors in [CLPS17] showed that the issue of "false local minima" occurs when the initial guess corresponds to a gauge transformation with vortex-like discontinuities, which may prevent the convergence of the MV optimization algorithm. Moreover, they proposed an algorithm based on the theoretical works [CHN16, FMP16] which is easy to implement. Supported by numerical tests, the algorithm in [CLPS17] is conjectured to produce continuous gauge transformation, but no theoretical proof of this conjecture is available yet. Moreover, the resulting Wannier functions are only algebraically decaying (and not exponentially). Nevertheless, in practice, the algorithm of [CLPS17] provides a good initial guess for the MV procedure.

Let us lastly point out that the MV algorithm does not exploit the symmetries of the crystal [SMV01, THJ05]. A constructive procedure to obtain symmetry-adapted

MLWFs has been proposed by Sakuma [Sak13], based on the theoretical works by des Cloizeaux. Sakuma's procedure consists in minimizing the functional Ω under suitable symmetry constraints on the transformation U. This algorithm was recently implemented in WANNIER90.

4.3.3 Applications

The Wannier functions are widely used in several contexts. We discuss here briefly some aspects related to their use in tight-binding approximations and refer the reader to [MMY+12] for an exhaustive list of applications.

The tight-binding method (TB) is an approximation method to calculate electronic band structures and other interesting properties. It is similar to the "classical" method of Linear Combination of Atomic Orbitals (LCAO) used by chemists to construct molecular orbitals. It often produces accurate results for complicated structures for which the first-principles calculations are too costly.

From a physical point of view, the main assumption in TB models is that electrons are tightly bound to the atomic sites. From a mathematical point of view, a TB model can be seen as the approximation of the Bloch states by a combination of localized functions. One particular choice for these localized functions is Wannier functions.

One particular example of application of Wannier functions is the study of the energy bands of heterogenuous structures composed of multiple (stacked) layers of 2D materials as shown in Figure 4.3. Due to the loss of periodicity caused by the twist angles between the different layers, the Bloch Floquet theory does not apply in general. Other methods involving super cells may require large calculation times. A reasonable TB approximation was recently proposed in [FK16] to study the electronic structure of such heterogeneous stacked layers. The idea is to consider each monolayer independently : build its Wannier functions and compute the TB matrix elements. Then, explicit expressions are proposed to model the interlayer couplings. These expressions involve empirical parameters and depend on geometrical parameters such as the distance between the layers and twist angles. This model allows one to study the variation of the energy bands (and other physically interesting quantities) as a function of the geometrical parameters.

The TB matrix elements are obtained from the evaluation of integrals involving Wannier functions. It is crucial to rapidly and efficiently compute these integrals in order to be able to investigate a large number of configurations (varying the number of layers, the types of materials, the relative distances and the twist angles). In practice, these integrals are approximated numerically since the Wannier functions (generated by WANNIER90) are provided on grids. The computation cost of such a numerical integration is of order M^3 where M is the number of grid points.

4.3.4 Contribution of the Thesis (Wannier Compression)

Our contribution consists in the development of a greedy algorithm for the compression of Wannier functions into Gaussian-polynomials type orbitals. Our procedure takes into account the symmetry of the Wannier function (if any) and allows one to store only a small number of parameters instead of storing all the M values of the Wannier function

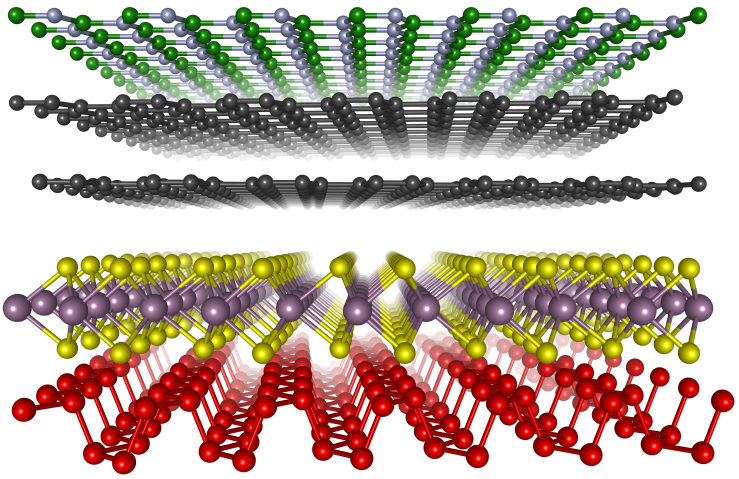


Figure 4.3 – Visualization of a 2D heterostructure. From top to bottom : hBN, Graphene, Graphene, MoS2, Phosphorene. Courtesy of Paul Cazeaux (Department of Mathematics, University of Kansas, Lawrence, USA).

on a grid, where M is the number of grid points. A second added value of the compression using Gaussian-type functions is to allow one to derive closed formulas for several quantities that involve Wannier functions such as tight-binding matrix elements. This aims in particular at accelerating electronic structure calculations for 2D heterogeneous layers, and thus allowing one to explore a larger number of configurations.

More precisely, we consider that we are given a real-valued Wannier function $W : \mathbb{R}^3 \to \mathbb{R}$ centered at a point $A \in \Gamma$ and assume that A corresponds to a one-dimensional representation of a symmetry point-group G_A leaving A invariant. Thus, W satisfies the property

$$\forall \Theta \in G_A, \quad (\Theta W)(\mathbf{r}) = \chi(\Theta) W(\mathbf{r}), \quad \forall \mathbf{r} \in \mathbb{R}^3 \tag{4.22}$$

where χ is the character of this one-dimensional representation. Consider symmetry-adapted Gaussian-type orbitals (SAGTO) of the form

$$\phi^{\mathrm{SA}}_{\boldsymbol{\alpha},\sigma,\Lambda}(\mathbf{r}) = \frac{1}{|G_A|} \sum_{\Theta \in G_A} \chi(\Theta)\, (\Theta \varphi_{\boldsymbol{\alpha},\sigma,\Lambda})(\mathbf{r}) = \frac{1}{|G_1|} \sum_{\Theta \in G_1} \chi(\Theta)\, \varphi_{\boldsymbol{\alpha},\sigma,\Lambda}(\Theta^{-1}\mathbf{r}), \tag{4.23}$$

where $|G_A|$ is the order of the group G_A, and where

$$\varphi_{\boldsymbol{\alpha},\sigma,\Lambda}(\mathbf{r}) = \left(\sum_{(n_x,n_y,n_z) \in \mathcal{I}} \lambda_{n_x,n_y,n_z} (r_x - \alpha_x)^{n_x} (r_y - \alpha_y)^{n_y} (r_z - \alpha_z)^{n_z} \right) \exp\left(-\frac{1}{2\sigma^2} |\mathbf{r} - \boldsymbol{\alpha}|^2 \right)$$

is a Gaussian-polynomial function centered at $\boldsymbol{\alpha} \in \mathbb{R}^3$ with standard deviation $\sigma > 0$. The set $\mathcal{I}$ is a subset of $\left\{ (n_x, n_y, n_z) \in \mathbb{N}^3 \mid n_x + n_y + n_z \leq L, \quad L \in \mathbb{N}^* \right\}$ determined by the symmetries of W.

The goal is to approximate the Wannier function W by a function $\widetilde{W}$ which is a finite sum of SAGTOs $\widetilde{W}(\mathbf{r}) = \sum_{n=1}^{p} \phi^{\mathrm{SA},(\mathrm{n})}_{\boldsymbol{\alpha},\sigma,\Lambda}(\mathbf{r})$ so that the H^s error $\|W - \widetilde{W}\|_{H^s(\mathbb{R}^3)}$ is minimized. To do so, we use a greedy algorithm that allows us to (iteratively) construct a sequence of approximations $\widetilde{W}_0$, $\widetilde{W}_1$, $\widetilde{W}_2, \cdots$ such that the error $\|W - \widetilde{W}_p\|_{H^s}$ tends to 0 as p goes to $+\infty$. We implemented our algorithm in the Fourier space

so that we can minimize the H^s error for any value of s. This work is the topic of the paper [BCC+17] written in collaboration with Eric Cancès (*Université Paris Est, CERMICS (ENPC), INRIA Paris*), Paul Cazeaux (*Department of Mathematics, University of Kansas, Lawrence, USA*), Shiang Fang and Efthimios Kaxiras (*Department of Physics, Harvard University, Cambridge, USA*). As a preview result, we show in Figure 4.4 a comparison between a Wannier function obtained with WANNIER90 and its approximation by a finite sum of SAGTOs where the H^1 error is minimized. The content of [BCC+17] is reported in Chapter 6.

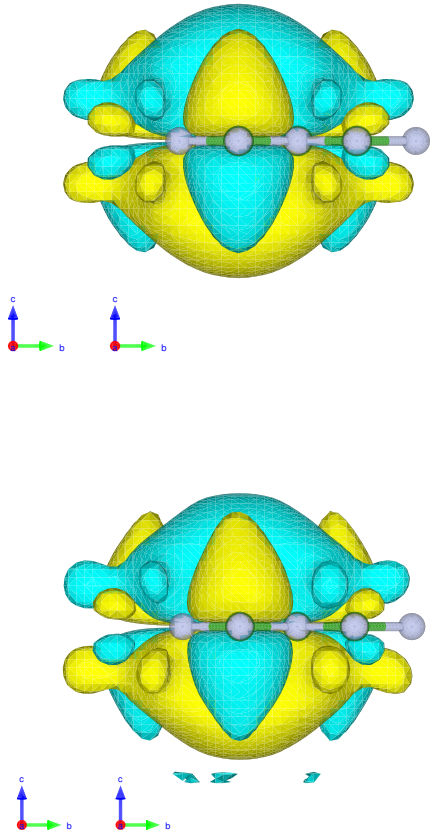


Figure 4.4 – Wannier function of single-layer hBN generated by Wannier90 (top), and its compression into SAGTOs (bottom). Positive and negative iso-surfaces corresponding to 15% of the maximum value are plotted. Visualization using VESTA [MI08].

4.4 Appendix : Numerical Approximation of the Band Structure

We present here one of the most popular numerical methods for the approximation of the spectrum of periodic Schrödinger operators: the **plane-wave discretization**. The periodic Schrödinger operator $A = -\Delta + V$ on L^2 is characterized by the spectra of the Bloch operators $A_q = -(\mathrm{i}\nabla + q)^2 + V$ for $q \in \Gamma^*$ acting on L^2_{per} with domain H^2_{per} and form domain H^1_{per}. Each operator A_q is self-adjoint and has compact resolvent and thus admits only discrete eigenvalues. The goal of this appendix is to show how to numerically solve the family of eigenvalue problems

$$A_q u_q = \varepsilon_q^V u_q, \quad \text{and} \quad \|u_q\|_{L^2_{\mathrm{per}}} = 1, \quad q \in \Gamma^* \tag{4.24}$$

For every $q \in \Gamma^*$, the eigenvalue problem (4.24) can be written under the variational form : find $(u_q, \varepsilon_q^V) \in H^1_{\mathrm{per}} \times \mathbb{R}$ such that

$$a_q(w, u_q) = \varepsilon_q^V \langle w, u_q \rangle \quad \forall w \in H^1_{\mathrm{per}} \quad \text{and} \quad \|u_q\|_{L^2_{\mathrm{per}}} = 1. \tag{4.25}$$

where the bilinear form a_q is defined for every $w, v \in H^1_{per} \times H^1_{\mathrm{per}}$ by

$$a_q(w, v) = \int_\Gamma \left[\overline{(\nabla + \mathrm{i}q)w}\right] \cdot [(\nabla + \mathrm{i}q)v] + \int_\Gamma V \overline{w} v.$$

The plane-wave method is a Galerkin approximation of the variational problem (4.25) in the Fourier space. More precisely, for all $k \in \mathcal{R}^*$, let $\mathbf{e}_k(x) := |\Gamma|^{-1/2} e^{\mathrm{i}k\cdot x}$ denote the plane-wave associated with the wave-vector $k \in \mathcal{R}^*$. For a given $s > 0$, let us define the finite dimensional space $X_s \subset H^1_{\mathrm{per}}$ as follows

$$X_s := \mathrm{Span}\left\{\mathbf{e}_k | \quad k \in \mathcal{R}^*, \ |k|^2 \leq s\right\} \tag{4.26}$$

and denote by N_s its dimension and by $\Pi_{X_s} : L^2_{\mathrm{per}} \to X_s$ the L^2_{per} orthogonal projector onto X_s.

Problem (4.25) is approximated by the discrete version : find $(u_q^s, \varepsilon_q^{V,s}) \in X_s \times \mathbb{R}$ such that

$$a_q(w, u_q^s) = \varepsilon_q^{V,s} \langle w, u_q^s \rangle \quad \forall w \in X_s \quad \text{and} \quad \|u_q^s\|_{L^2_{\mathrm{per}}} = 1. \tag{4.27}$$

where for every $x \in \mathbb{R}^d$,

$$u_q^s(x) = (\Pi_{X_s} u_q)(x) := \sum_{k \in \mathbb{Z}^d, |k|^2 \leq s} \hat{u}^s_{q,n,k} \, \mathbf{e}_k(x) \quad \text{with} \quad \sum_{k \in \mathbb{Z}^d, |k|^2 \leq s} |\hat{u}^s_{q,n,k}|^2 = 1.$$

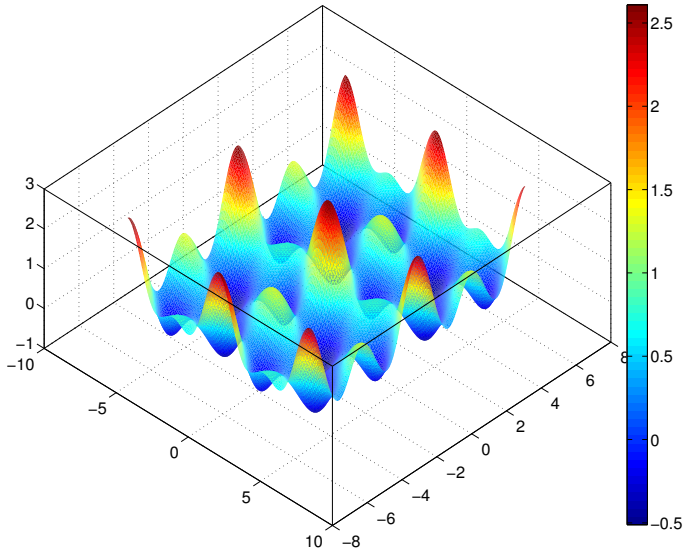
We denote by $U_q^s \in \mathbb{C}^{N_s}$ the vector of Fourier coefficients $(\hat{u}^s_{q,n,k})_{|k|^2 \leq s}$ and introduce the Hamiltonian matrix $\mathcal{H}_q^s \in \mathbb{C}^{N_s \times N_s}$ as follows

$$\left(\mathcal{H}_q^s\right)_{k,l} := a_q^p(\mathbf{e}_l, \mathbf{e}_k) = \begin{cases} |k+q|^2 + \hat{V}_0 & \text{if} \quad k = l, \\ \hat{V}_{k-l} & \text{if} \quad k \neq l. \end{cases} \tag{4.28}$$

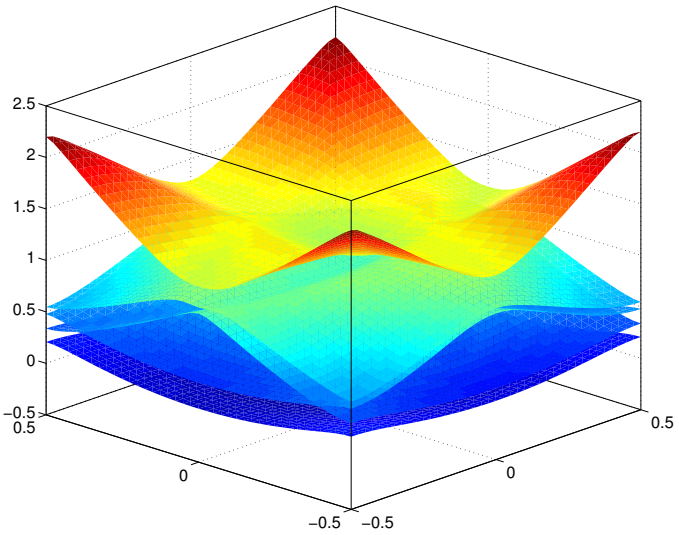
Finally, Problem (4.27) can be written as the matrix eigenvalue problem: find $(U_q^s, \varepsilon_q^{V,s}) \in \mathbb{C}^{N_s} \times \mathbb{R}$ such that

$$\mathcal{H}_q^s U_q^s = \varepsilon_q^{V,s} U_q^s \quad \|u_q^s\|_2 = 1 \tag{4.29}$$

It suffices lastly to compute the eigenvalues $\varepsilon_{q,n}^{V,s}$ and normalized eigenvectors $U_{q,n}^{s}$ for $1 \leq n \leq N_s$ of the Hermitian matrix $\mathcal{H}_q^s$. Several numerical methods allow to do this efficiently [Saa03]. To illustrate the method, let us calculate the low-energy bands associated to a two-dimensional real-valued $2\pi\mathbb{Z}^2-$periodic potential. The result is given in Figure 4.5.



(a) Potential V



(b) Lowest energy bands of $-\Delta + V$

Figure 4.5 – Numerical computation of a 2D band structure with the plane-wave discretization. The first Brillouin zone $[-\frac{1}{2}, \frac{1}{2}]^2$ is uniformly discretized using 41×41 points. The dimension of the used approximation space X_s is $N_s = 81$.

CHAPTER 5

RECONSTRUCTION OF THE FIRST BAND(S) IN AN INVERSE HILL'S PROBLEM

We report in this chapter the results of [BEG17] obtained with Virginie Ehrlacher and David Gontier.

Abstract. This paper concerns an inverse band structure problem for one dimensional periodic Schrödinger operators (Hill's operators). Our goal is to find a potential for the Hill's operator in order to reproduce as best as possible some given target bands, which may not be realisable. We recast the problem as an optimisation problem, and prove that this problem is well-posed when considering singular potentials (Borel measures). We then propose different algorithms to tackle the problem numerically.

Contents

5.1 Introduction

The aim of this article is to present new considerations on an inverse band structure problem for periodic one-dimensional Schrödinger operators, also called Hill's operators. A Hill operator is a self-adjoint, bounded from below operator of the form $A^V := -\frac{d^2}{dx^2} + V$, acting on $L^2(\mathbb{R})$, and where V is a periodic real-valued potential. Its spectrum is composed of a reunion of intervals, which can be characterised using Bloch-Floquet theory as the reunion of the spectra of a family of self-adjoint compact resolvent operators A_q^V, indexed by an element $q \in \mathbb{R}$ called the *quasi-momentum* or *k-point* (see [**?**, Chapter XIII] and Section 5.2.1). The m^{th} band function associated to a periodic potential is the function which maps $q \in \mathbb{R}$ to the m^{th} lowest eigenvalue of A_q^V. The properties of these band functions are well-known, especially in the one-dimensional case (see e.g. [RS78b, Chapter XIII]).

The inverse band structure problem is an interesting mathematical question of practical interest, which can be roughly formulated as follows: *is it possible to find a potential V so that its first bands are close to some target functions?*

A wide mathematical literature answers the question when the target functions are indeed the bands of some Hill's operator, corresponding to some V_{ref}. In this case, we need to *recover* a potential V that reproduces the bands of V_{ref}. We refer to [Esk89, ERT84a, ERT84b, PT87, FY01, Vel15] for the case when V_{ref} is a regular potential, and to [HM03a, HM04a, HM04b, HM03b, HM06] when V_{ref} is singular (see also the review [Kuc16]). The main ideas of the previous references are as follows. First, the band structure of a Hill's operator can be seen as the transformation of an analytic function. In particular, the knowledge of any band on an open set is enough to recover *theoretically* the whole band structure. A potential is then reconstructed from the high energy asymptotics of the bands.

The previous methods use the knowledge of the behaviour of the high energy bands, and therefore are unsuitable for practical purpose (material design) since we usually have no accurate and numerically stable information about these high energy bands. Moreover, in practice, only the low energy bands are usually of interest. The fact that there exists no explicit characterisation of the set of the first band functions associated to a given admissible set of periodic potentials is an additional numerical difficulty. For applications, it is therefore interesting to know how to construct a potential such that only its first bands are close to some given target functions, which may not be realisable (for instance not analytic). In this present work, we therefore adopt a different point of view, which, up to our best knowledge, has not been studied: we recast the inverse problem as an optimisation problem.

The outline of the paper is as follows. In Section 5.2, we recall basic properties about Hill's operators with singular potentials. and we state our main result (Theorem 5.3). Its proof is given in Section 5.3. Finally, we present in Section 5.4 some numerical tests and propose an adaptive optimisation algorithm, which is observed to converge faster than the standard one. This adaptive algorithm relies on the use of an a posteriori error estimator for discretised eigenvalue problems, whose computation is detailed in the Appendix.

5.2 Spectral decomposition of periodic Schrödinger operators, and main results

In this section, we recall some properties of Hill's operators with singular potentials. Elementary notions on the Bloch-Floquet transform [RS78b] are gathered in Section 5.2.1. The spectral decomposition of one-dimensional periodic Schrödinger operators with singular potentials is detailed in Section 5.2.2, building on the results of [Kat72, HM01, GZ06, MM08, DJP16]. We state our main results in Section 5.2.3.

5.2.1 Bloch-Floquet transform

We need some notation. Let $\mathcal{D}'$ denotes the Schwartz space of complex-valued distributions, and let $\mathcal{D}'_{\rm per} \subset \mathcal{D}'$ be the space of distributions that are 2π-periodic. In the sequel, the unit cell is $\Gamma := [-\pi, \pi)$, and the reciprocal unit cell (or Brillouin zone) is $\Gamma^* := [-1/2, 1/2]$. For $u \in \mathcal{D}'_{\rm per}$ and $k \in \mathbb{Z}$, the k^{th} normalised Fourier coefficient of u is denoted by $\widehat{u}(k)$. For $s \in \mathbb{R}$, we denote by

$$H^s_{\rm per} := \left\{ u \in \mathcal{D}'_{\rm per},\ \|u\|^2_{H^s_{\rm per}} := \sum_{k\in\mathbb{Z}} (1+|k|^2)^s |\widehat{u}(k)|^2 < +\infty \right\}$$

the complex-valued periodic Sobolev space, which is a Hilbert space when endowed with its natural inner product. We write $H^s_{{\rm per},r}$ for the *real-valued* periodic Sobolev space, i.e.

$$H^s_{{\rm per},r} := \left\{ u \in H^s_{\rm per}, \quad \forall k \in \mathbb{Z},\ \widehat{u}(-k) = \overline{\widehat{u}(k)} \right\}.$$

We also let $L^2_{\rm per} := H^{s=0}_{\rm per}$. From our normalisation, it holds that

$$\forall v, w \in L^2_{\rm per},\ \langle v, w\rangle_{L^2_{\rm per}} = \int_\Gamma \overline{v} w \quad \text{and} \quad \forall v, w \in H^1_{\rm per},\ \langle v, w\rangle_{H^1_{\rm per}} = \int_\Gamma \frac{d\overline{v}}{dx}\frac{dw}{dx} + \int_\Gamma \overline{v} w.$$

Lastly, we denote by $C^0_{\rm per}$ the space of 2π-periodic continuous functions, and by C^∞_c the space of C^∞ functions over $\mathbb{R}$, with compact support.

To introduce the Bloch-Floquet transform, we let $\mathcal{H} := L^2(\Gamma^*, L^2_{\rm per})$. For any element $f \in \mathcal{H}$, we denote by $f_q(x)$ its value at the point $(q, x) \in \Gamma^* \times \Gamma$. The space $\mathcal{H}$ is an Hilbert space when endowed with its inner product

$$\forall f, g \in \mathcal{H}, \quad \langle f, g\rangle_{\mathcal{H}} := \int_{\Gamma^*}\int_\Gamma \overline{f_q(x)} g_q(x) dx\ dq.$$

The Bloch-Floquet transform is the map $\mathcal{B} : L^2(\mathbb{R}) \to \mathcal{H}$ defined, for smooth functions $\varphi \in C^\infty_c(\mathbb{R})$, by

$$\phi_q(x) := (\mathcal{B}\varphi)_q(x) := \sum_{R\in\mathbb{Z}} \varphi(x+R) e^{-{\rm i}q(R+x)}.$$

It is an isometry from $L^2(\mathbb{R})$ to $\mathcal{H}$, whose inverse is given by

$$\left(\mathcal{B}^{-1}\phi\right)(x) := \int_{\Gamma^*} \phi_q(x) e^{iqx}\, dq = \varphi(x).$$

The Bloch theorem states that if A is a self-adjoint operator on $L^2(\mathbb{R})$ with domain $D(A)$ that commutes with $\mathbb{Z}$-translations, then $\mathcal{B}A\mathcal{B}^{-1}$ is diagonal in the q-variable.

More precisely, there exists a unique family of self-adjoint operators $(A_q)_{q\in\Gamma^*}$ on L^2_{per} such that for all $\varphi \in L^2(\mathbb{R}) \cap D(A)$,

$$(A\varphi)(x) = \int_{\Gamma^*} (A_q\phi_q)(x)\, dq.$$

In this case, we write

$$A = \int_{\Gamma^*}^{\oplus} A_q dq.$$

5.2.2 Hill's operators with singular potentials

Giving a rigorous mathematical sense to a Hill's operator of the form $-\frac{d^2}{dx^2} + V$ on $L^2(\mathbb{R})$, when the potential V is singular is not an obvious task. In the present paper, we consider $V \in H^{-1}_{\mathrm{per,r}}$, which is a case that was first tackled in [Kat72] (see also [HM01, DJP16, GZ06, MM08] for recent results).

The results which are gathered in this section are direct corollaries of results which were proved in these earlier works, particularly in [HM01].

Proposition 5.1. *[Theorem 2.1 and Lemma 3.2 of [HM01]] For all $V \in H^{-1}_{\mathrm{per,r}}$, there exists $\sigma_V \in L^2_{\mathrm{per}}$ and $\kappa_V \in \mathbb{R}$ such that*

$$V = \sigma_V' + \kappa_V \text{ in } \mathcal{D}'_{\mathrm{per}}. \tag{5.1}$$

Moreover, if $a^V : H^1(\mathbb{R}) \times H^1(\mathbb{R}) \to \mathbb{C}$ is the sesquilinear form defined by

$$\forall v, w \in H^1(\mathbb{R}), \quad a^V(v,w) = \int_{\mathbb{R}} \frac{d\overline{v}}{dx}\frac{dw}{dx} + \int_{\mathbb{R}} \kappa_V \overline{v} w - \int_{\mathbb{R}} \sigma_V \left(\frac{d\overline{v}}{dx} w + \overline{v}\frac{dw}{dx}\right), \tag{5.2}$$

then a^V is a symmetric, continuous sesquilinear form on $H^1(\mathbb{R})\times H^1(\mathbb{R})$, which is closed and bounded from below. Besides, a^V is independent of the choice of $\sigma_V \in L^2_{\mathrm{per}}$ and $\kappa_V \in \mathbb{R}$ satisfying (5.1).

Remark 5.2. *The expression (5.2) makes sense whenever $v, w \in H^1(\mathbb{R})$. This can be easily seen with the Cauchy-Schwarz inequality, and the embedding $H^1(\mathbb{R}) \hookrightarrow L^\infty(\mathbb{R})$. It is not obvious how to extend this result to higher dimension.*

A direct consequence of Proposition 5.1 is that one can consider the Friedrichs operator on $L^2(\mathbb{R})$ associated to a^V, which is denoted by A^V in the sequel. The operator A^V is thus a densely defined, self-adjoint, bounded from below operator on $L^2(\mathbb{R})$, with form domain $H^1(\mathbb{R})$ and whose domain is dense in $L^2(\mathbb{R})$. Formally, it holds that

$$A^V = -\frac{\partial^2}{\partial x^2} + V.$$

The spectral properties of the operator A^V can be studied (like in the case of regular potentials) using Bloch-Floquet theory.

The previous result, together with Bloch-Floquet theory, allows to study the operator A^V via its Bloch fibers $\left(A^V_q\right)_{q\in\Gamma^*}$. For $q \in \Gamma^*$, it holds that A^V_q is the self-adjoint extension of the operator

$$\left|-i\frac{d}{dx} + q\right|^2 + V.$$

It holds that A_q^V is a bounded from below self-adjoint operator acting on L^2_{per}, whose form domain is H^1_{per}, and with associated quadratic form a_q^V, defined by (recall that H^1_{per} is an algebra)

$$\forall v, w \in H^1_{\mathrm{per}}, \quad a_q^V(v,w) := \int_\Gamma \left[\overline{\left(-i\frac{d}{dx} + q\right) v} \left(-i\frac{d}{dx} + q\right) w \right] + \langle V, \overline{v}w \rangle_{H^{-1}_{\mathrm{per}}, H^1_{\mathrm{per}}}. \tag{5.3}$$

In other words, we have

$$A^V = \int_{\Gamma^*}^{\oplus} A_q^V \, dq.$$

The fact that L^2_{per} is compactly embedded in H^1_{per} implies that A_q^V is compact-resolvent. As a consequence, there exists a non-decreasing sequence of real eigenvalues $\left(\varepsilon^V_{q,m}\right)_{m\in\mathbb{N}^*}$ going to $+\infty$ and a corresponding orthonormal basis $(u^V_{q,m})_{m\in\mathbb{N}^*}$ of L^2_{per} such that

$$\forall m \in \mathbb{N}^*, \quad A_q^V u^V_{q,m} = \varepsilon^V_{q,m} u^V_{q,m}. \tag{5.4}$$

The map $\Gamma^* \ni q \mapsto \varepsilon^V_{q,m}$ is called the m^{th} band. Since the potential V is real-valued, it holds that $A^V_{-q} = \overline{A^V_q}$, so that $\varepsilon^V_{-q,m} = \varepsilon^V_{q,m}$ for all $q \in \Gamma^*$ and $m \in \mathbb{N}^*$. This implies that it is enough to study the bands on $[0, 1/2]$. Actually, we have

$$\sigma(A^V) = \bigcup_{q\in[0,1/2]} \bigcup_{m\in\mathbb{N}^*} \{\varepsilon^V_{q,m}\}.$$

In the sequel, we mainly focus on the first band. We write $\varepsilon_q^V := \varepsilon^V_{q,1}$ for the sake of clarity. Thanks to the knowledge of the form domain of A_q^V, we know that

$$\varepsilon_q^V := \min_{\substack{v\in H^1_{\mathrm{per}} \\ \|v\|_{L^2_{\mathrm{per}}}=1}} a_q^V(v,v). \tag{5.5}$$

This characterisation will be the key to our proof. When the potential V is smooth (say $V \in L^2_{\mathrm{per}}$), then the map $\Gamma^* \ni q \mapsto \varepsilon^V_{q,m}$ is analytic on $(-1/2, 1/2)$. Besides, it is increasing on $[0, 1/2]$ if m is odd, and decreasing if m is even (see e.g. [RS78b, Chapter XIII]).

5.2.3 Main results

The goal of this article is to find a potential V so that the bands of the corresponding Hill's operator are close to some given target functions. In order to do so, we recast the problem as a minimisation one, of the form

$$V^* \in \mathrm{argmin}_{V\in\mathcal{V}} \mathcal{J}(V).$$

Unfortunately, we were not able to consider the full setting where the minimisation set $\mathcal{V}$ is the whole set $H^{-1}_{\mathrm{per},r}$. The problem was that we were unable to control the negative part of V. To bypass this difficulty, we chose to work with potentials that are bounded from below. Such a distribution is necessary a measure (see e.g. [LL01]).

Hence measure-valued potentials provide a natural setting for band reconstruction. We recall here some basic properties about measures.

We denote by $\mathcal{M}^+_{\mathrm{per}}$ the space of non-negative $2\pi-$periodic regular Borel measures on $\mathbb{R}$, in the sense that for all $\nu \in \mathcal{M}^+_{\mathrm{per}}$, and all Borel set $S \in \mathcal{B}(\mathbb{R})$, it holds that $\nu(S) = \nu(S + 2\pi) \geq 0$, and $\nu(\Gamma) < \infty$. For all $\epsilon > 0$, from the Sobolev embedding $H^{1/2+\varepsilon}_{\mathrm{per}} \hookrightarrow C^0_{\mathrm{per}}$, we deduce that $\mathcal{M}^+_{\mathrm{per}} \hookrightarrow H^{-1/2-\varepsilon}_{\mathrm{per}} \hookrightarrow H^{-1}_{\mathrm{per}}$, where the last embedding is compact. For $\nu \in \mathcal{M}^+_{\mathrm{per}}$, we denote by $V_\nu \in H^{-1}_{\mathrm{per},r}$ the unique corresponding potential, which is defined by duality through the relation:

$$\forall \phi \in H^1_{\mathrm{per}}, \quad \int_\Gamma \phi d\nu = \langle V_\nu, \phi \rangle_{H^{-1}_{\mathrm{per}}, H^1_{\mathrm{per}}}.$$

For $B \in \mathbb{R}$, we define the set of B-bounded from below potentials

$$\mathcal{V}_B := \left\{ V \in H^{-1}_{\mathrm{per},r} | \quad \exists \nu \in \mathcal{M}^+_{\mathrm{per}}, \quad V = V_\nu - B \right\} \subset H^{-1}_{\mathrm{per,r}}.$$

This will be our minimisation space for our optimisation problem. Note that $\mathcal{V}_{B_1} \subset \mathcal{V}_{B_2}$ for $B_1 \geq B_2$.

We now introduce the functional $\mathcal{J}$ to minimise. First, we introduce the set $\mathcal{T}$ of allowed target functions:

$$\mathcal{T} := \left\{ b \in C^0(\Gamma^*), \quad b \text{ is even and } b \text{ is increasing on } [0, 1/2] \right\}. \tag{5.6}$$

Of course, for all $V \in H^{-1}_{\mathrm{per},r}$, it holds that $\Gamma^* \ni q \mapsto \varepsilon^V_q \in \mathcal{T}$. Finally, in order to quantify the quality of reconstruction of a band $b \in \mathcal{T}$, we introduce the error functional $\mathcal{J}_b : H^{-1}_{\mathrm{per,r}} \to \mathbb{R}$ defined by

$$\forall V \in H^{-1}_{\mathrm{per,r}}, \quad \mathcal{J}_b(V) := \frac{1}{2} \int_{\Gamma^*} |b(q) - \varepsilon^V_q|^2 \, dq = \int_0^{1/2} |b(q) - \varepsilon^V_q|^2 \, dq. \tag{5.7}$$

The main result of the present paper is the following.

Theorem 5.3. *Let $b \in \mathcal{T}$, and denote by $b^* := \fint_{\Gamma^*} b(q)\, dq \in \mathbb{R}$. Then, for all $B > 1/4 - b^*$, there exists a solution $V_{b,B} \in \mathcal{V}_B$ to the minimisation problem*

$$V_{b,B} \in \mathrm{argmin}_{V \in \mathcal{V}_B} \mathcal{J}_b(V). \tag{5.8}$$

The proof of Theorem 5.3 relies on the following proposition, which is central to our analysis. Both the proofs of Theorem 5.3 and Proposition 5.4 are provided in the next section.

Proposition 5.4. *Let $B \in \mathbb{R}$ and let $(V_n)_{n \in \mathbb{N}^*} \subset \mathcal{V}_B$. For all $n \in \mathbb{N}^*$, let $\nu_n \in \mathcal{M}^+_{\mathrm{per}}$ such that $V_n := V_{\nu_n} - B$. Let us assume that the sequence $\left(\varepsilon^{V_n}_0\right)_{n \in \mathbb{N}^*}$ is bounded and such that $\nu_n(\Gamma) \underset{n \to +\infty}{\longrightarrow} +\infty$. Then, up to a subsequence (still denoted n), the functions $q \mapsto \varepsilon^{V_n}_q$ converge uniformly to a constant function $\varepsilon \in \mathbb{R}$, with $\varepsilon \geq \frac{1}{4} - B$. In other words, there is $\varepsilon \geq \frac{1}{4} - B$ such that*

$$\max_{q \in [0,1/2]} \left| \varepsilon^{V_n}_q - \varepsilon \right| \underset{n \to \infty}{\longrightarrow} 0. \tag{5.9}$$

Conversely, for all $\varepsilon \geq \frac{1}{4} - B$, there is a sequence $(V_n)_{n \in \mathbb{N}^*} \subset \mathcal{V}_B$ such that (5.9) holds.

This result implies that the first band of the sequence of operators $\left(A^{V_n}\right)_{n\in\mathbb{N}^*}$, where $(V_n)_{n\in\mathbb{N}^*}$ satisfies the assumptions of Proposition 5.4, *becomes flat.*

Remark 5.5. *Here we have a sequence of first bands $\left(\varepsilon_q^{V_n}\right)_{n\in\mathbb{N}^*}$ that converges uniformly to a constant function. However, as the first band of any Hill's operator must be increasing and analytic, the limit* is not *the first band of a Hill's operator.*

5.3 Proof of Theorem 5.3 and Proposition 5.4

5.3.1 Preliminary lemmas

We first prove some intermediate useful lemmas before giving the proof of Proposition 5.4 and Theorem 5.3. We start by recording a spectral convergence result.

Proposition 5.6. *[Theorem 4.1 [HM01]] Let $(V_n)_{n\in\mathbb{N}^*} \subset H^{-1}_{\mathrm{per},r}$ be a sequence such that $(V_n)_{n\in\mathbb{N}^*}$ converges strongly in H^{-1}_{per} to some $V \in H^{-1}_{\mathrm{per},r}$. Then,*

$$\forall m \in \mathbb{N}^*, \quad \max_{q\in[0,1/2]} \left|\varepsilon_{q,m}^{V_n} - \varepsilon_{q,m}^{V}\right| \xrightarrow[n\to\infty]{} 0.$$

In our case, since we are working with potentials that are measures, we deduce the following result.

Proposition 5.7. *Let $B \in \mathbb{R}$ and $(V_n)_{n\in\mathbb{N}^*} \subset \mathcal{V}_B$ be a bounded sequence, in the sense*

$$\sup_{n\in\mathbb{N}} \langle V_n, \mathbb{1}_\Gamma\rangle_{H^{-1}_{\mathrm{per}},H^1_{\mathrm{per}}} < \infty.$$

For all $n \in \mathbb{N}^*$, let $\nu_n \in \mathcal{M}^+_{\mathrm{per}}$ such that $V_n = V_{\nu_n} - B$. Then, there exists $\nu \in \mathcal{M}^+_{\mathrm{per}}$ such that, up to a subsequence (still denoted n), $(\nu_n)_{n\in\mathbb{N}}$ converges weakly- to ν in $\mathcal{M}_{\mathrm{per}}$, and $(V_n)_{n\in\mathbb{N}^*}$ converges strongly in H^{-1}_{per} to $V := V_\nu - B \in \mathcal{V}_B$. Moreover, it holds that*

$$\forall m \in \mathbb{N}^*, \quad \max_{q\in[0,1/2]} \left|\varepsilon_{q,m}^{V_n} - \varepsilon_{q,m}^{V}\right| \xrightarrow[n\to\infty]{} 0.$$

Proof. The fact that we can extract from the bounded sequence $(\nu_n)_{n\in\mathbb{N}^*}$ a weakly-* convergent sequence in $\mathcal{M}^+_{\mathrm{per}}$ is the Prokhorov's theorem applied in the torus Γ^*. The second part comes from the compact embedding $\mathcal{M}_{\mathrm{per}} \hookrightarrow H^{-1}_{\mathrm{per}}$. The final part is the direct application of Proposition 5.6. □

Remark 5.8. *This proposition explains our choice to consider measure-valued potentials. Note that a similar result does not hold in the L^1_{per} setting for instance.*

We now give a lemma which is standard in the case of regular potentials V (see [Eva98]).

Lemma 5.9. *Let $V \in \mathcal{V}_B$ for some $B \in \mathbb{R}$. The first eigenvector $u^V_{q=0} \in H^1_{\mathrm{per}}$ of $A^V_{q=0}$ is unique up to a global phase. It can be chosen real-valued and positive.*

Proof. We use the min-max principle (5.5), and the fact that, for $u \in H^1_{\mathrm{per}}$, the following holds

$$\left|\frac{d}{dx}|u|\right| \leq \left|\frac{d}{dx}u\right| \quad \text{a.e.}$$

We see that if u is an eigenvector corresponding to the first eigenvalue, then so is $|u|$. We now consider a non-negative eigenvector $u \geq 0$, and prove that it is positive. The usual argument is Harnack's inequality. However, it is a priori unclear that it works in our singular setting. To prove it, we write $V = V_\nu - B$ for $\nu \in \mathcal{M}^+_{\mathrm{per}}$, and consider the repartition function F_ν of ν: $F_\nu(x) := \nu((0,x])$. This function is not periodic, but the function $f_\nu(x) := F_\nu(x) - \nu(\Gamma)\frac{x}{|\Gamma|}$ is. Since F_ν is an non decreasing, right-continuous function, we deduce that $f_\nu \in L^\infty_{\mathrm{per}}$. Moreover, it holds, in the H^{-1}_{per} sense, that $f'_\nu = V_\nu - |\Gamma|^{-1}\nu(\Gamma) = V + B - |\Gamma|^{-1}\nu(\Gamma)$. As a result, we see that u is solution to the minimisation problem

$$u \in \operatorname{argmin}_{\substack{v \in H^1_{\mathrm{per},r} \\ \|v\|_{L^2_{\mathrm{per}}}=1}} \left\{ \int_\Gamma \left|\frac{dv}{dx}\right|^2 + \left(\frac{\nu(\Gamma)}{|\Gamma|} - B\right) - 2\int_\Gamma f_\nu \left(v\frac{dv}{dx}\right) \right\}.$$

There exists $\lambda \in \mathbb{R}$ so that the corresponding Euler-Lagrange equations can be written in the weak-form:

$$\operatorname{div} F(x,u,u') + B(x,u,u') = 0,$$

with

$$F(x,u,p) = p - f_\nu u \quad \text{and} \quad B(x,u,p) = f_\nu p + \lambda u.$$

We are now in the settings of [Tru67, Theorem 1.1], and we deduce that $u > 0$. The rest of the proof is standard. □

5.3.2 Proof of Proposition 5.4

We now prove Proposition 5.4. Let $B \in \mathbb{R}$ and let $V_n = V_{\nu_n} - B \in \mathcal{V}_B$ with $\nu_n \in \mathcal{M}^+_{\mathrm{per}}$, be a sequence such that the sequence $\left(\varepsilon^{V_n}_{q=0}\right)_{n\in\mathbb{N}^*}$ is bounded and $\nu_n(\Gamma)$ goes to $+\infty$. Since $\left(\varepsilon^{V_n}_0\right)_{n\in\mathbb{N}^*}$ is bounded, then up to a subsequence (still denoted by n), there exists $\varepsilon \in \mathbb{R}$ such that $\varepsilon^{V_n}_0$ converges to ε. Our goal is to prove that the convergence also holds uniformly in $q \in \Gamma^*$.

Let $u^{V_n}_0 \in H^1_{\mathrm{per}}$ be the L^2_{per}-normalised positive eigenvector of $A^{V_n}_0$ associated to the eigenvalue $\varepsilon^{V_n}_0$ (see Lemma 5.9). We denote by $\alpha_n := \min_{x\in\Gamma} u^{V_n}_0(x) > 0$. Let us first prove that the following convergences hold:

$$\alpha_n \int_\Gamma u^{V_n}_0 \, d\nu_n \xrightarrow[n\to+\infty]{} 0 \quad \text{and} \quad \alpha_n^2 \nu_n(\Gamma) \xrightarrow[n\to+\infty]{} 0. \tag{5.10}$$

From the equality

$$\int_\Gamma \left|\frac{d}{dx}\left(u^{V_n}_0\right)\right|^2 + \int_\Gamma |u^{V_n}_0|^2 d\nu_n = \varepsilon^{V_n}_0 + B,$$

we get

$$\alpha_n^2 \nu_n(\Gamma) \leq \alpha_n \int_\Gamma u^{V_n}_0 d\nu_n \leq \int_\Gamma |u^{V_n}_0|^2 d\nu_n \leq \varepsilon^{V_n}_0 + B. \tag{5.11}$$

As the right-hand side is bounded, and $\nu_n(\Gamma) \to +\infty$ by hypothesis, this implies $\alpha_n \to 0$. Moreover, we have

$$0 \leq \int_\Gamma u^{V_n}_0 \, d\nu_n = a^{V_n}_0(u^{V_n}_0, \mathbb{1}_\Gamma) + B\int_\Gamma u^{V_n}_0 = (\varepsilon^{V_n}_0 + B)\int_\Gamma u^{V_n}_0 \leq (\varepsilon^{V_n}_0 + B)|\Gamma|^{1/2},$$

where we used the Cauchy-Schwarz inequality for the last part. As a result, we deduce that the sequence $\left(\int_\Gamma u_0^{V_n}\, d\nu_n\right)_{n\in\mathbb{N}^*}$ is bounded. The first convergence of (5.10) follows. The second convergence is a consequence of the first inequality in (5.11).

Let $x_n \in \Gamma = [0, 2\pi)$ be such that $\alpha_n = u_0^{V_n}(x_n)$. The fact that $\alpha_n \to 0$ implies that $l_n := \|u_0^{V_n}(x_n + \cdot) - \alpha_n\|^2_{L^2_{\rm per}} \to 1$ and we can thus define for n large enough

$$v_n := \frac{u_0^{V_n}(x_n + \cdot) - \alpha_n}{\|u_0^{V_n}(x_n + \cdot) - \alpha_n\|_{L^2_{\rm per}}}.$$

It holds that $v_n \in H^1_{\rm per}$, $\|v_n\|_{L^2_{\rm per}} = 1$. Besides, it holds that $v_n(0) = 0$. For $q \in \Gamma^*$, we introduce the function $v_{q,n}$ defined by:

$$\forall x \in \mathbb{R}, \quad v_{q,n}(x) := v_n(x)e^{-iq[x]}, \quad \text{where we set} \quad [x] := x \bmod 2\pi.$$

Thanks to the equality $v_n(0) = 0$, it holds that $v_{q,n} \in H^1_{\rm per}$, and that $\|v_{q,n}\|_{L^2_{\rm per}} = 1$. This function is therefore a valid test function for our min-max principle[1].

From the min-max principle (5.5) and the expression (5.3), we obtain

$$\begin{aligned}
B + \varepsilon_q^{V_n} &\le B + a_q^{V_n}(v_{q,n}, v_{q,n}) \\
&= \int_\Gamma \left|\left(-i\frac{d}{dx} + q\right) v_{q,n}\right|^2 + \int_\Gamma |v_{q,n}|^2\, d\nu_n = \int_\Gamma \left|\frac{dv_n}{dx}\right|^2 + \int_\Gamma |v_n|^2\, d\nu_n \\
&= \frac{1}{l_n}\left(\int_\Gamma \left|\frac{d}{dx}\left(u_0^{V_n}(x_n + \cdot)\right)\right|^2 + \int_\Gamma |u_0^{V_n}(x_n + \cdot) - \alpha_n|^2\, d\nu_n\right) \\
&= \frac{1}{l_n}\left(\int_\Gamma \left|\frac{d}{dx}\left(u_0^{V_n}\right)\right|^2 + \int_\Gamma |u_0^{V_n}|^2\, d\nu_n - 2\alpha_n \int_\Gamma u_0^{V_n}\, d\nu_n + \alpha_n^2 \nu_n(\Gamma)\right) \\
&= \frac{1}{l_n}\left(B + \varepsilon_0^{V_n} - 2\alpha_n \int_\Gamma u_0^{V_n}\, d\nu_n + \alpha_n^2 \nu_n(\Gamma)\right).
\end{aligned}$$

We infer from these inequalities, and from (5.10) that

$$0 \le \max_{q\in\Gamma^*} \left|\varepsilon_q^{V_n} - \varepsilon_0^{V_n}\right| \le \left(B + \varepsilon_0^{V_n}\right)\left(\frac{1}{l_n} - 1\right) + \frac{1}{l_n}\left(-2\alpha_n \int_\Gamma u_0^{V_n}\, d\nu_n + \alpha_n^2 \nu_n(\Gamma)\right) \xrightarrow[n\to+\infty]{} 0.$$

This already proves the convergence (5.9).

To see that $\varepsilon \ge \frac{1}{4} - B$, we write, for $V = V_\nu - B$ with $\nu \in \mathcal{M}^+_{\rm per}$ that

$$\forall q \in [-1/2, 1/2], \quad A_q^V = \left|-i\frac{d}{dx} + q\right|^2 + V_\nu - B \ge \left|-i\frac{d}{dx} + q\right|^2 - B \ge q^2 - B,$$

where we used the fact that the lowest eigenvalue of $\left|-i\frac{d}{dx} + q\right|^2$ is q^2 for $q \in [-1/2, 1/2]$ (this can be seen with the Fourier representation of the operator). As a consequence, for $q = \frac{1}{2}$, we obtain that for all $V \in \mathcal{V}_B$, $\varepsilon^V_{q=1/2} \ge \frac{1}{4} - B$. The result follows.

[1] This construction only works in one dimension. We do not know how to construct similar test functions in higher dimension.

To prove the converse, we exhibit an explicit sequence of measures $(\nu_n)_{n\in\mathbb{N}^*} \subset \mathcal{M}^+_{\rm per}$ such that $\varepsilon_q^{V_{\nu_n}} \to \frac{1}{4}$. The general result will follow by taking sequences of the form $V_n = V_{\nu_n} + \left(\varepsilon - \frac{1}{4}\right) - B$. We denote by δ_x the Dirac mass at $x \in \mathbb{R}$, and consider, for $\lambda > 0$, the measure

$$\nu_\lambda := \lambda \sum_{k\in\mathbb{Z}} \delta_{2\pi k} \in \mathcal{M}^+_{\rm per}. \tag{5.12}$$

From the first part of the Proposition, it is enough to check the convergence for $q = 0$. We are looking for a solution to (we denote by $\omega_\lambda^2 := \varepsilon_0^{V_{\nu_\lambda}} \geq 0$ for simplicity)

$$-u'' + \lambda\delta_0 u(0) = \omega_\lambda^2 u, \quad u \geq 0, \quad u(2\pi) = u(0). \tag{5.13}$$

On $(0, 2\pi)$, u satisfies the elliptic equation $-u'' = \omega_\lambda^2 u$, hence is of the form

$$u(x) = Ce^{i\omega_\lambda x} + De^{-i\omega_\lambda x},$$

for some $C, D \in \mathbb{R}$. The continuity of u at 2π implies $Ce^{2i\pi\omega_\lambda} + De^{-2i\pi\omega_\lambda} = C + D$. Moreover, integrating (5.13) between 0^- and 0^+ leads to the jump of the derivative $-u'(0) + u'(2\pi) + \lambda u(0) = 0$, or

$$i\omega_\lambda\left(D - C\right) + i\omega_\lambda\left(Ce^{2i\pi\omega_\lambda} - De^{-2i\pi\omega_\lambda}\right) + \lambda(C + D) = 0.$$

We deduce that (C, D) is solution to the 2×2 matrix equation

$$\begin{pmatrix} 1 - e^{2i\pi\omega_\lambda} & 1 - e^{-2i\pi\omega_\lambda} \\ -i\omega_\lambda\left(1 - e^{2i\pi\omega_\lambda}\right) + \lambda & i\omega_\lambda\left(1 - e^{-2i\pi\omega_\lambda}\right) + \lambda \end{pmatrix} \begin{pmatrix} C \\ D \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

The determinant of the matrix must therefore vanish, which leads to

$$1 = \cos(2\pi\omega_\lambda) + \frac{\lambda}{2}\frac{\sin(2\pi\omega_\lambda)}{\omega_\lambda}.$$

As $\lambda \to \infty$, one must have $\omega_\lambda \to 1/2$, or equivalently $\varepsilon_0^{V_{\nu_\lambda}} \to 1/4$. The result follows.

5.3.3 Proof of Theorem 5.3

We are now in position to give the proof of Theorem 5.3. Let $b \in \mathcal{T}$ and $B > 1/4 - b^*$ where $b^* := \fint_{\Gamma^*} b(q)\,dq$. Let $V_n = V_{\nu_n} - B \subset \mathcal{V}_B$ be a minimising sequence associated to problem (5.8).

Let us first assume by contradiction that $\nu_n(\Gamma) \to \infty$. Then, according to Proposition 5.4, up to a subsequence (still denoted by n), there exists $\varepsilon \geq \frac{1}{4} - B$ such that $\varepsilon_q^{V_n}$ converges uniformly in $q \in \Gamma^*$ to the constant function ε. Also, from the second part of Proposition 5.4, the fact that $B > \frac{1}{4} - b^*$ and the fact that b^* is the unique minimiser to

$$\inf_{c\in\mathbb{R}} \mathcal{K}_b(c), \tag{5.14}$$

where $\mathcal{K}_b(c) := \int_{[0,1/2]} |b(q) - c|^2\,dq$ for all $c \in \mathbb{R}$, it must hold that $\varepsilon = b^*$.

We now prove that

$$\inf_{V\in\mathcal{V}_B} \mathcal{J}_b(V) \neq \inf_{c\in\mathbb{R}} \mathcal{K}_b(c) = \mathcal{K}_b(b^*).$$

To this aim, we exhibit a potential $W \in \mathcal{V}_B$ such that $\mathcal{J}_b(W) < \mathcal{K}_b(b^*)$. Since b is continuous and increasing on $[0, 1/2]$, there exists a unique $q^* \in (0, 1/2)$ such that $b(q^*) = b^*$. We choose $\delta > 0$ small enough such that $0 < q^* - \delta < q^* + \delta < 1/2$, and set

$$\eta^{\mathrm{ext}} := \int_0^{q^*-\delta} |b(q) - b^*|^2 \, dq + \int_{q^*+\delta}^{1/2} |b(q) - b^*|^2 \, dq \quad \text{and} \quad \eta^{\mathrm{int}} := \int_{q^*-\delta}^{q^*+\delta} |b(q) - b^*|^2 \, dq,$$

so that $\mathcal{K}_b(b^*) = \eta^{\mathrm{ext}} + \eta^{\mathrm{int}}$. Since b is increasing and continuous, it holds that $\eta^{\mathrm{int}} > 0$ and $\eta^{\mathrm{ext}} > 0$, and that $b(q^* - \delta) < b^* < b(q^* + \delta)$.

We now choose a constant $\sigma > 0$ such that

$$0 < \sigma < \min\left\{ \frac{\eta^{\mathrm{int}}}{8\delta}, B + b^* - \frac{1}{4}, b^* - b(q^* - \delta), b(q^* + \delta) - b^* \right\}.$$

Let ν_n be the measure defined in (5.12) for $\lambda = n \in \mathbb{N}$, and let

$$\widetilde{W}_n := V_{\nu_n} + b^* - \frac{1}{4}.$$

Since $\varepsilon_q^{\widetilde{W}_n}$ converges to b^* uniformly in Γ^*, there exists $n_0 \in \mathbb{N}^*$ large enough such that

$$\forall q \in \Gamma^*, \quad \left| \varepsilon_q^{\widetilde{W}_{n_0}} - b^* \right| < \sigma/2.$$

We then define

$$W := \widetilde{W}_{n_0} + b^* - \varepsilon_{q^*}^{\widetilde{W}_{n_0}} = V_{\nu_n} + \left[\left(B + b^* - \frac{1}{4} \right) - \left(\varepsilon_{q^*}^{\widetilde{W}_{n_0}} - b^* \right) \right] - B.$$

Since $\sigma < B + b^* - 1/4$, it holds that $W \in \mathcal{V}_B$. Moreover, it holds that $b^* - \sigma < \varepsilon_q^W < b^* + \sigma$ for all $q \in \Gamma^*$. Finally, for $q = q^*$, we have $\varepsilon_{q^*}^W = b^*$.

Let us evaluate $\mathcal{J}_b(W)$. We get

$$\mathcal{J}_b(W) = \int_0^{q^*-\delta} |b(q) - \varepsilon_q^W|^2 \, dq + \int_{q^*-\delta}^{q^*+\delta} |b(q) - \varepsilon_q^W|^2 \, dq + \int_{q^*+\delta}^{1/2} |b(q) - \varepsilon_q^W|^2 \, dq.$$

For the first part, we notice that for $0 \le q < q^* - \delta$, we have

$$b(q) < b(q^* - \delta) < b^* - \sigma < \varepsilon_q^W < \varepsilon_{q^*}^W = b^*.$$

This yields that

$$\forall\ 0 \le q < q^* - \delta, \quad |b(q) - \varepsilon_q^W| = \varepsilon_q^W - b(q) < b^* - b(q) = |b(q) - b^*|.$$

Integrating this inequality leads to

$$\int_0^{q^*-\delta} |b(q) - \varepsilon_q^W|^2 \, dq < \int_0^{q^*-\delta} |b(q) - b^*|^2 \, dq.$$

Similarly, we obtain that

$$\int_{q^*+\delta}^{1/2} |b(q) - \varepsilon_q^W|^2 \, dq < \int_{q^*+\delta}^{1/2} |b(q) - b^*|^2 \, dq.$$

Lastly, for the middle part, we have

$$\int_{q^*-\delta}^{q^*+\delta} |b(q) - \varepsilon_q^W|^2 \, dq < 2\delta \left[\varepsilon_{q^*+\delta}^W - \varepsilon_{q^*-\delta}^W\right] \leq 4\delta\sigma \leq \frac{\eta^{\text{int}}}{2} < \int_{q^*-\delta}^{q^*+\delta} |b(q) - b^*|^2 \, dq.$$

Combining all these inequalities yields that $\mathcal{J}_b(W) < \mathcal{K}_b(b^*)$. This contradicts the minimising character of the sequence $(V_n)_{n\in\mathbb{N}^*}$.

Hence the sequence $(\nu_n(\Gamma))_{n\in\mathbb{N}^*}$ is bounded. The proof of Theorem 5.3 then follows from Proposition 5.7.

5.4 Numerical tests

In this section, we present some numerical results obtained on different toy inverse band structure problems. We propose an adaptive optimisation algorithm in which the different discretisation parameters are progressively increased. Such an approach, although heuristic, shows a significant gain in computational time on the presented test cases in comparison to a naive optimisation approach.

In Section 5.4.1, we present the discretised version of the inverse band problem for multiple target bands. We present the different optimisation procedures used for this problem (direct and adaptive) in Section 5.4.2. Numerical results on different test cases are given in Section 5.4.3. The reader should keep in mind that although the proof given in the previous section only works for the reconstruction of the first band, it is possible to numerically look for methods that reproduce several bands.

5.4.1 Discretised inverse band structure problem

For $k \in \mathbb{Z}$, we let $e_k(x) := \frac{1}{\sqrt{2\pi}} e^{ikx}$ be the k-th Fourier mode. For $s \in \mathbb{N}^*$, we define by

$$X_s := \text{Span}\left\{e_k, \ k \in \mathbb{Z}, \ |k| \leq s\right\} \tag{5.15}$$

the finite dimensional space of L^2_{per} consisting of the $N_s := 2s+1$ lowest Fourier modes. We denote by $\Pi_{X_s} : L^2_{\text{per}} \to X_s$ the L^2_{per} orthogonal projector onto X_s. In practice, the solutions of the eigenvalue problem (5.4) are approximated using a Galerkin method in X_s. We denote by $\varepsilon_{q,1}^{V,s} \leq \cdots \leq \varepsilon_{q,N_s}^{V,s}$ the eigenvalues (ranked in increasing order, counting multiplicity) of the operator $A_q^{V,s} := \Pi_{X_s} A_q^V \Pi_{X_s}^*$. We also denote by $(u_{q,1}^{V,s}, \cdots, u_{q,N_s}^{V,s})$ an orthonormal basis of X_s composed of eigenvectors associated to these eigenvalues so that

$$\forall 1 \leq j \leq N_s, \quad A_q^{V,s} u_{q,j}^{V,s} = \varepsilon_{q,j}^{V,s} u_{q,j}^{V,s}. \tag{5.16}$$

An equivalent variational formulation of (5.16) is the following:

$$\forall 1 \leq j \leq N_s, \quad \forall v \in X_s, \quad a_q^V\left(u_{q,j}^{V,s}, v\right) = \varepsilon_{q,j}^{V,s} \left\langle u_{q,j}^{V,s}, v \right\rangle_{L^2_{\text{per}}}.$$

As s goes to $+\infty$, it holds that $\varepsilon_{q,m}^{V,s} \underset{s\to+\infty}{\longrightarrow} \varepsilon_{q,m}^V$.

In order to perform the integration in (5.7), we discretise the Brillouin zone. We use a regular grid of size $Q \in \mathbb{N}^*$, and set

$$\Gamma_Q^* := \left\{ -\frac{1}{2} + \frac{j}{Q},\ j \in \{0, \cdots, Q-1\} \right\}.$$

We emphasise that since the maps $q \mapsto \varepsilon_{q,m}$ are analytic and periodic, the discretisation error coming from the integration will be exponentially small with respect to Q. In practice, we fix $Q \in \mathbb{N}^*$.

Let $M \in \mathbb{N}^*$ be a desired number of targeted bands and $b_1, \cdots, b_M \in C^0_{\mathrm{per}}$ be real-valued even functions, and such that b_m is increasing when m is odd and decreasing when m is even. Our cost functional is therefore $\mathcal{J} : H^{-1}_{\mathrm{per},r} \to \mathbb{R}$, defined by

$$\forall V \in H^{-1}_{\mathrm{per,r}}, \quad \mathcal{J}(V) := \frac{1}{Q} \sum_{q \in \Gamma_Q^*} \sum_{m=1}^{M} |b_m(q) - \varepsilon^V_{q,m}|^2.$$

Its discretised version, when the eigenvalues problems are solved with a Galerkin approximation, is

$$\forall s \in \mathbb{N}^*, \quad \forall V \in H^{-1}_{\mathrm{per,r}}, \quad \mathcal{J}^s(V) := \frac{1}{Q} \sum_{q \in \Gamma_Q^*} \sum_{m=1}^{M} |b_m(q) - \varepsilon^{V,s}_{q,m}|^2.$$

Recall that our goal is to find a potential $V \in H^{-1}_{\mathrm{per,r}}$ which minimise the functional $\mathcal{J}^s$. In practice, an element $V \in H^{-1}_{\mathrm{per,r}}$ is approximated with a finite set of Fourier modes. For $p \in \mathbb{N}^*$, we denote by

$$Y_p := \mathrm{Span} \left\{ \sum_{k \in \mathbb{Z},\ |k|^2 \leq p} \widehat{V}_k e_k,\ \forall k \in \mathbb{Z},\ |k| \leq p,\ \overline{\widehat{V}_{-k}} = \widehat{V}_k \right\}. \tag{5.17}$$

Altogether, we want to solve

$$V^{s,p} := \mathrm{argmin}_{V \in Y_p} \mathcal{J}^s(V).$$

5.4.2 Algorithms for optimisation procedures

Naive algorithm

We first present a naive optimisation procedure, using a gradient descent method, where the parameters s and p are fixed beforehand. We tested three different versions of the gradient descent algorithm: steepest descent (**SD**), conjugate gradient with Polak Ribiere formula (**PR**) and quasi Newton with the Broyden-Fletcher-Goldfarb-Shanno formula (**BFGS**). We do not detail here these classical descents and corresponding line search routines for the sake of conciseness and refer the reader to [**?**, NW06].

For all $V \in H^{-1}_{\mathrm{per,r}}$, there exists real-valued coefficients $\left(c_k^V\right)_{k \in \mathbb{N}}$ and $\left(d_k^V\right)_{k \in \mathbb{N}^*}$ such that

$$V(x) = c_0^V + \sum_{k \in \mathbb{N}^*} c_k^V \cos(kx) + d_k^V \sin(kx), \quad \text{and} \quad \sum_{k \in \mathbb{N}^*} (1+|k|^2)^{-1} \left(|c_k^V|^2 + |d_k^V|^2 \right) < +\infty.$$

For all $k \in \mathbb{N}$ (respectively $k \in \mathbb{N}^*$), we can express the derivative $\partial_{c_k^V} \mathcal{J}^s(V)$ (respectively $\partial_{d_k^V} \mathcal{J}^s(V)$) exactly in terms of the Bloch eigenvectors $u_{q,m}^{V,s}$. Indeed, it holds that

$$\partial_{c_k^V} \mathcal{J}^s(V) = \frac{1}{Q} \sum_{q \in \Gamma_Q^*} \sum_{m=1}^{M} 2 \left(\varepsilon_{q,m}^{V,s} - b_m(q) \right) \partial_{c_k^V} \left(\varepsilon_{q,m}^{V,s} \right).$$

On the other hand, from the Hellman-Feynman theorem, it holds that

$$\partial_{c_k^V} \left(\varepsilon_{q,m}^{V,s} \right) = \left\langle u_{q,m}^{V,s}, \partial_{c_k^V} A_q^V, u_{q,m}^{V,s} \right\rangle = \langle u_{q,m}^{V,s}, \cos(k\cdot) u_{q,m}^{V,s} \rangle_{L^2_{\mathrm{per}}}.$$

Similarly, for all $k \in \mathbb{N}^*$,

$$\partial_{d_k^V} \left(\varepsilon_{q,m}^{V,s} \right) = \left\langle u_{q,m}^{V,s}, \partial_{d_k^V} A_q^V, u_{q,m}^{V,s} \right\rangle = \langle u_{q,m}^{V,s}, \sin(k\cdot) u_{q,m}^{V,s} \rangle_{L^2_{\mathrm{per}}}.$$

In the rest of the article, for all $p \in \mathbb{N}^*$, we will denote by $\nabla \mathcal{J}^s(V)|_{Y^p}$ the $2p+1$-dimensional real-valued vector so that

$$\nabla \mathcal{J}^s(V)\big|_{Y^p} = \left(\partial_{d_p^V} \mathcal{J}^s(V), \partial_{d_{p-1}^V} \mathcal{J}^s(V), \cdots, \partial_{d_1^V} \mathcal{J}^s(V), \partial_{c_0^V} \mathcal{J}^s(V), \partial_{c_1^V} \mathcal{J}^s(V), \cdots, \partial_{c_p^V} \mathcal{J}^s(V) \right).$$

In order for the reader to better compare our adaptive algorithm with this naive one, we provide its pseudo-code below (Algorithm 1).

Input:
$p, s \in \mathbb{N}^*$;
$W_0 \in Y_p$: initial guess;
$\varepsilon > 0$: prescribed global precision;
$\nu > 0$: tolerance for the norm of the gradient;

Output:
$W_* \in Y_p$ such that $\|\nabla \mathcal{J}^s(W_*)\big|_{Y_p}\| \leq \nu$;

Instructions:
$n = 0$, $W = W_0$;
while $\|\nabla \mathcal{J}^s(W)\big|_{Y_p}\| > \nu$ **do**
  compute a descent direction $D \in Y_p$ at $\mathcal{J}^s(W)$ (using **SD** / **PR** / **BFGS**);
  choose $t \in \mathbb{R}$ so that $t \in \underset{\bar{t} \in \mathbb{R}}{\operatorname{argmin}} \mathcal{J}^s(W + \bar{t} D)$;
  set $W \leftarrow W + tD$;
end
return $W_* = W$.

Algorithm 1: Naive optimisation algorithm

Although this method gives satisfactory numerical optimisers as shown in Section 5.4.3, its computational time grows very quickly with the discretisation parameters p and s. Besides, it is not clear how these parameters should be chosen a priori, given some target bands. This motivates the design of an adaptive algorithm.

Adaptive algorithm

In order to improve on the efficiency of the numerical optimisation procedure, we propose an adaptive algorithm, where the discretisation parameters s or p are increased during the optimisation process. To describe this procedure, we introduce two criteria to determine whether s or p need to be increased during the algorithm.

As the parameter s is increased, the approximated eigenvalues $\varepsilon_{q,m}^{V,s}$ becomes more accurate, and the discretised cost functional $\mathcal{J}^s$ gets closer to the true one $\mathcal{J}$. Our criterion for s relies on the use of an a posteriori error estimator for the eigenvalue problem (5.16). More precisely, assume we can calculate at low numerical cost an estimator $\Delta_{q,m}^{V,s} \in \mathbb{R}_+$ such that

$$|\varepsilon_{m,q}^{V} - \varepsilon_{m,q}^{V,s}| \leq \Delta_{q,m}^{V,s},$$

(see Appendix 5.5), then we would have that

$$\begin{aligned}
|\mathcal{J}(V) - \mathcal{J}^s(V)| &= \left| \frac{1}{Q} \sum_{q \in \Gamma_Q^*} \sum_{m=1}^{M} \left(|b_m(q) - \varepsilon_{q,m}^{V}|^2 - |b_m(q) - \varepsilon_{q,m}^{V,s}|^2 \right) \right| \\
&= \left| \frac{1}{Q} \sum_{q \in \Gamma_Q^*} \sum_{m=1}^{M} \left(2b_m(q) - \varepsilon_{q,m}^{V} - \varepsilon_{q,m}^{V,s} \right) \left(\varepsilon_{q,m}^{V,s} - \varepsilon_{q,m}^{V} \right) \right| \\
&\leq \frac{1}{Q} \sum_{q \in \Gamma_Q^*} \sum_{m=1}^{M} \left(2 \left| b_m(q) - \varepsilon_{q,m}^{V,s} \right| + \Delta_{q,m}^{V,s} \right) \Delta_{q,m}^{V,s} =: \mathcal{S}_V^s.
\end{aligned}$$

The quantity $\mathcal{S}_V^s$ estimates the error between $\mathcal{J}(V)$ and $\mathcal{J}^s(V)$ and therefore gives information on the necessity to adapt the value of the discretisation parameter s.

We now derive a criterion for the parameter p. When this parameter is increased, the minimisation space Y_p gets larger. A natural way to decide whether or not to increase p is therefore to consider the gradient of $\mathcal{J}^s$, at the current minimisation point $W \in Y_p$, but calculated on a larger subspace $Y_{p'} \supset Y_p$ with $p' > p$.

In practice, the natural choice $p' = p + 1$ is inefficient. This is not a surprise, as there is no reason a priori to expect a sudden change at exactly the next Fourier mode. We therefore took the heuristic choice $p' = 2p$. More specifically, we define

$$\mathcal{P}_V^p := \left\| \nabla_V \mathcal{J}^s(V) \big|_{Y_{2p}} \right\|.$$

Note that this estimator needs to be computed only when V is a local minimum of $\mathcal{J}^s$ on Y_p. When this estimator is larger than some threshold, we increase p so that the new space Y_p contains the Fourier mode which provides the highest contribution in $(\nabla_V \mathcal{J}^s(V)) \big|_{Y_{2p}}$.

The adaptive procedure we propose is described in details in Algorithm 2:

Input:
$p_0, s_0 \in \mathbb{N}^*$: initial discretisation parameters;
$W_0 \in Y_{p_0}$: initial guess;
$\eta > 0$: global discretisation precision;
$\nu > 0$: gradient norm precision;

Output:
$p \geq p_0$, $s \geq s_0$: final discretisation parameters;
$W_* \in Y_p$ such that $\|\nabla \mathcal{J}^s(W_*)\big|_{Y_p}\| \leq \nu$, $\mathcal{S}^s_{W_*} \leq \eta$ and $\mathcal{P}^p_{W_*} \leq \eta$;

Instructions:
$n = 0$, $W = W_0$;
while $\|\nabla \mathcal{J}^s(W)\big|_{Y_p}\| > \nu$ *or* $\mathcal{S}^s_W > \eta$ *or* $\mathcal{P}^p_W > \eta$ **do**
  while $\|\nabla \mathcal{J}^s_p(W)\big|_{Y_p}\| > \nu$ **do**
    compute a descent direction $D \in Y_p$ at $\mathcal{J}^s(W)$ (using **SD** / **PR** / **BFGS**);
    choose $t \in \mathbb{R}$ so that $t \in \operatorname{argmin}_{\bar{t}\in\mathbb{R}} \mathcal{J}^s(W + \bar{t}D)$;
    set $W \leftarrow W + tD$;
  end
  if $\mathcal{S}^s_W > \eta$ **then**
    set $s \leftarrow s + 1$;
  end
  else if $\mathcal{P}^p_W > \eta$ **then**
    set $p \leftarrow \operatorname{argmax}_{p<\bar{p}\leq 2p} \quad \max\left(\left|\partial_{d\frac{V}{\bar{p}}} \mathcal{J}^s(W)\right|, \left|\partial_{c\frac{V}{\bar{p}}} \mathcal{J}^s(W)\right|\right)$;
  end
end
return $W_* = W$.

Algorithm 2: Adaptive optimisation algorithm

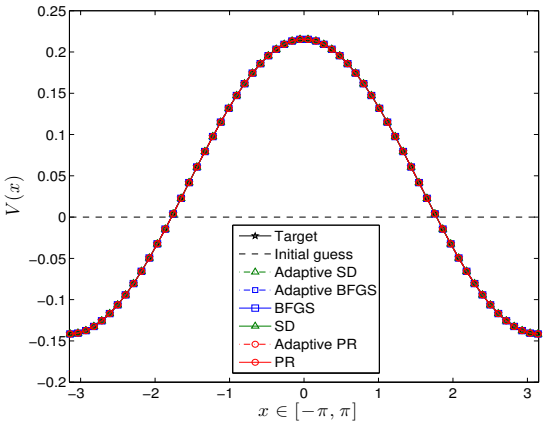
5.4.3 Numerical results

In this section, we illustrate the different algorithms presented above.

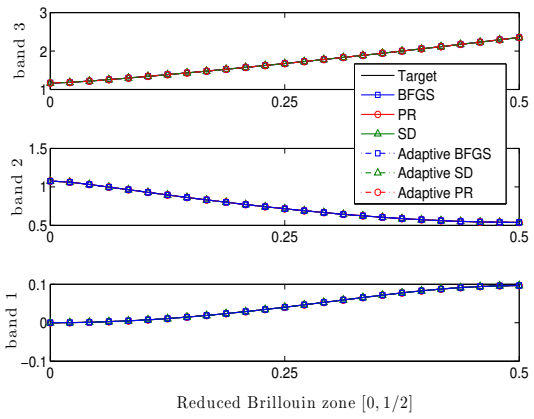
We consider the case where the target functions come from a target potential $V_{\mathrm{t}} \in Y_{p_{\mathrm{t}}}$, whose Fourier coefficients are randomly chosen for some $p_t \in \mathbb{N}^*$. We therefore take $b_m(q) := \varepsilon_{q,m}^{V_{\mathrm{t}},s_{\mathrm{t}}}$, and try to recover the first M functions b_m. The numerical parameters are $M = 3$, $Q = 25$, $\nu = 10^{-5}$, $\eta = 10^{-6}$ and $s_{\mathrm{t}} = 20$. The initial guess is $W_0 = 0$. The naive algorithms are run with $s = s_{\mathrm{t}}$ and $p = p_{\mathrm{t}}$, while the adaptive algorithms start with $s_0 = p_0 = 1$. In addition, the a posteriori estimator is obtained with $s_{\mathrm{ref}} = 250$ and $\theta = 0.01$ (see Appendix 5.5). All tests are done with the naive and adaptive algorihms, with steepest descent (**SD**), conjugate gradient with Polak Ribiere formula (**PR**) and quasi Newton with the Broyden-Fletcher-Goldfarb-Shanno formula (**BFGS**).

In our first test, we try to recover a simple shifted cosine function (i.e. $p_{\mathrm{t}} = 1$). Results are shown in Figure 5.1. We observe that the bands and the potential are well reconstructed. We also notice that the adaptive algorithm takes more iterations to

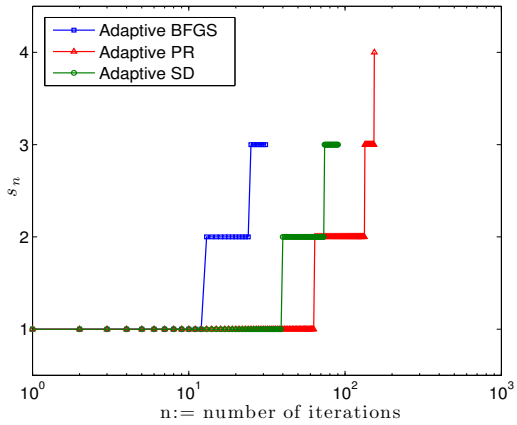
converge. However, as we will see later, most iterations are performed for low values of the parameters s and p, and therefore are usually faster in terms of CPU time (see Table 5.1 below).



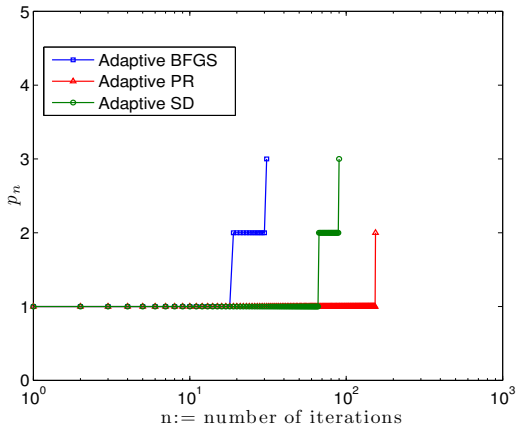
(a) Potentials



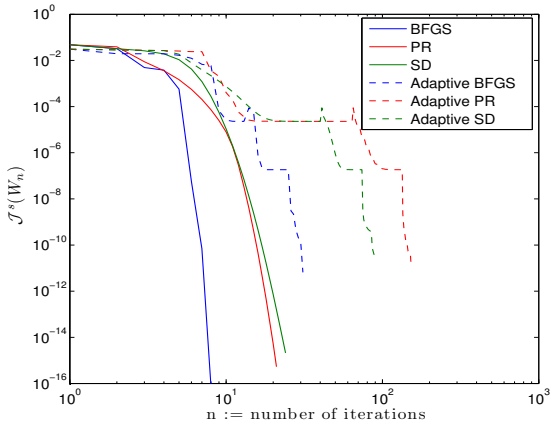
(b) Bands



(c) Evolution of s



(d) Evolution of p

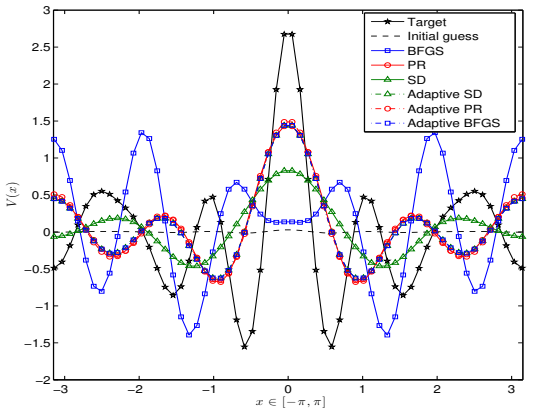


(e) Convergence of the algorithms

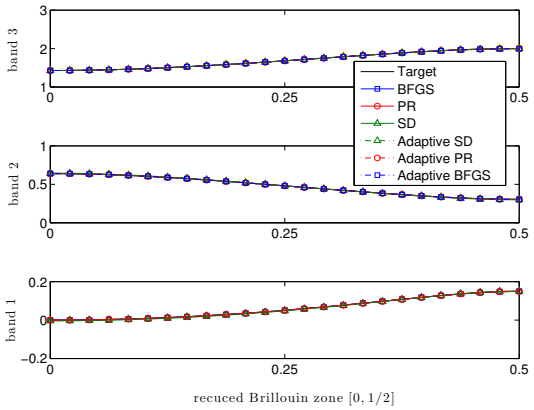
Figure 5.1 – Recovery of the cosine potential.

In the second test case, we try to recover a more complex potential with $p_t = 8$ (see Figure 5.2). In this case, all the algorithms reproduce well the first bands, but fail to

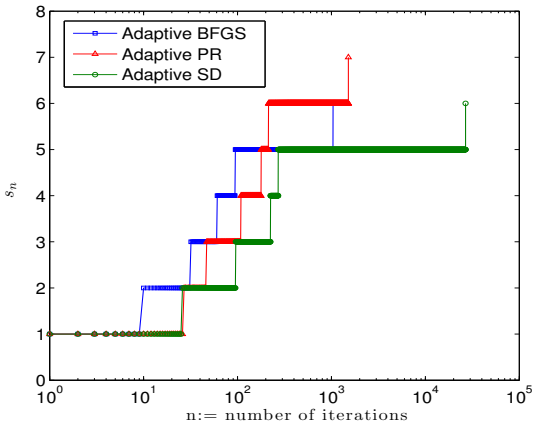
recover the potential. Actually, we see how different methods can lead to different local minima for the functional $\mathcal{J}$. This reflects the complex landscape of this function.



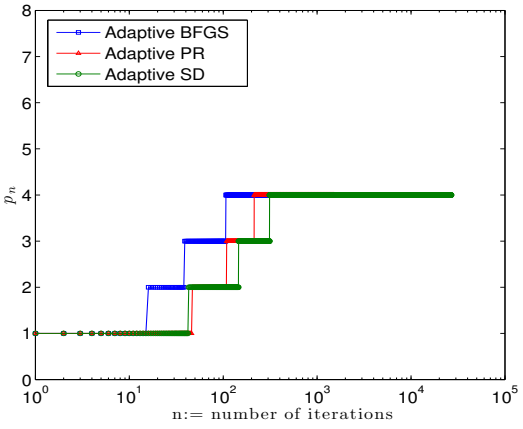
(a) Potentials



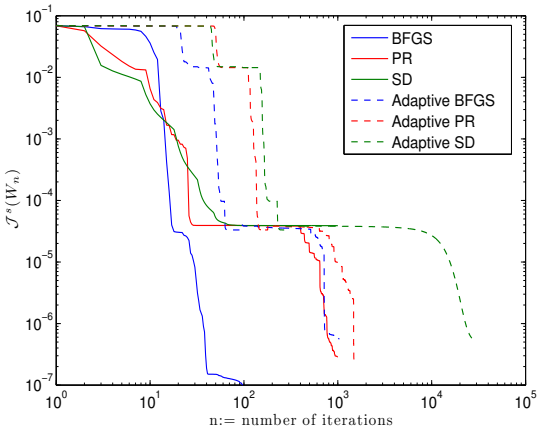
(b) Bands



(c) Evolution of s



(d) Evolution of p



(e) Convergence of the algorithms

Figure 5.2 – Recovery of an oscillating potential.

We end this section by reporting results obtained with the different algorithms, and for different target potential $V_{\mathrm{t}} \in Y_{p_{\mathrm{t}}}$ with $p_{\mathrm{t}} = 1, 4, 8, 12$ (see Table 5.1). In this table, N denotes the number of iterations, s_N and p_N are the values of the parameters s and p at the last iteration (in particular, for the naive algorithms, we have $s_N = s_{\mathrm{t}} = 20$ and $p_N = p_{\mathrm{t}}$). Lastly, for each algorithm **algo**, we define a relative CPU time

$$\tau_{\mathbf{algo}} = \frac{t_{\mathbf{algo}}}{t_{\mathbf{SD}}},$$

where $t_{\mathbf{algo}}$ is the CPU time consumed by the algorithm **algo** and $t_{\mathbf{SD}}$ is the CPU time consumed by the classical steepest descent. In particular, $\tau_{\mathbf{SD}} = 1$.

p_{t}	-	**BFGS**		**PR**		**SD**	
	-	**naive**	**adaptive**	**naive**	**adaptive**	**naive**	**adaptive**
1	τ	0.259	1.176	0.929	1.320	1	1.255
	N	8	31	21	154	24	90
	s_N	20	3	20	4	20	3
	p_N	1	3	1	2	1	3
4	τ	0.070	0.009	0.464	0.281	1	0.259
	N	54	1424	1927	7091	8453	19095
	s_N	20	8	20	7	20	5
	p_N	4	5	4	3	4	3
8	τ	0.470	0.151	1.090	0.144	1	0.519
	N	553	1041	1023	1515	7326	26783
	s_N	20	6	20	7	20	6
	p_N	8	4	8	4	8	4
12	τ	0.007	0.001	0.054	0.004	1	0.044
	N	765	2474	2413	2727	50312	34865
	s_N	20	9	20	9	20	9
	p_n	12	8	12	8	12	8

Table 5.1 – Results for recovery test with different algorithms. Red values are reference values.

We notice that although the adaptive approach requires more iterations to converge, it is usually faster than the naive one. As we already mentioned, this is due to the fact that most of the iterations are performed with small values of p and s, and are therefore faster. Moreover, we notice that the adaptive algorithms tend to find an optimised potential which $p_N \leq p_{\mathrm{t}}$, i.e. a less oscillatory potential than the target one.

Ackowledgements

The authors heartily thank Éric Cancès, Julien Vidal, Damiano Lombardi and Antoine Levitt for their great help in this work and for inspiring discussions. The IRDEP institute is acknowledged for funding.

5.5 Appendix: A posteriori error estimator for the eigenvalue problem

We present in this appendix the a posteriori error estimator for eigenvalue problems that we use in Section 5.4.3. More details about this estimator are given in [BL17].

Let $\mathcal{H}$ be a finite dimensional space of size N_{ref} and let A be a self-adjoint operator on $\mathcal{H}$. In our case, $\mathcal{H}$ is some $X_{s_{\text{ref}}}$ (see definition (5.15)) for some large $s_{\text{ref}} \gg 1$, and $A = A_q^{V,s_{\text{ref}}}$. The eigenvalues of A, counting multiplicities are denoted by $\varepsilon_1 \leq \varepsilon_2 \leq \cdots \leq \varepsilon_{N_{\text{ref}}}$.

For $N \ll N_{\text{ref}}$, we consider X_N a finite dimensional subspace of $\mathcal{H}$. We denote by Π_{X_N} the orthogonal projection on X_N, and by $A^N := \Pi_{X_N} A \Pi_{X_N}^*$. The eigenvalues of A^N are denoted by $\varepsilon_1^N \leq \varepsilon_2^N \leq \cdots \leq \varepsilon_N^N$. Let us also denote by $\left(u_m^N\right)_{1\leq j\leq N}$ a corresponding orthogonal basis of X^N, so that

$$\forall 1 \leq m \leq N, \quad A^N u_m^N = \varepsilon_m^N u_m^N.$$

We recall that, from the min-max principle, it holds that $\varepsilon_m \leq \varepsilon_m^N$. A certified a posteriori error estimator for the m-th eigenvalue is a non-negative real number $\Delta_m^N \in \mathbb{R}_+$ such that

$$\varepsilon_m^N - \varepsilon_m \leq \Delta_m^N.$$

We also require that the expression of Δ_m^N only involves the approximated eigenpair ε_m^N and u_m^N (and not ε_m).

Proposition 5.10. *Assume that ε_m (resp. ε_m^N) is a non-degenerate eigenvalue of A (resp. A^N), and that*

$$0 < \varepsilon_m^N - \varepsilon_m < \operatorname{dist}\left(\varepsilon_m^N, \sigma(A) \setminus \{\varepsilon_m\}\right). \tag{5.18}$$

Let $\lambda_m < \varepsilon_m$. Then there exists $\delta_m > 0$ such that, for all $0 \leq \delta < \delta_m$, we have

$$\varepsilon_m^N - \varepsilon_m \leq \left\langle r_m^N, (A - c_\delta)^{-1} (A - d_\delta) (A - c_\delta)^{-1} r_m^N \right\rangle, \tag{5.19}$$

where we set $c_\delta := \varepsilon_m^N + \delta$, $d_\delta := \lambda_m + \delta$, and where $r_m^N := \left(A - \varepsilon_m^N\right) u_m^N$ is the residual.

Proof. Assumption (5.18) implies that $\varepsilon_m^N \notin \sigma(A)$, so that $\left(A - \varepsilon_m^N\right)$ is invertible. From the fact that $\langle u_m^N, A u_m^N \rangle = \varepsilon_m^N$, and the definition of the residual, it holds that

$$\varepsilon_m^N - \varepsilon_m = \left\langle r_m^N, \left(A - \varepsilon_m^N\right)^{-1} (A - \varepsilon_m) \left(A - \varepsilon_m^N\right)^{-1} r_m^N \right\rangle. \tag{5.20}$$

Thus, a sufficient condition for (5.19) to hold is that

$$(A - c_\delta)^{-1} (A - d_\delta) (A - c_\delta)^{-1} \geq \left(A - \varepsilon_m^N\right)^{-1} (A - \varepsilon_m) \left(A - \varepsilon_m^N\right)^{-1}.$$

Thanks to the spectral decomposition of A, this is the case if and only if,

$$\forall 1 \leq \widetilde{m} \leq N_{\text{ref}}, \quad \frac{\varepsilon_{\widetilde{m}} - d_\delta}{\left(\varepsilon_{\widetilde{m}} - c_\delta\right)^2} \geq \frac{\varepsilon_{\widetilde{m}} - \varepsilon_m}{\left(\varepsilon_{\widetilde{m}} - \varepsilon_m^N\right)^2}.$$

Denoting by $\eta := \operatorname{dist}\left(\varepsilon_m^N, \sigma(A) \setminus \{\varepsilon_m\}\right) - \left(\varepsilon_m^N - \varepsilon_m\right)$, this holds true as soon as $\delta \leq \delta_m := \min\left(\varepsilon_m - \lambda_m, \eta\right)$. The result follows. □

In order to use the left-side of (5.19) as an a posteriori estimator, we need to choose $\lambda_m < \varepsilon_m$ and $\delta_m > 0$. For the choice of λ_m, we follow [WS80], and notice that

$$\varepsilon_m \geq \lambda_m := \mu - \left(\frac{N_{\mathrm{ref}} - m - 1}{m+1}\right)^{1/2} \sigma,$$

where we set

$$\mu := \frac{1}{N_{\mathrm{ref}}} \operatorname{Tr} A \quad \text{and} \quad \sigma^2 := \frac{1}{N_{\mathrm{ref}}} \operatorname{Tr} A^2 - \mu^2.$$

For the choice of δ_m, we chose the simple rule

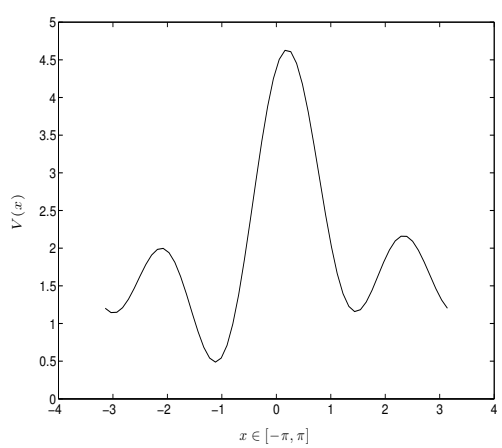
$$\delta_m = \theta \left(\varepsilon_m^N - \kappa\right) \quad \text{with} \quad 0 < \theta \ll 1 \quad \text{and } \kappa \in \mathbb{R} \quad \text{independent of } m.$$

The real number κ is chosen to be an a priori lower bound of the lowest eigenvalue ε_1 of A. This choice is heuristic in the sense that we cannot guarantee that the assumptions of Proposition 5.10 are satisfied. However, the encouraging numerical results we obtain below motivated our choice to use such an estimator (see Section 5.5).

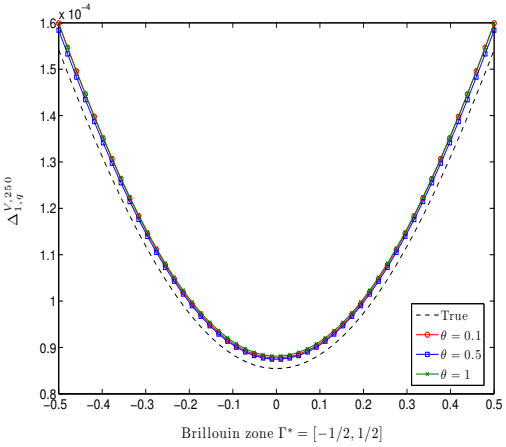
Numerical test To illustrate the efficiency of our heuristic, we tested it to compute the first bands of the Hill's operator A^V with

$$V(x) = \sum_{k=-3}^{3} \hat{V}_k e_k, \quad \text{where} \quad \hat{V}_0 = 2 \quad \text{and} \quad \overline{\hat{V}_{-1}} = \overline{\hat{V}_{-2}} = \hat{V}_1 = \hat{V}_2 = 1 + 0.5\, i.$$

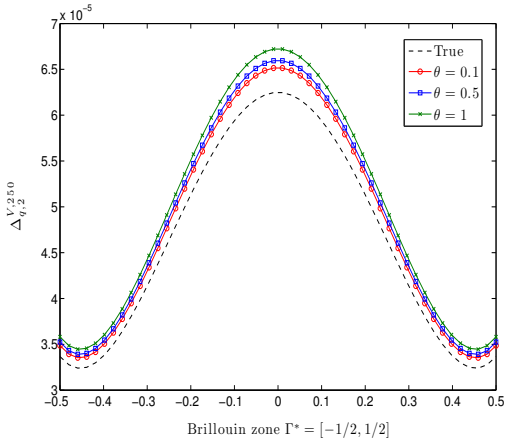
The reference operator is $A := A_q^{V,s_{\mathrm{ref}}}$ with $s_{\mathrm{ref}} = 250$, and the first three bands are computed on the space X^s defined in 5.15 with $s = 6$. We plot in Figure 5.3 the true error $\varepsilon_{q,m}^{V,s} - \varepsilon_{q,m}^{V,s_{\mathrm{ref}}}$ for $m = 1, 2, 3$, and the corresponding a posteriori error with $\kappa = 0$ and different values of θ (namely $\theta = 0.1, 0.5, 1$). We observe that our estimator is sharp for a large range of θ.



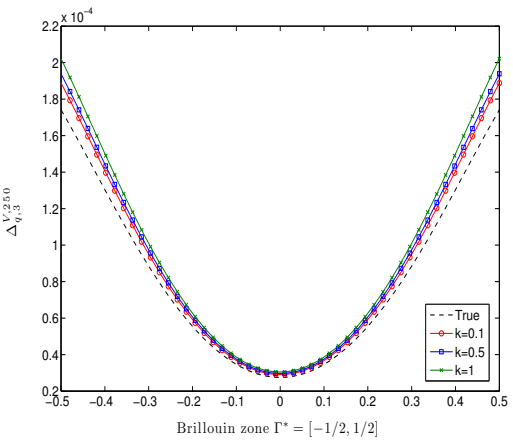
(a) Potential V



(b) $m = 1$



(c) $m = 2$



(d) $m = 3$

Figure 5.3 – Numerical validation of the a posteriori error estimator proposed in Appendix 5.5.

CHAPTER 6

COMPRESSION OF WANNIER FUNCTIONS INTO GAUSSIAN-TYPE ORBITALS

The work presented in this chapter is done in collaboration with Eric Cancès, Paul Cazeaux, Shiang Fang and Efthimios Kaxiras. It is part of the article [BCC+17]

Abstract. We propose a greedy algorithm for the compression of Wannier functions into Gaussian-polynomials orbitals. The so-obtained compressed Wannier functions can be stored in a very compact form, and can be used to efficiently parameterize effective tight-binding Hamiltonians for multilayer 2D materials for instance. The compression method preserves the symmetries (if any) of the original Wannier function. Algorithmic details are provided, and the performance of our implementation are illustrated on several examples (graphene, hexagonal boron-nitride, single-layer FeSe, diamond-phase silicon)

Contents

6.1 Introduction

Since their introduction in 1937 [Wan37], Wannier functions have become a widely used tool in solid state physics and materials science. Theses functions provide insights on

chemical bonding in crystalline material [MMY+12], they play an essential role in the modern theory of polarization [KSV93], and they can be used to parametrized tight-binding Hamiltonians for the calculation of electronic properties [FKDS+15]. Other applications of Wannier functions are presented in the review paper [MMY+12].

Maximally localized Wannier functions (MLWFs) were introduced by Marzari and Vanderbilt [MV97] and are obtained by minimizing some spread functional [MV97, SMV01, MMY+12]. Several algorithms for generating MLWFs are implemented in the Wannier90 computer program [MYL+08]. In the general case, MLWFs obtained by the standard Marzari-Vanderbilt procedure are not centered at high-symmetry points of the crystal (typically atoms or centers of chemical bonds), and do not fulfill any symmetry properties [SMV01, THJ05], which complicates their physical interpretation. Symmetry-adapted Wannier functions (SAWFs) are Wannier functions centered at high-symmetry points and are associated with irreducible representations of a non-trivial subgroup of the space group of the crystal (precise definitions are given in Appendix). They can be seen as the solid-state counterparts of symmetry-adapted molecular orbitals [Lad16] fruitfully used in quantum chemistry. SAWFs were investigated in [DC63, Koh73, VBC79, Krü87, SB94, ES12, SU01, SE05, PBMM02, CZWP06] from both theoretical and numerical points of view. An algorithm for generating maximally-localized SAWFs was recently proposed by Sakuma [Sak13]; it allows one to enforce the center and symmetries of the Wannier functions during the spread minimization procedure. It is now implemented in the Wannier90 computer program.

In this work, we propose a numerical method for compressing Wannier functions into a finite sum of Gaussian-polynomial functions (referred to as Gaussian-type orbitals - GTOs - in the sequel), which preserves the centers and the possible symmetries of the original Wannier functions. Such compressed representations enable the characterization of a Wannier function by a small number of parameters (the shape parameters of the Gaussians and the polynomial coefficients) rather than by its values on a potentially very large grid. In addition, they can be used to accelerate the parameterization of tight-binding Hamiltonians or more advanced reduced models from Wannier functions computed from Density Functional Theory. Indeed, matrix elements of effective Hamiltonians can be computed very efficiently for GTOs; this fundamental remark by Boys [Boy50] was instrumental for the development of numerical methods for quantum chemistry. Gaussian-type approximate Wannier functions should be particularly useful for simulating multilayer two-dimensional materials [JM13, FK16], especially when Fock exchange terms are considered, which is the case for hybrid functionals.

This article is organized as follows. In Section 6.2, we describe our approach for compressing a given symmetry-adapted Wannier function W into a finite sum of GTOs $\widetilde{W}_p$ sharing the same center and symmetries as W. Note that our procedure is also valid if the Wannier function has no symmetry (in the case, the symmetry group is reduced to the identity matrix). The main idea is to construct a sequence $\widetilde{W}_0, \widetilde{W}_1, \widetilde{W}_2, \cdots$ of better and better approximations of W (for the relevant metric, see Section 6.2.1), by means of an orthogonal greedy algorithm [Tem08, TZ11]. The basics of greedy algorithms and symmetry-adapted Wannier functions are briefly recalled in Sections 6.2.2 and 6.2.3 respectively. An overall description of our algorithm is given in Section 6.2.4 and implementation details are provided in Section 6.2.5. We strongly believe that greedy methods are very well adapted to the compressing problem under consideration; on the

other hand, we do not claim that our implementation is optimal: many variants of the numerical scheme described in Section 6.2.5 can be considered, and there is clearly room for improvement to reduce the number of GTOs necessary to reach a given accuracy. The purpose of this contribution is to assess the efficiency of greedy methods in this setting, and to stimulate further work in this direction. The performance of our current implementation is illustrated in Section 6.3 on four examples: three two-dimensional materials, namely graphene, hexagonal boron-nitride (hBN), and FeSe, and bulk silicon (in the diamond phase).

6.2 Theory

6.2.1 Error control

Consider a real-valued Wannier function $W : \mathbb{R}^3 \to \mathbb{R}$, which we would like to approximate by a finite sum of well-chosen Gaussian-polynomial functions. First, we have to specify the norm with which the error between W and its approximation $\widetilde{W}$ will be measured. We will consider here the L^2 and H^1 norms respectively defined by

$$\|u\|_{L^2} = \left(\int_{\mathbb{R}^3} |u(\mathbf{r})|^2 \, d\mathbf{r} \right)^{1/2}$$

and

$$\|u\|_{H^1} = \left(\int_{\mathbb{R}^3} |u(\mathbf{r})|^2 \, d\mathbf{r} + \int_{\mathbb{R}^3} |\nabla u(\mathbf{r})|^2 \, d\mathbf{r} \right)^{1/2}. \tag{6.1}$$

Requesting that $\|W - \widetilde{W}\|_{H^1}$ is small is far more demanding than simply requesting that $\|W - \widetilde{W}\|_{L^2}$ is small, since in the former case, both $\|W - \widetilde{W}\|_{L^2}$ and $\|\nabla W - \nabla\widetilde{W}\|_{L^2}$ must be small. In the perspective of using approximate Wannier functions to calibrate tight-binding models, it is important to request $\|W - \widetilde{W}\|_{H^1}$ to be small. Indeed, while the errors on the overlap integrals can be controlled by L^2-norms:

$$\left| \int_{\mathbb{R}^3} W_i(\mathbf{r}) W_j(\mathbf{r}) \, d\mathbf{r} - \int_{\mathbb{R}^3} \widetilde{W}_i(\mathbf{r}) \widetilde{W}_j(\mathbf{r}) \, d\mathbf{r} \right| \leq \|W_i\|_{L^2} \|W_i - \widetilde{W}_i\|_{L^2} + \|\widetilde{W}_i\|_{L^2} \|W_j - \widetilde{W}_j\|_{L^2},$$

the errors on the kinetic energy integrals appearing in effective one-body Hamiltonians matrix elements

$$\langle W_i | H | W_j \rangle = \frac{1}{2} \int_{\mathbb{R}^3} \nabla W_i(\mathbf{r}) \cdot \nabla W_j(\mathbf{r}) \, d\mathbf{r} + \int_{\mathbb{R}^3} \mathcal{V}(\mathbf{r}) W_i(\mathbf{r}) W_j(\mathbf{r}) \, d\mathbf{r}$$

are controlled by the L^2-norms of the gradients, hence by the H^1-norms of the functions. The H^1-norm also allows one to control the errors on the potential integrals, even in presence of Coulomb singularities.

Note that the L^2 and H^1-norms are particular instances of the Sobolev norms H^s, $s \in \mathbb{R}$, defined on the Solobev spaces

$$H^s(\mathbb{R}^3) = \left\{ u : \mathbb{R}^3 \to \mathbb{R} \text{ s.t. } \int_{\mathbb{R}^3} (1 + |\mathbf{k}|^2)^s |\widehat{u}(\mathbf{k})|^2 \, d\mathbf{k} < \infty \right\},$$

where $\widehat{u}$ is the Fourier transform of u, by

$$\|u\|_{H^s} := \left(\int_{\mathbb{R}^3} (1 + |\mathbf{k}|^2)^s |\widehat{u}(\mathbf{k})|^2 \, d\mathbf{k} \right)^{1/2}. \tag{6.2}$$

The L^2-norm corresponds to $s = 0$, due to the isometry property of the Fourier transform:

$$\int_{\mathbb{R}^3} |\widehat{u}(\mathbf{k})|^2 \, d\mathbf{k} = \int_{\mathbb{R}^3} |u(\mathbf{r})|^2 \, d\mathbf{r}.$$

Likewise, definition (6.2) agrees with definition(6.1) for $s = 1$ since

$$\int_{\mathbb{R}^3} |\mathbf{k}|^2 |\widehat{u}(\mathbf{k})|^2 \, d\mathbf{k} = \int_{\mathbb{R}^3} |i\mathbf{k}\widehat{u}(\mathbf{k})|^2 \, d\mathbf{k} = \int_{\mathbb{R}^3} \left|\widehat{\nabla u}(\mathbf{k})\right|^2 \, d\mathbf{k} = \int_{\mathbb{R}^3} |\nabla u(\mathbf{r})|^2 \, d\mathbf{r}.$$

It can be useful to consider other kinds of Sobolev norms in some particular applications. For instance, $\|W - \widetilde{W}\|_{H^1}$ being small does not guarantee that the pointwise values of the function $(W - \widetilde{W})$ are small. On the other hand, if $\|W - \widetilde{W}\|_{H^2}$ is small, then $|W(\mathbf{r}) - \widetilde{W}(\mathbf{r})|$ is small for each $\mathbf{r} \in \mathbb{R}^3$.

Our greedy algorithm has been implemented in the Fourier representation, and can therefore minimize the error between the Wannier function W and its GTO representation for any value of the Sobolev exponent s. In the numerical examples reported in Section 6.3, we will consider the cases $s = 0$ and $s = 1$.

6.2.2 Greedy algorithms in a nutshell

Greedy algorithms [Tem08, TZ11] are iterative algorithms allowing one, among other things, to construct sequences of approximations $\widetilde{W}_0$, $\widetilde{W}_1$, $\widetilde{W}_2$, ... of some target function $W \in H^s(\mathbb{R}^3)$, with the following properties:

- each approximate function $\widetilde{W}_p$ is a sum of p "simple" functions belonging to some prescribed *dictionary* $\mathcal{D} \subset H^s(\mathbb{R}^3)$:

$$\widetilde{W}_p(\mathbf{r}) = \sum_{j=1}^{p} \phi_j^{(p)}(\mathbf{r}),$$

  with $\phi_j^{(p)} \in \mathcal{D}$. In our case, $\mathcal{D}$ will be a set of symmetry-adapted Gaussian-polynomial functions;

- the errors $\|W - \widetilde{W}_p\|_{H^s}$ decay to 0 when $p \to \infty$.

Greedy algorithms therefore provide systematic ways to approximate a given function $W \in H^s(\mathbb{R}^3)$ by a finite sum of simple functions with an arbitrary accuracy. Of course, the set $\mathcal{D}$ of elementary functions cannot be any subset $H^s(\mathbb{R}^3)$ (for instance $\mathcal{D}$ cannot be chosen as the set of radial functions since only radial functions can be well approximated by finite sums of radial functions). The convergence property $\|W - \widetilde{W}_p\|_{H^s} \to 0$ is guaranteed provided the set $\mathcal{D}$ is a *dictionary* of $H^s(\mathbb{R}^3)$, that is a family of functions $H^s(\mathbb{R}^3)$ satisfying the following three conditions:

1. $\mathcal{D}$ is a cone, that is if $\phi \in \mathcal{D}$, then $t\phi \in \mathcal{D}$ for any $t \in \mathbb{R}$;

2. Span$(\mathcal{D})$ is dense in the Sobolev space $H^s(\mathbb{R}^3)$. This means that any function $W \in H^s(\mathbb{R}^3)$ can be approximated with an arbitrary accuracy $\epsilon > 0$ by a finite linear combination of functions of $\mathcal{D}$, and therefore by a finite sum of functions

of $\mathcal{D}$ since $\mathcal{D}$ is a cone: for any $\epsilon > 0$, there exists a finite integer $p \in \mathbb{N}^*$, and p functions $\phi_1^{(p)}$, ... $\phi_p^{(p)}$ in $\mathcal{D}$ such that

$$\left\| W - \left(\sum_{j=1}^{p} \phi_j^{(p)} \right) \right\|_{H^s} \leq \epsilon.$$

Greedy algorithms provide practical ways to construct such approximations;

3. $\mathcal{D}$ is weakly closed in $H^s(\mathbb{R}^3)$. This technical assumption ensures the convergence of the greedy algorithm [Tem08].

Given a dictionary $\mathcal{D}$, the greedy method then consists in

- initializing the algorithm with (for instance) $\widetilde{W}_0 = 0$;
- constructing iteratively a sequence $\widetilde{W}_1, \widetilde{W}_2, \widetilde{W}_3, \cdots$ of more and more accurate approximations of the target Wannier function W of the forms

$$\widetilde{W}_p(\mathbf{r}) = \sum_{j=1}^{p} \phi_j^{(p)}(\mathbf{r}), \tag{6.3}$$

  where $\phi_j^{(p)}$ are functions of the dictionary $\mathcal{D}$;
- stopping the iterative process when $\|W - \widetilde{W}_p\|_{H^s} \leq \epsilon$, where $\epsilon > 0$ is the desired accuracy (for the chosen H^s-norm).

We will use here the so-called orthogonal greedy algorithm for constructing $\widetilde{W}_{p+1}$ from $\widetilde{W}_p$, which is defined as follows.

Algorithm 6.1 (Orthogonal greedy algorithm)**.**

Step 1: *Compute the residual at iteration p:*

$$R_p(\mathbf{r}) = W(\mathbf{r}) - \widetilde{W}_p(\mathbf{r});$$

Step 2: *find a local minimizer* ϕ_{p+1} *to the optimization problem*

$$\min_{\phi \in \mathcal{D}} J_p(\phi), \quad \textit{where} \quad J_p(\phi) := \|R_p - \phi\|_{H^s}^2; \tag{6.4}$$

Step 3: *solve the unconstrained quadratic optimization problem*

$$(c_j^{(p+1)})_{1 \leq j \leq p+1} \in \operatorname{argmin} \left\{ \left\| W - \left(\sum_{j=1}^{p+1} c_j \phi_j^{(p)} + c_{p+1} \phi_{p+1} \right) \right\|_{H^s}^2, \ (c_j)_{1 \leq j \leq p+1} \in \mathbb{R}^{p+1} \right\}; \tag{6.5}$$

Step 4: *set* $\phi_j^{(p+1)} = c_j^{(p+1)} \phi_j^{(p)}$, $1 \leq j \leq p$, *and* $\phi_{p+1}^{(p+1)} = c_{p+1}^{(p+1)} \phi_{p+1}$.

Note that Step 3 is easy to perform since (6.5) is nothing but a least square problem in dimension $(p+1)$ (p is of the order of 10 to 10^3 in practice). Step 2 will be described in detail in Sections 6.2.4 and 6.2.5. The next section is concerned with the choice of the dictionary $\mathcal{D}$.

6.2.3 Symmetry-adapted Wannier functions and Gaussian-type orbitals

For the reader's convenience, the basics of the theory of symmetry-adapted Wannier functions we make use of in this section are recalled in Appendix.

We assume from now on that we are dealing with a periodic material with space group $G = \mathcal{R} \rtimes G_{\rm p}$, where $\mathcal{R}$ is a Bravais lattice embedded in $\mathbb{R}^3$, and $G_{\rm p}$ a finite point group (a finite subgroup of the orthogonal group $O(3)$). The Bravais lattice $\mathcal{R}$ is two-dimensional for 2D materials such as graphene or hBN, and three-dimensional for usual 3D crystals.

We also assume that we are given a symmetry-adapted Wannier function W centered at a high-symmetry point $\mathbf{q} \in \mathbb{R}^3$ of the crystalline lattice, and corresponding to a one-dimensional representation of the subgroup

$$G^0_{\mathbf{q}} := \{\Theta \in G_{\rm p} \mid \Theta\mathbf{q} \in \mathbf{q} + \mathcal{R}\}$$

of $G_{\rm p}$. Note that our method can straightforwardly be extended to the case of two-dimensional irreducible representations of $G^0_{\mathbf{q}}$. We now translate the origin of the Cartesian frame to point $\mathbf{q}$. Setting $G^0 := G^0_{\mathbf{q}}$ to simplify the notation, the function W satisfies in this new frame the invariance property

$$\forall \Theta \in G^0, \quad (\Theta W)(\mathbf{r}) = W(\Theta^{-1}\mathbf{r}) = \chi(\Theta) W(\mathbf{r}), \tag{6.6}$$

where χ is the character of this one-dimensional representation.

Our goal is to approximate the Wannier function W by a finite sum of GTOs. In order to reduce the number of GTOs necessary to obtain the desired accuracy, while enforcing the symmetries of the approximate Wannier functions $\widetilde{W}_p$, we use a dictionary consisting of symmetry-adapted Gaussian-type orbitals (SAGTOs) of the form

$$\phi^{\rm SA}_{\boldsymbol{\alpha},\sigma,\Lambda}(\mathbf{r}) = \frac{1}{|G^0|} \sum_{\Theta \in G^0} \chi(\Theta)\, (\Theta \varphi_{\boldsymbol{\alpha},\sigma,\Lambda})(\mathbf{r}) = \frac{1}{|G^0|} \sum_{\Theta \in G^0} \chi(\Theta)\, \varphi_{\boldsymbol{\alpha},\sigma,\Lambda}(\Theta^{-1}\mathbf{r}), \tag{6.7}$$

where $|G^0|$ is the order of the group G^0, and where

$$\varphi_{\boldsymbol{\alpha},\sigma,\Lambda}(\mathbf{r}) = \left(\sum_{(n_x,n_y,n_z) \in \mathcal{I}} \lambda_{n_x,n_y,n_z} (r_x - \alpha_x)^{n_x} (r_y - \alpha_y)^{n_y} (r_z - \alpha_z)^{n_z} \right) \exp\left(-\frac{1}{2\sigma^2} |\mathbf{r} - \boldsymbol{\alpha}|^2 \right)$$

is a Gaussian-polynomial function centered at $\boldsymbol{\alpha} \in \mathbb{R}^3$ with standard deviation $\sigma > 0$. The set $\mathcal{I}$ is a carefully chosen subset of $\big\{(n_x, n_y, n_z) \in \mathbb{N}^3 \mid n_x + n_y + n_z \leq L\big\}$ (total degree lower or equal to L) determined by the symmetries of the SAWF. Note that for 2D materials laying in the xy plane, it is more appropriate to chose $\mathcal{I} \subset \big\{(n_x, n_y, n_z) \in \mathbb{N}^3 \mid n_x + n_y \leq L_\parallel,\ n_z \leq L_\perp\big\}$.

Any function $\phi^{\rm SA}_{\boldsymbol{\alpha},\sigma,\Lambda}$ of the dictionary thus satisfies the same symmetry property

$$\forall \Theta \in G^0, \quad (\Theta \phi^{\rm SA}_{\boldsymbol{\alpha},\sigma,\Lambda})(\mathbf{r}) = \phi^{\rm SA}_{\boldsymbol{\alpha},\sigma,\Lambda}(\Theta^{-1}\mathbf{r}) = \chi(\Theta) \phi^{\rm SA}_{\boldsymbol{\alpha},\sigma,\Lambda}(\mathbf{r})$$

as the Wannier function W to be approximated.

6.2.4 A greedy algorithm for compressing SAWF into SAGTO

It can be shown that the set

$$\mathcal{D}^{\mathrm{SA}} := \left\{ \phi^{\mathrm{SA}}_{\boldsymbol{\alpha},\sigma,\Lambda},\ \boldsymbol{\alpha} \in \mathbb{R}^3,\ \sigma \in [\sigma_{\min}, \sigma_{\max}],\ \Lambda \in \mathbb{R}^{\mathcal{I}_{\boldsymbol{\alpha}}} \right\}, \tag{6.8}$$

where $0 < \sigma_{\min} < \sigma_{\max} < \infty$ are given parameters (chosen by the user), and $\mathcal{I}_{\boldsymbol{\alpha}}$ is a carefully chosen nonempty subset of $\mathbb{N}^3$ depending on the center $\boldsymbol{\alpha}$ of the SAGTO, is a dictionary for the closed subspace

$$H^{s,\mathrm{SA}}(\mathbb{R}^3) := \left\{ f \in H^s(\mathbb{R}^3) \mid \forall \Theta \in G^0,\ (\Theta f)(\mathbf{r}) = f(\Theta^{-1}\mathbf{r}) = \chi(\Theta) f(\mathbf{r}) \right\}$$

of $H^s(\mathbb{R}^3)$ for any $s \in \mathbb{R}_+$.

For example, in the case of Graphene and hBN (see Section 6.3), we use the same set for each $\boldsymbol{\alpha} \in \mathbb{R}^3$:

$$\mathcal{I}_{\boldsymbol{\alpha}} = \{(0,0,1), (0,0,3), (0,0,5)\}, \quad \forall \boldsymbol{\alpha} \in \mathbb{R}^3.$$

More refine strategies will be considered in future works.

The main practical difficulty in Algorithm 6.1 is the computation of a local minimum to Problem (6.4). This problem can be formulated in our case as

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^3,\ \sigma\in[\sigma_{\min},\sigma_{\max}],\ \Lambda\in\mathbb{R}^{\mathcal{I}}} \mathcal{J}_p(\boldsymbol{\alpha},\sigma,\Lambda), \quad \text{where} \quad \mathcal{J}_p(\boldsymbol{\alpha},\sigma,\Lambda) := \|R_p - \phi_{\boldsymbol{\alpha},\sigma,\Lambda}\|^2_{H^s}. \tag{6.9}$$

The above minimization problem can in turn be written as:

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^3,\ \sigma\in[\sigma_{\min},\sigma_{\max}]} \widetilde{\mathcal{J}}_p(\boldsymbol{\alpha},\sigma), \tag{6.10}$$

where

$$\widetilde{\mathcal{J}}_p(\boldsymbol{\alpha},\sigma) = \min_{\Lambda\in\mathbb{R}^{\mathcal{I}}} \mathcal{J}_p(\boldsymbol{\alpha},\sigma,\Lambda). \tag{6.11}$$

Since the map $\Lambda \mapsto \mathcal{J}_p(\boldsymbol{\alpha},\sigma,\Lambda)$ is quadratic in Λ, problem (6.11) can be solved explicitly at a very low computational cost, and the gradient of $\widetilde{\mathcal{J}}_p(\boldsymbol{\alpha},\sigma)$ with respect to both $\boldsymbol{\alpha}$ and σ can be easily computed from the solution of problem (6.11) by the chain rule.

We can then use an off-the-shelf constrained optimization solver to find a local minimizer to the four-dimensional optimization problem (6.10).

6.2.5 Algorithmic details

Construction of MLWFs

The Wannier functions considered in this work are MLWFs constructed using VASP [KF96a, KF96b] and WANNIER90 [MYL+08]. Let us briefly describe the construction procedure.

First, the Bloch energy bands and wave-functions of the periodic Kohn-Sham Hamiltonian are obtained using VASP with pseudo-potentials of the Projector Augmented Wave (PAW) type [Blö94], the PBE exchange-correlation functional [PBE96], a plane-wave energy cutoff E_{c} and a grid $\mathcal{Q}$ of the Brillouin zone Γ^*. For 2D materials, the height η of the supercell is chosen sufficiently large to eliminate the spurious interactions between the material and its periodic images.

Next, the Bloch eigenfunctions belonging to the energy bands of interest are combined into a basis of MLWFs using the Marzari-Vanderbilt algorithm [MV97] as implemented in the Wannier90 computer program [MYL+08]. The final output is a set of Wannier functions which are known to be localized at a certain point and exponentially decaying for materials which suitable topological properties such as the ones considered in Section 6.3 (see [PP13]). In practice, one chooses a sufficiently large rectangular box,

$$\Omega := [x_{\min}, x_{\max}] \times [y_{\min}, y_{\max}] \times [z_{\min}, z_{\max}] \subset \mathbb{R}^3,$$

such that we can safely neglect the exponentially vanishing values of the Wannier function under consideration outside the box. The numerical values of the Wannier function W are given on a Cartesian grid $\mathcal{M}$ spanning the box and containing $M = M_x M_y M_z$ points.

Note that the Wannier functions obtained in this manner are in general not perfectly symmetry-adapted. Indeed, the classical Marzari-Vanderbilt algorithm does not take symmetries into account. The current implementation of Sakuma's method in WANNIER90 being not compatible with the outputs of VASP, we were not able to use it for our simulations. However, in practice, the MLWFs we generated are close enough to SAWFs so that it was possible to identify a high-symmetry center and an associated point group. To test our compression method, we symmetrize the MLWFs according to the identified point group before applying the greedy procedure.

Optimization Procedure in the Discrete Setting

Let us now focus on the practical implementation of the second step of the greedy algorithm presented above. We present in this section the discrete formulation of problem (6.11). The discrete data representing the Wannier function W centered at $\mathbf{q} \in \mathbb{R}^3$ are composed of: i) the symmetry group G^0 and ii) the point values $(W(\mathbf{r}))_{\mathbf{r}\in\mathcal{M}}$ at each point of the cartesian grid $\mathcal{M}$.

Because we seek to minimize in particular the H^1-norm of the residual, we introduce an auxiliary Fourier representation of the data. Indeed, computing gradients is a fast (diagonal) operation in momentum space. The Fast Fourier Transform algorithm (FFT) can be used to efficiently transform data from position to momentum space. In particular, we obtain the unnormalized discrete representation of the Fourier transform $\widehat{u}$ of any function u as point values $(\widehat{u}(\mathbf{k}))_{\mathbf{k}\in\mathcal{K}}$ on a secondary Cartesian momentum-space grid that we denote by $\mathcal{K}$, containing the same number of points as the real-space grid, i.e $|\mathcal{K}| = |\mathcal{M}| = M$. Let us recall that the FFT algorithm requires M_x, M_y and M_z to be even numbers so that the momentum grid $\mathcal{K}$ is centered at zero. The H^s–norm (6.2) of u then has a discrete approximation given by

$$\|u\|_{H^s}^2 \approx \frac{|\Omega|}{M^2} \sum_{\mathbf{k}\in\mathcal{K}} \left(1 + |\mathbf{k}|^2\right)^s |\widehat{u}(\mathbf{k})|^2 . \tag{6.12}$$

At every greedy iteration $p \geq 0$, the exact cost functional $\mathcal{J}_p$ is approximated in the discrete setting by the functional $\mathcal{J}_p^{\mathcal{M}}$ defined as:

$$\mathcal{J}_p^{\mathcal{M}}(\boldsymbol{\alpha}, \sigma, \Lambda) := \frac{|\Omega|}{M^2} \sum_{\mathbf{k}\in\mathcal{K}} \left(1 + |\mathbf{k}|^2\right)^s \left| \widehat{R_p}(\mathbf{k}) - \widehat{\phi_{\boldsymbol{\alpha},\sigma,\Lambda}}(\mathbf{k}) \right|^2 , \tag{6.13}$$

where we recall that the residual R_p is computed from the approximation $\widetilde{W}_p$ at step p of the target Wannier function W,

$$R_p(\mathbf{r}) = W(\mathbf{r}) - \widetilde{W}_p(\mathbf{r}).$$

Note that while the Fourier transform of the SAGTO function $\phi_{\boldsymbol{\alpha},\sigma,\Lambda}$ which appears in this expression can be analytically computed, it is faster and more consistent to evaluate directly the Fourier transform of the residual numerically using the FFT algorithm.

Let us now focus on the implementation of the minimization problem (6.10) with the discrete error functional (6.13). As mentioned above, we use an off-the-shelf constrained optimization solver to find a local minimizer to the non-convex minimization problem

$$\min_{\boldsymbol{\alpha}\in\Omega,\ \sigma\in[\sigma_{\min},\sigma_{\max}]} \widetilde{\mathcal{J}}_p^{\mathcal{M}}(\boldsymbol{\alpha},\sigma), \tag{6.14}$$

the minimization over the coefficients Λ of the SAGTO being performed explicitly for fixed $\boldsymbol{\alpha}$, σ by solving the least-square problem

$$\widetilde{\mathcal{J}}_p^{\mathcal{M}}(\boldsymbol{\alpha},\sigma) = \min_{\Lambda\in\mathbb{R}^{\mathcal{I}}} \mathcal{J}_p^{\mathcal{M}}(\boldsymbol{\alpha},\sigma,\Lambda). \tag{6.15}$$

We tested both the *Sequential Quadratic Programming* (SQP) and the *Interior-Point* (IP) specializations of the *fmincon* optimization routine implemented in the Matlab Optimization Toolbox [MAT16]. To accelerate the computation, the gradient (but not the Hessian matrix) is also provided to the optimizer routine. Note that it is straightforward to compute explicitely the gradient by the chain rule in the case of the discrete error functional in (6.14) from the solution of the inner problem in (6.15); however its expression is quite cumbersome and will be omitted here for the sake of conciseness. The iterative procedure is stopped when one of the following two convergence criteria is met: (i) the norm of the gradient is smaller than $\delta = 10^{-10}$; (ii) the relative step size between two successive iterations is smaller than $\tau_{\min} = 10^{-12}$. In practice, our numerical tests show that both optimization routines (SQP or IP) provide similar results, with the IP method being slightly faster.

As usual with non-convex optimization problems, it is very important to provide a suitable initial guess for the parameters, namely here the center of the Gaussian $\boldsymbol{\alpha}^0 \in \Omega$ and its variance $\sigma_{\min} \leq \sigma^0 \leq \sigma_{\max}$. We propose here the following initialization procedure. First, the initial center position $\boldsymbol{\alpha}^0$ is chosen as a maximizer of the absolute value of the residual R_p:

$$\boldsymbol{\alpha}^0 \in \operatorname*{argmax}_{\mathbf{r}\in\Omega} |R_p(\mathbf{r})|. \tag{6.16}$$

Next, two different heuristic guesses are proposed to determine a suitable initial value σ^0, assuming that the function $|R_p|$ resembles locally a Gaussian function centered at $\boldsymbol{\alpha}^0$,

$$|R_p(\mathbf{r})| \approx |R_p(\boldsymbol{\alpha}^0)| \exp\left(-\frac{|\mathbf{r}-\boldsymbol{\alpha}^0|^2}{2\sigma^2}\right). \tag{6.17}$$

A first guess for σ^0 is obtained by a local data fit,

$$\sigma_1^0 = \operatorname*{argmin}_{\sigma>0} \sum_{\mathbf{r}\in\mathcal{M}\cap B(\boldsymbol{\alpha}^0)} \left(\frac{1}{2\sigma^2}\left|\mathbf{r}-\boldsymbol{\alpha}^0\right|^2 + \log\left|\frac{R_p(\mathbf{r})}{R_p(\boldsymbol{\alpha}^0)}\right|\right)^2,$$

where $B(\boldsymbol{\alpha}^0)$ is a cubic box centered at $\boldsymbol{\alpha}^0$ of side length $2r_{\text{cutoff}}$, with r_{cutoff} a user-defined parameter. This is in fact a linear least-squares fit, yielding the explicit formula:

$$\sigma_1^0 = \left(\frac{\displaystyle\sum_{\mathbf{r}\in\mathcal{M}\cap B(\boldsymbol{\alpha}^0)} \left|\mathbf{r}-\boldsymbol{\alpha}^0\right|^4}{\displaystyle -2\sum_{\mathbf{r}\in\mathcal{M}\cap B(\boldsymbol{\alpha}^0)} \left|\mathbf{r}-\boldsymbol{\alpha}^0\right|^2 \log\left|\frac{R_p(\mathbf{r})}{R_p(\boldsymbol{\alpha}^0)}\right|} \right)^{1/2}. \tag{6.18}$$

A second guess is provided by a property linking the variance of the standard normalized Gaussian $g(\mathbf{r}) = (2\pi\sigma^2)^{-1/2}\exp\left(-\frac{1}{2\sigma^2}|\mathbf{r}|^2\right)$ to its *full width at half maximum*, denoted ω_{h}:

$$\frac{\omega_{\text{h}}[g]}{\sigma} = 2\sqrt{2\log 2}.$$

The full width at half maximum is not well defined for arbitrary (non-radial) functions. We choose here to sample the full-width at half maximum along one-dimensional slices in all three directions x, y, z around $\boldsymbol{\alpha}^0$ and retain the smallest value. For an arbitrary function u assumed to have its maximum magnitude at the origin, we let:

$$\omega_h[u] := \min_{d\in\{x,y,z\}} \inf\left\{ |\gamma_+ - \gamma_-| : \qquad \gamma_- < 0 < \gamma_+ \text{ and } \left|\frac{u(\gamma_\pm \mathbf{e}_d)}{u(\mathbf{0})}\right| \leq \frac{1}{2} \right\},$$

where $\mathbf{e}_d$ is the standard unit vector in the direction $d \in \{x, y, z\}$. This leads to a second initial guess for the variance:

$$\sigma_2^0 = \frac{\omega_{\text{h}}\left[R_p(\cdot - \boldsymbol{\alpha}^0)\right]}{2\sqrt{2\log 2}}. \tag{6.19}$$

In practice, we project the values σ_1^0 given by (6.18) and σ_2^0 given by (6.19) on the interval $[\sigma_{\min}, \sigma_{\max}]$ and choose

$$\sigma^0 = \operatorname*{argmin}_{i=1,2} \mathcal{J}_p(\boldsymbol{\alpha}^0, \sigma_i^0, \Lambda^0). \tag{6.20}$$

Again, we do not claim that this procedure is optimal; it however gives satisfactory results for all the test cases we ran.

6.3 Numerical results

Our greedy algorithm allows us to compress a SAWF defined on a cartesian grid with $M = M_x M_y M_z$ points into a sum of SAGTOs parameterized by $p(4+|\mathcal{I}|)$ real numbers, where p is the number of SAGTOs in the expansion

$$\widetilde{W}_p^{\text{SA}}(\mathbf{r}) = \sum_{j=1}^{p} \phi^{\text{SA}}_{\boldsymbol{\alpha}_j,\sigma_j,\Lambda_j}(\mathbf{r}),$$

and where each $\phi^{\text{SA}}_{\boldsymbol{\alpha}_j,\sigma_j,\Lambda_j}$ is characterized by $(4+|\mathcal{I}|)$ real parameters. The compression gains for the four numerical examples detailed below, namely three 2D materials (single-layer graphene, hBN, and FeSe), and one bulk crystal (diamond-phase silicon), are collected in Table 6.1. The numerical parameters used in the construction of the original Wannier functions (as described in Section 6.2.5) are given in Table 6.2 for the sake of completeness.

Material	M	$\lvert\mathcal{I}\rvert$	ϵ	p	$p(4+\lvert\mathcal{I}\rvert)$	Compression ratio
Graphene	3237696	3	0.1 0.02	115 1036	805 7252	4022 446
hBN	4021248	3	0.1 0.03	137 1500	959 10500	4193 383
Si	110592	3	0.1 0.02	424 1500	2968 10500	38 10
FeSe	4032000	2	0.1 0.02	133 1610	798 9660	5052 417

Table 6.1 – Compression gains obtained with our implementation of the orthogonal greedy minimizing the H^1-norm of the residual for Wannier functions of graphene, hBN, FeSe, and bulk silicon, for different tolerance levels ϵ.

Material	$E_c[eV]$	$\mathcal{Q}$	η [Å]	$\mathcal{M}$
Graphene	500	$25 \times 25 \times 1$	20	$168 \times 132 \times 146$
hBN	500	$25 \times 25 \times 1$	20	$192 \times 154 \times 136$
FeSe	500	$19 \times 19 \times 1$	25	$120 \times 120 \times 280$
Si	300	$7 \times 7 \times 7$	–	$48 \times 48 \times 48$

Table 6.2 – Numerical parameters used for the construction of the original Wannier functions using VASP and Wannier90.

6.3.1 Graphene and single-layer hBN

The space groups of graphene and single-layer hBN are respectively

$$\begin{aligned} G &= \mathrm{Dg80} := \mathcal{R} \rtimes \underline{\mathrm{D}}_{6h}, \qquad \text{(space group of graphene)}, \\ G &= \mathrm{P\bar{6}m2} := \mathcal{R} \rtimes \underline{\mathrm{D}}_{3h}, \qquad \text{(space group of single-layer hBN)}, \end{aligned}$$

where $\mathcal{R}$ is the 2D Bravais lattice embedded in $\mathbb{R}^3$ defined as

$$\mathcal{R} = \mathbb{Z}a \begin{pmatrix} \sqrt{3}/2 \\ 1/2 \\ 0 \end{pmatrix} + \mathbb{Z}a \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \tag{6.21}$$

where $a > 0$ is the lattice parameter (which takes different values for graphene and hBN). The group $\underline{\mathrm{D}}_{6h}$ is a group of order 24, and has 12 irreducible representations, while the group $\underline{\mathrm{D}}_{3h}$ is a group of order 12, and has 6 irreducible representations.

The points O, A, B and C represented in Figure 6.1 are high-symmetry points of graphene (left) and hBN (right); their symmetry groups are respectively

$$\begin{aligned} &G_O \equiv \underline{D}_{6h}, \quad G_A \equiv \underline{D}_{3h}, \quad G_B \equiv \underline{D}_{3h}, \quad G_C \equiv \underline{D}_{2h}, \quad \text{(graphene)}, \\ &G_O \equiv \underline{D}_{3h}, \quad G_A \equiv \underline{D}_{3h}, \quad G_B \equiv \underline{D}_{3h}, \quad G_C \equiv \underline{D}_{1h}, \quad \text{(single-layer hBN)}. \end{aligned}$$

Let σ_h be the reflection operator with respect to the horizontal plane containing the graphene sheet. The two irreducible representations of the subgroup $\underline{C}_s = (E, \sigma_h)$ of $\underline{D}_{6h}$ and $\underline{D}_{3h}$ give rise to the decomposition of $L^2(\mathbb{R}^3)$ as

$$L^2(\mathbb{R}^3) = L^2_+(\mathbb{R}^3) \oplus L^2_-(\mathbb{R}^3),$$

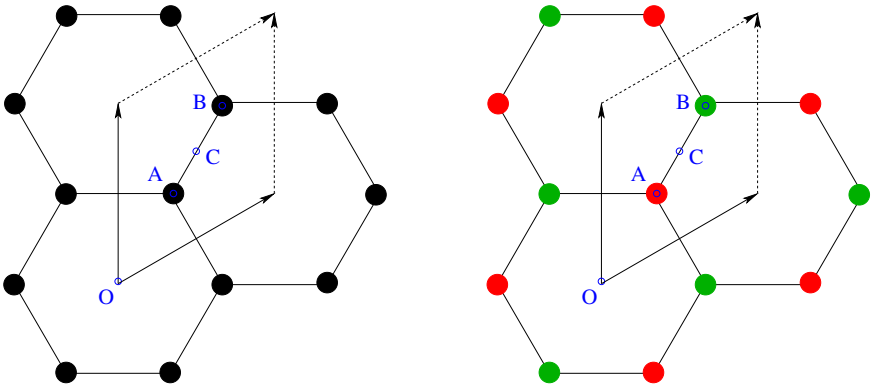


Figure 6.1 – The honeycomb lattices of graphene (left) and hBN (right). The black dots represent carbon atoms, the red dots boron atoms, and the green dots nitrogen atoms. The blue dots O, A, B, and C represent some high-symmetry points.

$\underline{D}_{3h}$	E	$2C_3$ (z)	$3C_2'$	σ_h (xy)	$2S_3$	$3\sigma_v$	linear functions	quadratic functions	cubic functions
A_2''	+1	+1	-1	-1	-1	+1	z	-	z^3, $z(x^2+y^2)$

Table 6.3 – Character of the A"$_2$ representation of the group $\underline{D}_{3h}$

where

$$L^2_+(\mathbb{R}^3) = \mathrm{Ker}(\sigma_h - 1), \quad L^2_-(\mathbb{R}^3) = \mathrm{Ker}(\sigma_h + 1).$$

The bands associated with $L^2_+(\mathbb{R}^3)$ are the σ bands, the ones associated with $L^2_-(\mathbb{R}^3)$ the π bands. The bands of interest for graphene and single-layer hBN are the valence and conduction bands closer to the Fermi level. For graphene, these are the π bands originating from the $2p_z$ orbitals of the carbon atoms.

The SAWF functions for graphene and single-layer hBN considered here are centered at point A and are transformed according to the (one-dimensional) A_2'' representation of $\underline{D}_{3h}$, whose character is given in Table 6.3.

Graphical representations of the original Wannier functions generated by Wannier90 and of their compressions into Gaussian orbitals obtained with the VESTA visualization package [MI08], are displayed in Figures 6.2 (graphene) and 6.3 (hBN). The decays of the L^2 and H^1-norms of the residuals along the iterations of our implementation of the orthogonal greedy algorithm aiming at minimizing the H^1-norm of the residual, are plotted in Figure 6.4.

6.3.2 Single-layer SeFe

The space group of single-layer FeSe is

$$G = \mathrm{P4/nmm} := \mathcal{R} \rtimes \underline{\mathrm{D}}_{4h},$$

where $\mathcal{R}$ is the 2D square lattice of $\mathbb{R}^3$ defined as

$$\mathcal{R} = \mathbb{Z}a \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + \mathbb{Z}a \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \tag{6.22}$$

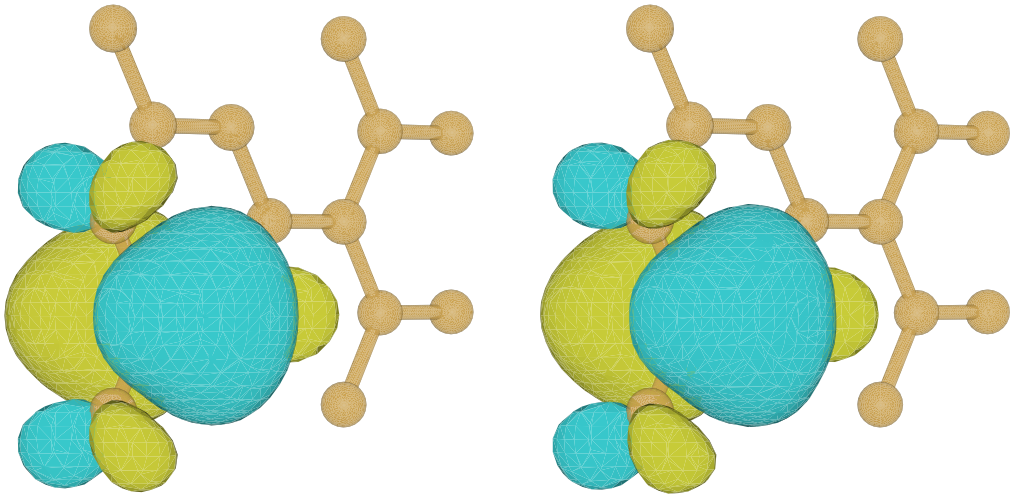
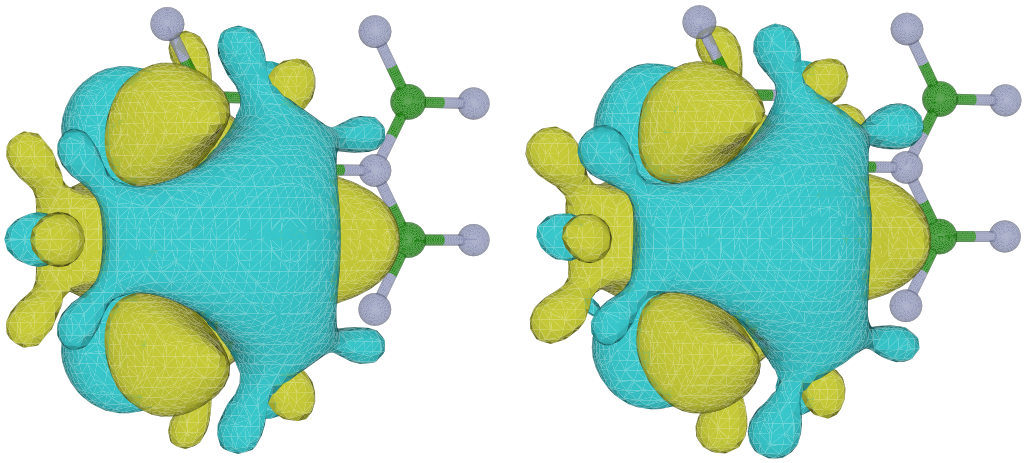


Figure 6.2 – Wannier function of graphene generated with VASP and Wannier90 (left), and its compression into Gaussian orbitals (right). Positive and negative iso-surfaces corresponding to 15% of the maximum value are plotted. .



(a) (b)

Figure 6.3 – Wannier function of single-layer hBN generated with VASP and Wannier90 (left), and its compression into Gaussian orbitals (right). Positive and negative iso-surfaces corresponding to 15% of the maximum value are plotted.

where $a > 0$ is the lattice parameter. The group $\underline{D}_{4h}$ is of order 16 and has 10 irreducible representations. The symmetry group of the high-symmetry point A represented in Figure 6.5 is $G_A = \underline{C}_{2v}$.

The Wannier function considered here corresponds to a d−type orbital on an Fe atom centered at point A and is transformed according to the one-dimensional A_1 representation of $\underline{C}_{2v}$, whose character is given in Table 6.4. Graphical representations of the original Wannier function and of its compression into Gaussian orbitals are given

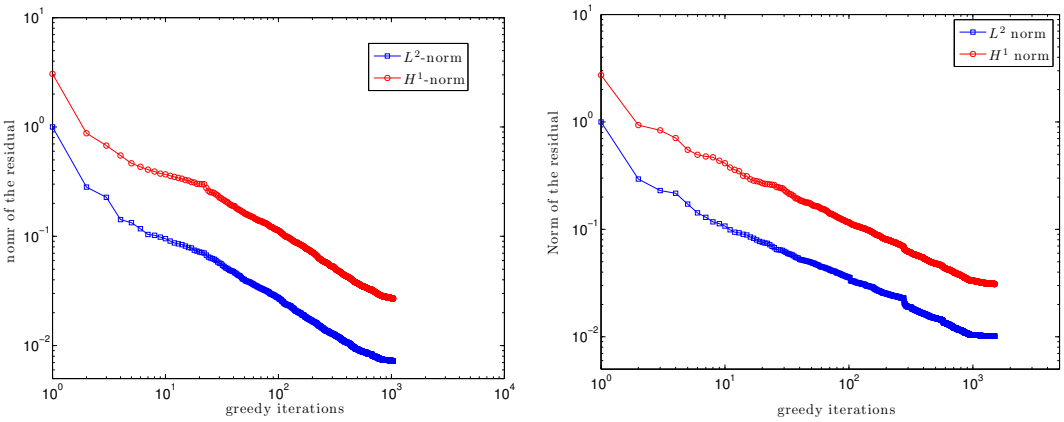
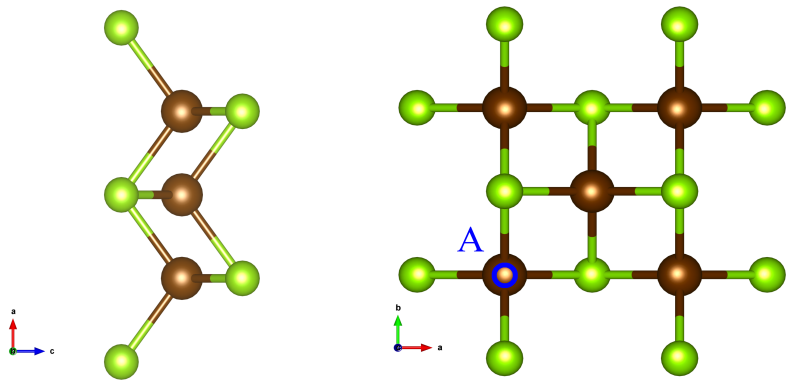


Figure 6.4 – Decays of the L^2 and H^1-norms of the residual for our implementation of the orthogonal greedy algorithm minimizing the H^1-norm of the residual (left: graphene, right: hBN)



(a) side view (b) top view

Figure 6.5 – Crystalline structure of FeSe (2D layer with a finite thickness). The brown balls represent Fe atoms and the green balls represent Se atoms. The spotted point A corresponds to the high-symmetry point at which the Wannier function is centered.

in Figure 6.6. The decays of the L^2 and H^1-norms of the residual along the iterations of our implementation of the orthogonal greedy algorithm minimizing the H^1-norm of the residual are plotted in Figure 6.9.

6.3.3 Diamond-phase silicon

The space group of diamond-phase silicon is

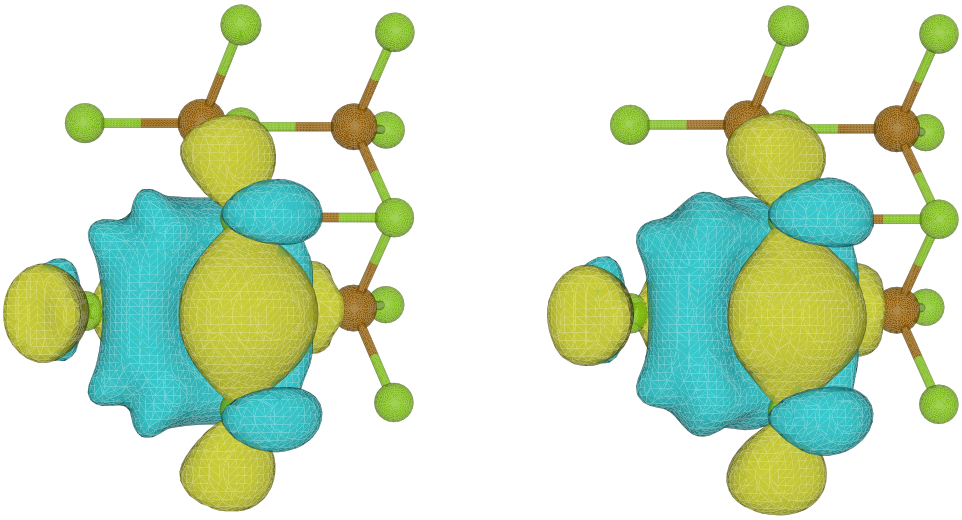
$$G = \mathrm{Fd3m} := \mathcal{R} \rtimes \underline{O}_h$$

where $\mathcal{R}$ is the Bravais lattice of $\mathbb{R}^3$ defined as

$$\mathcal{R} = \mathbb{Z}a \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} + \mathbb{Z}a \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} + \mathbb{Z}a \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}, \tag{6.23}$$

$\underline{C}_{2v}$	E	C_2 (z)	σ_v(xz)	σ_v(yz)	linear functions	quadratic functions	cubic functions
A_1	+1	+1	+1	+1	z	x^2, y^2, z^2	z^3, x^2z, y^2z

Table 6.4 – Character of the A_1 representation of the group $\underline{C}_{2v}$.



(a) (b)

Figure 6.6 – Wannier function of single-layer FeSe generated with VASP and Wannier90 (left), and its compression into Gaussian orbitals (right). Positive and negative iso-surfaces corresponding to 12% of the maximum value are plotted.

where $a > 0$ is the lattice parameter. The group $\underline{O}_h$ is of order 48 and has 10 irreducible representations. The Wannier function considered here corresponds a p_y-type orbital centered at the high-symmetry point A represented in Figure 6.7 whose symmetry group is $G_A = \underline{C}_{2v}$.

It is transformed according to the one-dimensional irreducible representation A_1 of the group $\underline{C}_{2v}$. Let us mention the following point : since the basis $\hat{\mathbf{x}} = (1, 0, 1)$, $\hat{\mathbf{y}} = (1, 1, 0)$ and $\hat{\mathbf{z}}(0, 1, 1)$ is not orthonormal in $\mathbb{R}^3$, the symmetry operators $C_2(z)$, $\sigma_v(xz)$ and $\sigma_v(yz)$ must be adapted to this geometry. Indeed, the two-fold rotation C_2 is about the axis of direction $(0, 1, 1)$ and the two reflexions σ_v are defined with respect to the planes $\mathcal{P}_1$ and $\mathcal{P}_2$ of cartesian equations $x + z = 0$ and $y + z = 0$ respectively. Graphical representations of the original Wannier function and of its compression into Gaussian orbitals are given in Figure 6.8. The decays of the L^2 and H^1-norms of the residual along the iterations of our implementation of the greedy algorithm aiming at constructing H^1-norm approximations of the Wannier function are plotted in Figure 6.9.

Acknowledgments

This work was supported in part by ARO MURI Award W911NF-14-1-0247.

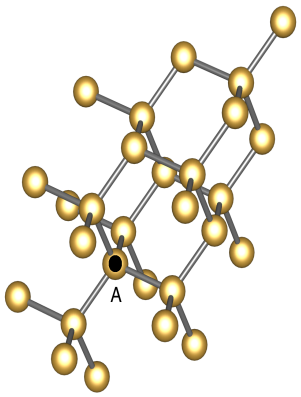
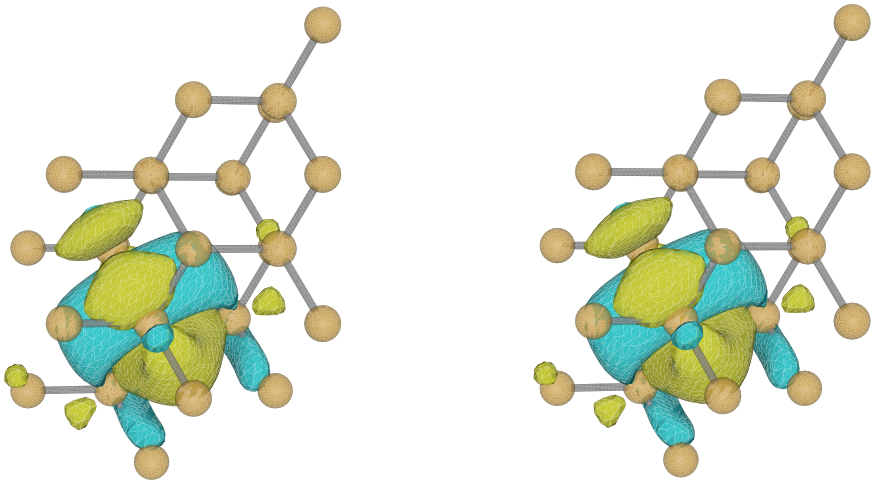


Figure 6.7 – Crystalline structure of Silicon. The brown balls represent Si atoms and the spotted point A corresponds to the high-symmetry point where the Wannier function is centered.



(a) (b)

Figure 6.8 – Wannier function of bulk Silicon (diamond phase) generated by Wannier90 (left), and its compression into Gaussian orbitals (right). Positive and negative iso-surfaces corresponding to 10% of the maximum value are plotted.

Appendix: symmetry-adapted Wannier functions

A.1 Space group of a periodic material

Consider a periodic material with M nuclei of charges $z_1, \cdots, z_M$ per unit cell. The nuclear charge distribution in the material is of the form

$$\nu = \sum_{\mathbf{R}\in\mathcal{R}} \sum_{m=1}^{M} z_m \delta_{\mathbf{R}_m+\mathbf{R}},$$

where $\mathcal{R}$ is the Bravais lattice of the crystal (embedded in $\mathbb{R}^3$ if the material is a 2D material), $\delta_{\mathbf{a}}$ the Dirac mass at point $\mathbf{a} \in \mathbb{R}^3$, and $\mathbf{R}_1, \cdots, \mathbf{R}_M \in \mathbb{R}^3$ the positions of

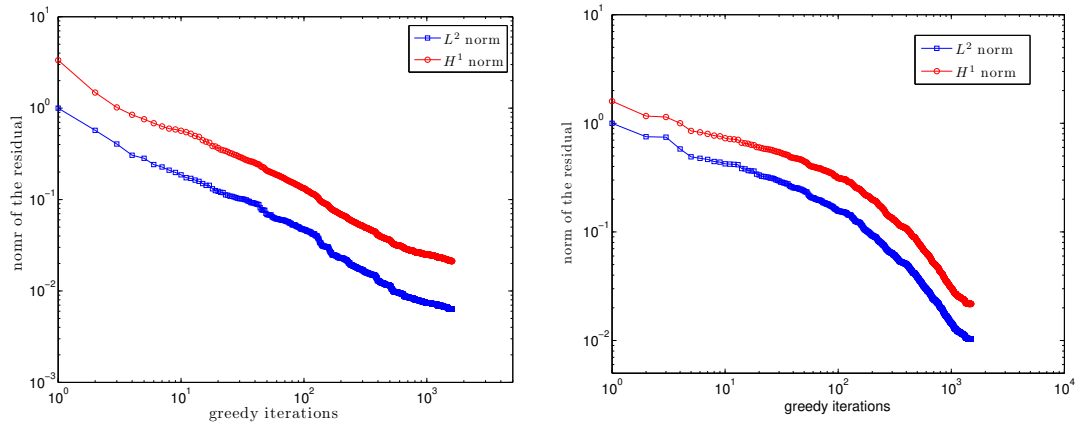


Figure 6.9 – Decays of the L^2 and H^1-norms of the residual for our implementation of the orthogonal greedy algorithm minimizing the H^1-norm of the residual (left: FeSe, right: diamond-phase sillicon)

the nuclei laying in the unit cell. The space group $G = \mathcal{R} \rtimes G_{\mathrm{p}}$ of the crystal is the semidirect product of $\mathcal{R}$ and a finite point group G_{p} (a finite subgroup of $O(3)$). Recall that the composition law in $\mathcal{R} \rtimes G_{\mathrm{p}}$ is defined as

$$\forall g_1 = (\mathbf{R}_1, \Theta_1), \quad g_2 = (\mathbf{R}_2, \Theta_2), \quad g_1 g_2 = (\Theta_1 \mathbf{R}_2 + \mathbf{R}_1, \Theta_1 \Theta_2),$$

and that the natural representation of G in $\mathbb{R}^3$ is given by

$$\forall g = (\mathbf{R}, \Theta) \in G, \quad \forall \mathbf{r} \in \mathbb{R}^3, \quad \hat{g}\mathbf{r} = \widehat{(\mathbf{R}, \Theta)}\mathbf{r} = \Theta \mathbf{r} + \mathbf{R}.$$

Note that

$$\forall g = (\mathbf{R}, \Theta) \in G, \quad g^{-1} = (-\Theta^{-1}\mathbf{R}, \Theta^{-1}) \quad \text{and} \quad \forall \mathbf{r} \in \mathbb{R}^3, \quad \hat{g}^{-1}\mathbf{r} = \Theta^{-1}(\mathbf{r} - \mathbf{R}).$$

The space group of the crystal is the largest group (for an optimal choice of the origin of the Cartesian frame) leaving ν invariant:

$$\forall g \in G, \quad \widehat{g}\nu := \sum_{\mathbf{R} \in \mathcal{R}} \sum_{m=1}^{M} z_m \delta_{\hat{g}(\mathbf{R}_m + \mathbf{R})} = \nu.$$

The group G has a natural unitary representation $\Pi = (\Pi_g)_{g \in G}$ on $L^2(\mathbb{R}^3)$ defined by

$$\forall g = (\mathbf{R}, \Theta) \in G, \quad (\Pi_g \psi)(\mathbf{r}) = \psi(\hat{g}^{-1}\mathbf{r}) = \psi(\Theta^{-1}(\mathbf{r} - \mathbf{R})).$$

Denoting by E the identity matrix of rank 3, and by $\tau = (\tau_{\mathbf{a}})_{\mathbf{a} \in \mathbb{R}^3}$ the natural unitary representation on $\mathbb{R}^3$ on $L^2(\mathbb{R}^3)$ defined by

$$\forall \mathbf{a} \in \mathbb{R}^3, \quad \forall \phi \in L^2(\mathbb{R}^3), \quad (\tau_{\mathbf{a}} \phi)(\mathbf{r}) = \phi(\mathbf{r} - \mathbf{a}),$$

we have $\Pi_{(\mathbf{R}, E)} = \tau_{\mathbf{R}}$ for all $\mathbf{R} \in \mathcal{R}$, so that $(\tau_{\mathbf{R}})_{\mathbf{R} \in \mathcal{R}}$ is an abelian subgroup of Π.

A.2 Bloch transform

Let us now recall the basics of Bloch theory. We denote by Γ a unit cell of the Bravais lattice $\mathcal{R}$, by

$$L^2_{\rm per}(\Gamma) := \left\{ u \in L^2_{\rm loc}(\mathbb{R}^3, \mathbb{C}),\ u\ \mathcal{R}\text{-periodic} \right\}, \quad \langle u|v\rangle_{L^2_{\rm per}} := \int_\Gamma \overline{u(\mathbf{r})}\, v(\mathbf{r})\, d\mathbf{r},$$

the Hilbert space of locally square-integrable $\mathcal{R}$-periodic functions $\mathbb{C}$-valued functions on $\mathbb{R}^3$, by $\mathcal{R}^*$ the dual lattice of $\mathcal{R}$ and by Γ^* the first Brillouin zone. The Bloch transform associated with $\mathcal{R}$ (see *e.g.* [RS78c, Section XIII.16]) is the unitary transform

$$L^2(\mathbb{R}^3, \mathbb{C}) \ni \phi \mapsto (\phi_{\mathbf{k}})_{\mathbf{k}\in\Gamma^*} \in \mathcal{H} = \fint_{\Gamma^*}^{\oplus} L^2_{\rm per}(\Gamma)\, d\mathbf{k}$$

where $\fint_{\Gamma^*}$ is a notation for the normalized integral $|\Gamma^*|^{-1} \int_{\Gamma^*}$, where $\mathcal{H}$ is endowed with the inner product

$$\langle (\phi_{\mathbf{k}})_{\mathbf{k}\in\Gamma^*} | (\psi_{\mathbf{k}})_{\mathbf{k}\in\Gamma^*} \rangle_{\mathcal{H}} = \fint_{\Gamma^*} \langle \phi_{\mathbf{k}} | \psi_{\mathbf{k}} \rangle_{L^2_{\rm per}}\, d\mathbf{k},$$

and where, for a smooth fast decaying function ϕ, the periodic function $\phi_{\mathbf{k}}$ is given by

$$\phi_{\mathbf{k}}(\mathbf{r}) = \sum_{\mathbf{R}\in\mathcal{R}} \phi(\mathbf{r}+\mathbf{R}) e^{-i\mathbf{k}\cdot(\mathbf{r}+\mathbf{R})}.$$

The original function ϕ is recovered from its Bloch transform using the inversion formula

$$\phi(\mathbf{r}) = \fint_{\Gamma^*} \phi_{\mathbf{k}}(\mathbf{r})\, e^{i\mathbf{k}\cdot\mathbf{r}}\, d\mathbf{k}.$$

Consider a one-body Hamiltonian

$$H = -\frac{1}{2}\Delta + V_{\rm per}, \quad V_{\rm per} \in L^2_{\rm per}(\Gamma),$$

describing the electronic properties of the material (we ignore spin for simplicity). In the absence of symmetry breaking, H commutes with all the unitary operators in $\Pi = (\Pi_g)_{g\in G}$. In particular, H commutes with the translations $\tau_{\mathbf{R}}$, $\mathbf{R} \in \mathcal{R}$, and is therefore decomposed by the Bloch transform:

$$H = \fint_{\Gamma^*} H_{\mathbf{k}}\, d\mathbf{k},$$

meaning that there exists a family $(H_{\mathbf{k}})_{\mathbf{k}\in\Gamma^*}$ of self-adjoint operators on $L^2_{\rm per}(\Gamma)$ such that for any ϕ in the domain of H, $\phi_{\mathbf{k}}$ is almost everywhere in the domain of $H_{\mathbf{k}}$ and

$$(H\phi)_{\mathbf{k}} = H_{\mathbf{k}}\phi_{\mathbf{k}}.$$

It is well-known that

$$H_{\mathbf{k}} = \frac{1}{2}(-i\nabla + \mathbf{k})^2 + V_{\rm per} = -\frac{1}{2}\Delta - i\mathbf{k}\cdot\nabla + \frac{1}{2}|\mathbf{k}|^2 + V_{\rm per}.$$

The operator $H_{\mathbf{k}}$ can in fact be defined for any $\mathbf{k} \in \mathbb{R}^3$, and it holds

$$\forall \mathbf{k} \in \mathbb{R}^3, \quad \forall \mathbf{K} \in \mathcal{R}^*, \quad H_{\mathbf{k}+\mathbf{K}} = V_{\mathbf{K}} H_{\mathbf{k}} V_{\mathbf{K}}^*, \tag{6.24}$$

where $V_{\mathbf{K}}$ is the unitary operator on $L^2_{\mathrm{per}}(\Gamma)$ defined by

$$\forall u \in L^2_{\mathrm{per}}(\Gamma), \quad (V_{\mathbf{K}} u)(\mathbf{r}) = e^{-i\mathbf{K}\cdot\mathbf{r}} u(\mathbf{r}).$$

As a consequence, for all $\mathbf{k} \in \mathbb{R}^3$ and $\mathbf{K} \in \mathcal{R}^*$, $H_{\mathbf{k}}$ and $H_{\mathbf{k}+\mathbf{K}}$ are unitary equivalent, and therefore have the same spectrum. Not every Π_g *a priori* commutes with the translation operators $\tau_{\mathbf{R}}$, $\mathbf{R} \in \mathcal{R}$. The operator Π_g is therefore not in general decomposed by the Bloch transform. On the other hand, denoting by $U = (U_\Theta)_{\Theta \in G_p}$ the natural unitary representation of G_p in $L^2_{\mathrm{per}}(\Gamma)$ defined by

$$\forall \Theta \in G_{\mathrm{p}}, \quad \forall u \in L^2_{\mathrm{per}}(\Gamma), \quad (U_\Theta u)(\mathbf{r}) = u(\Theta^{-1}\mathbf{r}),$$

the Bloch representation of the operator Π_g, $g = (\mathbf{R}, \Theta) \in G$, has a simple form:

$$[\Pi_{(\mathbf{R},\Theta)}]_{\mathbf{k},\mathbf{k}'} = e^{-i\mathbf{k}\cdot\mathcal{R}} U_\Theta \delta_{\mathbf{k}',\Theta^{-1}\mathbf{k}},$$

that is:

$$[\Pi_{(\mathbf{R},\Theta)}\phi]_{\mathbf{k}}(\mathbf{r}) = e^{-i\mathbf{k}\cdot\mathcal{R}} \phi_{\Theta^{-1}\mathbf{k}}(\Theta^{-1}\mathbf{r}).$$

Since H commutes with all the Π_g's, this implies that the family $(H_{\mathbf{k}})_{\mathbf{k}\in\Gamma^*}$ satisfies the covariance relation

$$\forall \mathbf{k} \in \mathbb{R}^d, \quad \forall \Theta \in G_p, \quad H_{\Theta\mathbf{k}} = U_\Theta H_{\mathbf{k}} U_\Theta^*.$$

For each $\mathbf{k} \in \mathbb{R}^3$, the operator $H_{\mathbf{k}}$ is self-adjoint on $L^2_{\mathrm{per}}(\Gamma)$ and is bounded below. If $\mathcal{R}$ is a three-dimensional lattice (3D crystal), then $H_{\mathbf{k}}$ has a compact resolvent and its spectrum is purely discrete. If $\mathcal{R}$ is a two-dimensional lattice (2D material), then the essential spectrum of $H_{\mathbf{k}}$ is a half-line $[\Sigma_{\mathbf{k}}, +\infty)$.

A.3 Symmetry-adapted Wannier functions

We assume here that H has a finite number $n \geq 1$ of bands isolated from the rest of the spectrum, that is that there exist two continuous $\mathbb{R}$-valued $\mathcal{R}$-periodic functions $\mathbf{k} \mapsto \mu_-(\mathbf{k})$ and $\mathbf{k} \mapsto \mu_+(\mathbf{k})$ such that $\mu_-(\mathbf{k}) < \mu_+(\mathbf{k})$, $\mu_\pm(\mathbf{k}) \notin \sigma(H_{\mathbf{k}})$ and $\mathrm{tr}\big(\mathbb{1}_{[\mu_-(\mathbf{k}),\mu_+(\mathbf{k})]}(H_{\mathbf{k}})\big) = n$ for all $\mathbf{k} \in \mathbb{R}^3$. We denote by $\epsilon_{1,\mathbf{k}} \leq \epsilon_{2,\mathbf{k}} \leq \cdots \leq \epsilon_{n,\mathbf{k}}$ the eigenvalues of $H_{\mathbf{k}}$ laying in the range $[\mu_-(\mathbf{k}), \mu_+(\mathbf{k})]$ (counting multiplicities). The functions $\mathbf{k} \mapsto \epsilon_{n,\mathbf{k}}$ are Lipschitz continuous, and, in view (6.24), are also $\mathcal{R}$-periodic.

A generalized Wannier function associated to these n bands is a function of the form

$$\forall \mathbf{r} \in \mathbb{R}^3, \quad W(\mathbf{r}) = \fint_{\Gamma^*} u_{\mathbf{k}}(\mathbf{r})\, e^{i\mathbf{k}\cdot\mathbf{r}}\, d\mathbf{k}, \quad u_{\mathbf{k}} \in \mathrm{Ran}(\mathbb{1}_{[\mu_-(\mathbf{k}),\mu_+(\mathbf{k})]}(H)), \quad \|u_{\mathbf{k}}\|_{L^2_{\mathrm{per}}} = 1.$$

Let $\mathbf{q}$ be a site of the unit cell of the crystalline lattice[1]. We denote by

$$G_{\mathbf{q}} = \{g = (\mathbf{R}, \Theta) \in G \mid \hat{g}\mathbf{q} = \Theta\mathbf{q} + \mathbf{R} = \mathbf{q}\}$$

[1]Here the lattice is not in general a Bravais lattice. For graphene and hBN, this is a honeycomb lattice.

the finite subgroup of G leaving $\mathbf{q}$ invariant. The point $\mathbf{q}$ is called a high-symmetry point if $G_{\mathbf{q}}$ is not trivial. Setting $\mathbf{R}_\Theta = \mathbf{q} - \Theta\mathbf{q}$, we have

$$G_{\mathbf{q}} = \left\{ g = (\mathbf{R}_\Theta, \Theta),\ \Theta \in G^0_{\mathbf{q}} \right\},$$

where $G^0_{\mathbf{q}}$ is a subgroup of G_{p}.

A symmetry-adapted Wannier function centered at a high-symmetry point $\mathbf{q}$ is a Wannier function W such that

1. the finite-dimensional space

$$\mathcal{H}_{W,\mathbf{q}} := \mathrm{Span}\left(\Pi_g W,\ g \in G_{\mathbf{q}}\right)$$

   is Π_g-invariant for any $g \in G_{\mathbf{q}}$;

2. $(\Pi_g|_{\mathcal{H}_{W,\mathbf{q}}})_{g\in G_{\mathbf{q}}}$ defines an irreducible unitary representation β of $G_{\mathbf{q}}$.

Let $n_\beta := \dim(\mathcal{H}_{W,\mathbf{q}})$ be the dimension of this representation and $(W^{(\beta)}_{i,1})_{1\le i\le n_\beta}$ be a basis of $\mathcal{H}_{W,\mathbf{q}}$ such that $W^{(\beta)}_{1,1} = W$. Let $(d^\beta(\Theta))_{\Theta\in G^0_{\mathbf{q}}} \in (\mathbb{C}^{n_\beta\times n_\beta})^{n_{\mathbf{q}}}$ be the matrix representation of the group $G^0_{\mathbf{q}}$ in

$$\mathcal{H}^0_{W,\mathbf{q}} := \mathrm{Span}\left(\Pi_\Theta \tau_{-\mathbf{q}} W,\ \Theta \in G^0_{\mathbf{q}}\right), \quad \text{where} \quad \Pi_\Theta := \Pi_{(\mathbf{0},\Theta)}.$$

We therefore have

$$\forall \Theta \in G^0_{\mathbf{q}}, \quad \Pi_\Theta \left(\tau_{-\mathbf{q}} W^{(\beta)}_{i,1}\right) = \sum_{i'=1}^{n_\beta} d^{(\beta)}_{i',i}(\Theta) \left(\tau_{-\mathbf{q}} W^{(\beta)}_{i',1}\right),$$

so that

$$\forall (\mathbf{R}_\Theta, \Theta) \in G_{\mathbf{q}}, \quad \Pi_{(\mathbf{R}_\Theta,\Theta)} W^{(\beta)}_{i,1} = \sum_{i'=1}^{n_\beta} d^{(\beta)}_{i',i}(\Theta) W^{(\beta)}_{i',1}.$$

If the representation β is one-dimensional ($n_\beta = 1$), then $(d^\beta(\Theta))_{\Theta\in G^0_{\mathbf{q}}}$ is the character of the corresponding representation of $G^0_{\mathbf{q}} \subset G_p$ in $\mathcal{H}^0_{W,\mathbf{q}}$.

Let $J = |G_{\mathrm{p}}|/|G_{\mathbf{q}}| \in \mathbb{N}^*$. Then, there exist $(g_j)_{1\le j\le J} \in G^J$ such that

$$G = \sum_{j=1}^{J} \sum_{\mathbf{R}\in\mathcal{R}} (\mathbf{R}|E) g_j G_{\mathbf{q}}.$$

More precisely, there exist $(g_j)_{1\le j\le J} \in G^J$ such that

- for each $1 \le j \le J$, $\mathbf{q}_j := \hat{g}_j \mathbf{q} \in \Gamma$;
- any $g \in G$ can be decomposed in a unique way as

$$g = (\mathbf{R}|E) g_j g_{\mathbf{q}}$$

  for a unique triplet $(\mathbf{R}, j, g_{\mathbf{q}}) \in \mathcal{R} \times |[1, J]| \times G_{\mathbf{q}}$.

For each $1 \leq i \leq n_\beta$, $1 \leq j \leq J$ and $\mathbf{R} \in T$, we set

$$W^{(\beta)}_{i,j,\mathbf{R}} = \Pi_{(\mathbf{R}|E)g_j} W^{(\beta)}_{i,1},$$

and we then define

$$\mathcal{H}_W = \overline{\mathrm{Span}\left(W^{(\beta)}_{i,j,\mathbf{R}},\ 1 \leq i \leq n_\beta,\ 1 \leq j \leq J,\ \mathbf{R} \in \mathcal{R}\right\}}.$$

In other words, $\mathcal{H}_W$ is the closure of the vector space generated by the mother SAWF W and all the SAWFs obtained by letting the elements of G act on W.

The space $\mathcal{H}_W \subset H^2(\mathbb{R}^3)$ is both H-invariant and Π-invariant, and for any $g \in G$, the action of Π_g on $W^{(\beta)}_{i,j,\mathbf{R}}$ can be computed as follows. Let $(\mathbf{R}', j', g'_{\mathbf{q}})$ the unique element of $\mathcal{R} \times |[1, J]| \times G_{\mathbf{q}}$ such that $g(\mathbf{R}|E)g_j = (\mathbf{R}'|E)g_{j'}g'_{\mathbf{q}}$. We have

$$\begin{aligned}
\Pi_g W^{(\beta)}_{i,j,\mathbf{R}} &= \Pi_g \Pi_{(\mathbf{R}|E)g_j} W^{(\beta)}_{i,1} = \Pi_{g(\mathbf{R}|E)g_j} W^{(\beta)}_{i,1} = \Pi_{(\mathbf{R}'|E)g_{j'}g'_{\mathbf{q}}} W^{(\beta)}_{i,1} \\
&= \Pi_{(\mathbf{R}'|E)g_{j'}} \Pi_{g'_{\mathbf{q}}} W^{(\beta)}_{i,1} = \Pi_{(\mathbf{R}'|E)g_{j'}} \left(\sum_{i'=1}^{n_\beta} d^{(\beta)}_{i',i}(\Theta'_q) W^{(\beta)}_{i',1} \right) \\
&= \sum_{i'=1}^{n_\beta} d^{(\beta)}_{i',i}(\Theta'_q) W^{(\beta)}_{i',j',\mathbf{R}'}.
\end{aligned}$$

The index j' is the unique integer in the range $|[1, J]|$ such that

$$\hat{g}(\mathbf{q}_j + \mathbf{R}) \in \mathbf{q}_{j'} + \mathcal{R}.$$

The explicit expressions of $\mathbf{R}'$ and $\Theta'_{\mathbf{q}}$ as functions of $(\mathbf{R}, j)$ and $g = (\mathbf{R}, \Theta)$ are the following

$$\Theta'_{\mathbf{q}} = \Theta_{j'}^{-1} \Theta \Theta_j, \quad \mathbf{R}' = \hat{g}\mathbf{q}_j - \mathbf{q}_{j'} + \Theta\mathbf{R}.$$

Constructing a basis of SAWFs for the n bands defined by the functions μ_- and μ_+ amounts to finding $s \in \mathbb{N}^*$ high-symmetry points $\mathbf{q}_1, \cdots, \mathbf{q}_s$, and s SAWFs Wannier functions $W_1, \cdots, W_s$ respectively centered at the points $\mathbf{q}_1, \cdots, \mathbf{q}_s$, such that

$$\fint_{\Gamma^*}^{\oplus} \mathrm{Ran}\left(\mathbb{1}_{[\mu_-(\mathbf{k}), \mu_+(\mathbf{k})]}(H)\right)\, d\mathbf{k} = \mathcal{H}_{W_1} \oplus \cdots \oplus \mathcal{H}_{W_s}.$$

This is the purpose of the numerical method introduced in [Sak13].

BIBLIOGRAPHY

[A+90] Herbert Amann et al. Dynamic theory of quasilinear parabolic equations. ii. reaction-diffusion systems. *Differential and Integral Equations*, 3(1):13–75, 1990.

[ABL00] Cécile Ané, Dominique Bakry, and Michel Ledoux. *Sur les inégalités de Sobolev logarithmiques*. Société mathématique de France Paris, 2000.

[Ali79] Nicholas D Alikakos. Lp bounds of solutions of reaction-diffusion equations. *Communications in Partial Differential Equations*, 4(8):827–868, 1979.

[Ama88] Herbert Amann. Dynamic theory of quasilinear parabolic equations—i. abstract evolution equations. *Nonlinear Analysis: Theory, Methods & Applications*, 12(9):895–919, 1988.

[Ama89] Herbert Amann. Dynamic theory of quasilinear parabolic systems. iii global existence. *Mathematische Zeitschrift*, 202(2):219–250, 1989.

[And04] Alan L Andrew. Numerical solution of inverse sturm–liouville problems. *ANZIAM Journal*, 45:326–337, 2004.

[AS78] JE Avron and B Simon. Analytic properties of band functions. *Annals of Physics*, 110(1):85–101, 1978.

[AUT00] P. Markowich A. Unterreiter, A. Arnold and G. Toscani. On Generalized Csiszár-Kullback Inequalities. *Monatshefte fuer Mathematik*, 131(3):235–253, 2000.

[BCC+17] Athmane Bakhta, Eric Cancès, Paul Cazeaux, Shiang Fang, and Efthimios Kaxirs. Compression of wannier functions into gaussian type orbitlas. 2017.

[BCL99] H. Brézis, P.G. Ciarlet, and J.L. Lions. *Analyse fonctionnelle: théorie et applications*. Collection Mathématiques appliquées pour la maîtrise. Dunod, 1999.

[BDFPS10] Martin Burger, Marco Di Francesco, Jan-Frederik Pietschmann, and Bärbel Schlake. Nonlinear cross-diffusion with size exclusion. *SIAM Journal on Mathematical Analysis*, 42(6):2842–2871, 2010.

[BE16] Athmane Bakhta and Virginie Ehrlacher. Cross-diffusion systems with non-zero-flux boundary conditions on a moving domain. *arXiv preprint arXiv:1611.07698*, 2016.

[BEG17] Athmane Bakhta, Virginie Ehrlacher, and David Gontier. Numerical reconstruction of the first band (s) in an inverse hill's problem. *arXiv preprint arXiv:1709.07023*, 2017.

[BGS12] Laurent Boudin, Bérénice Grec, and Francesco Salvarani. A mathematical and numerical analysis of the maxwell-stefan diffusion equations. *Discrete and Continuous Dynamical Systems-Series B*, 17(5):1427–1440, 2012.

[BL17] Athmane Bakhta and Damiano Lombardi. An a posteriori error estimator based on shifts for positive hermitian eigenvalue problems. 2017.

[BLMP09] Mostafa Bendahmane, Thomas Lepoutre, Americo Marrocco, and Benoît Perthame. Conservative cross diffusions and pattern formation through relaxation. *Journal de mathématiques pures et appliquées*, 92(6):651–667, 2009.

[Blo29] Felix Bloch. über die quantenmechanik der elektronen in kristallgittern. *Zeitschrift für Physik*, 52(7):555–600, 1929.

[Blö94] Peter E Blöchl. Projector augmented-wave method. *Physical review B*, 50(24):17953, 1994.

[Bot11] Dieter Bothe. On the maxwell-stefan approach to multicomponent diffusion. In *Parabolic problems*, pages 81–93. Springer, 2011.

[Boy50] S Francis Boys. Electronic wave functions. i. a general method of calculation for the stationary states of any molecular system. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, volume 200, pages 542–554. The Royal Society, 1950.

[BPC+07] Christian Brouder, Gianluca Panati, Matteo Calandra, Christophe Mourougane, and Nicola Marzari. Exponential localization of wannier functions in insulators. *Physical review letters*, 98(4):046402, 2007.

[CDA16] Xiuqing Chen, Esther S. Daus, and Juengel Ansgar. Global existence analysis of cross-diffusion population systems for multiple species. *arXiv:1608.03696*, 2016.

[CHN16] Horia D Cornean, Ira Herbst, and Gheorghe Nenciu. On the construction of composite wannier functions. In *Annales Henri Poincaré*, volume 17, pages 3361–3398. Springer, 2016.

[CJ04] Li Chen and Ansgar Juengel. Analysis of a multidimensional parabolic population model with strong cross-diffusion. *SIAM journal on mathematical analysis*, 36(1):301–322, 2004.

[CJ06] Li Chen and Ansgar Juengel. Analysis of a parabolic cross-diffusion population model without self-diffusion. *Journal of Differential Equations*, 224(1):39–59, 2006.

[CLBM06] Eric Cancès, Claude Le Bris, and Yvon Maday. *Mathematical methods in quantum chemistry. An introduction.* Springer-Verlag Berlin Heidelberg, 2006.

[CLPS17] Éric Cancès, Antoine Levitt, Gianluca Panati, and Gabriel Stoltz. Robust determination of maximally localized wannier functions. *Physical Review B*, 95(7):075114, 2017.

[CLY04] YS Choi, Roger Lui, and Yoshio Yamada. Existence of global solutions for the shigesada-kawasaki-teramoto model with strongly coupled cross-diffusion. *Discrete and Continuous Dynamical Systems*, 10(3):719–730, 2004.

[CP04] Wenyan Chen and Rui Peng. Stationary patterns created by cross-diffusion for the competitor–competitor–mutualist model. *Journal of Mathematical Analysis and Applications*, 291(2):550–564, 2004.

[CZWP06] Silvia Casassa, Claudio M Zicovich-Wilson, and Cesare Pisani. Symmetry-adapted localized wannier functions suitable for periodic local correlation methods. *Theoretical Chemistry Accounts: Theory, Computation, and Modeling (Theoretica Chimica Acta)*, 116(4):726–733, 2006.

[DC63] Jacques Des Cloizeaux. Orthogonal orbitals and generalized wannier functions. *Physical Review*, 129(2):554, 1963.

[DC64a] Jacques Des Cloizeaux. Analytical properties of n-dimensional energy bands and wannier functions. *Physical Review*, 135(3A):A698, 1964.

[DC64b] Jacques Des Cloizeaux. Energy bands and projection operators in a crystal: analytic and asymptotic properties. *Physical Review*, 135(3A):A685, 1964.

[Deu87] Paul Deuring. An initial-boundary-value problem for a certain density-dependent diffusion system. *Mathematische Zeitschrift*, 194(3):375–396, 1987.

[DFR08] Marco Di Francesco and Jesús Rosado. Fully parabolic keller–segel model for chemotaxis with prevention of overcrowding. *Nonlinearity*, 21(11):2715, 2008.

[DG93] Ennio De Giorgi. New problems on minimizing movements. *Ennio de Giorgi: Selected Papers*, pages 699–713, 1993.

[DGJ97] Pierre Degond, Stéphane Génieys, and Ansgar Juengel. A system of parabolic equations in nonequilibrium thermodynamics including thermal and electrical effects. *Journal de mathématiques pures et appliquées*, 76(10):991–1015, 1997.

[DJ12] Michael Dreher and Ansgar Juengel. Compact families of piecewise constant functions in lp (0, t; b). *Nonlinear Analysis: Theory, Methods & Applications*, 75(6):3072–3077, 2012.

[DJP16] N.C. Dias, C. Jorge, and J.N. Prata. One-dimensional Schrödinger operators with singular potentials: A Schwartz distributional formulation. *J. of Differential Equations*, 260(8):6548–6580, 2016.

[DLMT15] Laurent Desvillettes, Thomas Lepoutre, Ayman Moussa, and Ariane Trescases. On the entropic structure of reaction-cross diffusion systems. *Communications in Partial Differential Equations*, 40(9):1705–1747, 2015.

[DNS09] Jean Dolbeault, Bruno Nazaret, and Giuseppe Savaré. A new class of transport distances between measures. *Calculus of Variations and Partial Differential Equations*, 34(2):193–231, 2009.

[DS08] Sara Daneri and Giuseppe Savaré. Eulerian calculus for the displacement convexity in the wasserstein distance. *SIAM Journal on Mathematical Analysis*, 40(3):1104–1122, 2008.

[DT62] John Bruce Duncan and HL Toor. An experimental study of three component gas diffusion. *AIChE Journal*, 8(1):38–41, 1962.

[DT15] Laurent Desvillettes and Ariane Trescases. New results for triangular reaction cross diffusion system. *Journal of Mathematical Analysis and Applications*, 430(1):32–59, 2015.

[Dun00] Le Dung. Remarks on hölder continuity for parabolic equations and convergence to global attractors. *Nonlinear Analysis: Theory, Methods & Applications*, 41(7-8):921–941, 2000.

[EG02] Alexandre Ern and Jean-Luc Guermond. Éléments finis: théorie, applications, mise en œuvre, volume 36 of mathématiques & applications (berlin)[mathematics & applications], 2002.

[ERT84a] J. Eskin, J. Ralston, and E. Trubowiz. On isospectral periodic potential in $\mathbb{R}^n$. I. *Commun. Pure Appl. Maths.*, 37(5):647–676, 1984.

[ERT84b] J. Eskin, J. Ralston, and E. Trubowiz. On isospectral periodic potential in $\mathbb{R}^n$. II. *Commun. Pure Appl. Maths.*, 37(6):715–753, 1984.

[ES12] Robert Evarestov and Vyacheslav P Smirnov. *Site symmetry in crystals: theory and applications*, volume 108. Springer Science & Business Media, 2012.

[Esk89] G. Eskin. Inverse spectral problem for the Schrödinger equation with periodic vector potential. *Commun. Math. Phys*, 125(2):263–300, 1989.

[Eva98] L.C. Evans. *Partial Differential Equations*. Graduate studies in mathematics. American Mathematical Society, 1998.

[FK16] Shiang Fang and Efthimios Kaxiras. Electronic structure theory of weakly interacting bilayers. *Physical Review B*, 93(23):235153, 2016.

[FKDS+15] Shiang Fang, Rodrick Kuate Defo, Sharmila N. Shirodkar, Simon Lieu, Georgios A. Tritsaris, and Efthimios Kaxiras. Ab initio tight-binding hamiltonian for transition metal dichalcogenides. *Phys. Rev. B*, 92:205108, Nov 2015.

[Flo83] Gaston Floquet. Sur les équations différentielles linéaires à coefficients périodiques. *Annales de l'École Normale Supérieure*, 1883.

[FMP16] Domenico Fiorenza, Domenico Monaco, and Gianluca Panati. Construction of real-valued localized composite wannier functions for insulators. In *Annales Henri Poincaré*, volume 17, pages 63–97. Springer, 2016.

[FY01] G. Freiling and V. Yurko. *Inverse Sturm-Liouville problems and their applications*. Nova Science Publishers, 2001.

[Gaj94a] H. Gajewski. On a variant of monotonicity and its application to differential equations. *Nonlinear Anal.*, 22(1):73–80, January 1994.

[Gaj94b] Herbert Gajewski. On the uniqueness of solutions to the drift-diffusion model of semiconductor devices. *Mathematical Models and Methods in Applied Sciences*, 4(01):121–133, 1994.

[G.B46] G.Borg. Eine umkehrung der sturm-liouville eigenwertaufgabe. *Acta Math. 76*, 1946.

[GCH13] Qin Gao, Xiaoliang Cheng, and Zhengda Huang. Modified numerov's method for inverse sturm–liouville problems. *Journal of Computational and Applied Mathematics*, 253:181–199, 2013.

[Gio12] Vincent Giovangigli. Multicomponent flow modeling. *Science China Mathematics*, 55(2):285–308, 2012.

[GMSW17] Philip E. Gill, Walter Murray, Michael A. Saunders, and Elizabeth Wong. User's guide for SNOPT 7.6: Software for large-scale nonlinear programming. Center for Computational Mathematics Report CCoM 17-1, Department of Mathematics, University of California, San Diego, La Jolla, CA, 2017.

[GQZQXL08] Sun Gui-Quan, Jin Zhen, Liu Quan-Xing, and Li Li. Pattern formation induced by cross-diffusion in a predator–prey system. *Chinese Physics B*, 17(11):3936, 2008.

[GR10] Jens André Griepentrog and Lutz Recke. Local existence, uniqueness and smooth dependence for nonsmooth quasilinear parabolic problems. *Journal of Evolution Equations*, 10(2):341–375, 2010.

[GS82] Mariano Giaquinta and Michael Struwe. On the partial regularity of weak solutions of nonlinear parabolic systems. *Mathematische Zeitschrift*, 179(4):437–451, 1982.

[GZ06] F. Gesztesy and M. Zinchenko. On spectral theory for Schrödinger operators with strongly singular potentials. *Math. Nachr.*, 279(9-10):1041–1082, 2006.

[H.H76] B.Lieberman H.Hochstadt. An inverse sturm-liouville problem with mixed given data. *SIAM J. Appl. Math 34*, 1976.

[HM01] R.O. Hryniv and Y.V. Mykytyuk. 1-D Schrödinger operators with periodic singular potentials. *Methods Funct. Anal. Topology*, 7(4):31–42, 2001.

[HM03a] R.O. Hryniv and Y.V. Mykytyuk. Inverse spectral problems for Sturm-Liouville operators with singular potentials. *Inverse Problems*, 19(3):665, 2003.

[HM03b] R.O. Hryniv and Y.V. Mykytyuk. Inverse spectral problems for Sturm-Liouville operators with singular potentials. III. Reconstruction by three spectra. *J. Math. Anal. Appl.*, 284(2):626–646, 2003.

[HM04a] R.O. Hryniv and Y.V. Mykytyuk. Half-inverse spectral problems for Sturm-Liouville operators with singular potentials. *Inverse Problems*, 20(5):1423, 2004.

[HM04b] R.O. Hryniv and Y.V. Mykytyuk. Inverse spectral problems for Sturm-Liouville operators with singular potentials. II. Reconstruction by two spectra. *North-Holland Mathematics Studies*, 197:97–114, 2004.

[HM06] R.O. Hryniv and Y.V. Mykytyuk. Inverse spectral problems for Sturm-Liouville operators with singular potentials. IV. Potentials in the Sobolev space scale. *Proceedings of the Edinburgh Mathematical Society*, 49(2):309–329, 2006.

[HMPW17] Martin Herberg, Martin Meyries, Jan Pruess, and Mathias Wilke. Reaction–diffusion systems of maxwell–stefan type with reversible mass-action kinetics. *Nonlinear Analysis*, 159:264–284, 2017.

[HNP15] Luan T Hoang, Truyen V Nguyen, and Tuoc V Phan. Gradient estimates and global existence of smooth solutions to a cross-diffusion system. *SIAM Journal on Mathematical Analysis*, 47(3):2122–2177, 2015.

[HP09] Thomas Hillen and Kevin J Painter. A user's guide to pde models for chemotaxis. *Journal of mathematical biology*, 58(1-2):183–217, 2009.

[Jay57] Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

[JB02] Trachette L Jackson and Helen M Byrne. A mechanical model of tumor encapsulation and transcapsular spread. *Mathematical biosciences*, 180(1):307–328, 2002.

[JBS06] Claude Lemaréchal J.Frédéric Bonnans, Jean Charles Gilbert and Claudia Sagastizàbal. *Numerical Optimization. Theoretical and Practical Aspects*, volume 1. Springer-Verlag Berlin Heidelberg, 2006.

[JHW+15] Philip Jackson, Dimitrios Hariskos, Roland Wuerz, Oliver Kiowski, Andreas Bauer, Theresa Magorian Friedlmeier, and Michael Powalla. Properties of cu (in, ga) se2 solar cells with new record efficiencies up to 21.7%. *physica status solidi (RRL)-Rapid Research Letters*, 9(1):28–31, 2015.

[JKO98] Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the fokker–planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.

[JM13] Jeil Jung and Allan H. MacDonald. Tight-binding model for graphene π-bands from maximally localized wannier functions. *Phys. Rev. B*, 87:195450, May 2013.

[JS98] Oldiichi John and Jana Stará. On the regularity of weak solutions to parabolic systems in two spatial dimensions. *Communications in partial differential equations*, 23(7-8):437–451, 1998.

[JS12] Ansgar Juengel and Ines Viktoria Stelzer. Entropy structure of a cross-diffusion tumor-growth model. *Mathematical Models and Methods in Applied Sciences*, 22(07):1250009, 2012.

[JS13] Ansgar Juengel and Ines Viktoria Stelzer. Existence analysis of maxwell–stefan systems for multicomponent mixtures. *SIAM Journal on Mathematical Analysis*, 45(4):2421–2440, 2013.

[Jue15a] Ansgar Juengel. The boundedness-by-entropy method for cross-diffusion systems. *Nonlinearity*, 28(6):1963, 2015.

[Jue15b] Ansgar Juengel. *Entropy Methods for Diffusive Partial Differential Equations*. Springer Verlag, 2015.

[JWH+16] Philip Jackson, Roland Wuerz, Dimitrios Hariskos, Erwin Lotter, Wolfram Witte, and Michael Powalla. Effects of heavy alkali elements in cu (in, ga) se2 solar cells with efficiencies up to 22.6%. *physica status solidi (RRL)-Rapid Research Letters*, 10(8):583–586, 2016.

[JZ14] Ansgar Juengel and Nicola Zamponi. Boundedness of weak solutions to cross-diffusion systems from population dynamics. *arXiv preprint arXiv:1404.6054*, 2014.

[Kat72] T. Kato. Schrödinger operators with singular potentials. *Israel Journal of Mathematics*, 13(1):135–148, 1972.

[KF96a] Georg Kresse and Jürgen Furthmüller. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Computational materials science*, 6(1):15–50, 1996.

[KF96b] Georg Kresse and Jürgen Furthmüller. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Physical review B*, 54(16):11169, 1996.

[Kim84] Jong Uhn Kim. Smooth solutions to a quasi-linear system of diffusion equations for a certain population model. *Nonlinear Analysis: Theory, Methods & Applications*, 8(10):1121–1144, 1984.

[Kli15] Torben Klinkert. *Comprehension and optimisation of the co-evaporation deposition of Cu (In, Ga) Se2 absorber layers for very high efficiency thin film solar cells*. PhD thesis, Université Pierre et Marie Curie-Paris VI, 2015.

[Koh73] W Kohn. Construction of wannier functions and applications to energy bands. *Physical Review B*, 7(10):4388, 1973.

[Krü87] E Krüger. Symmetry-adapted wannier functions in perfect antiferromagnetic chromium. *Physical Review B*, 36(4):2263, 1987.

[KS88] Shuichi Kawashima and Yasushi Shizuta. On the normal form of the symmetric hyperbolic-parabolic systems associated with the conservation laws. *Tohoku Mathematical Journal, Second Series*, 40(3):449–464, 1988.

[KSV93] RD King-Smith and David Vanderbilt. Theory of polarization of crystalline solids. *Physical Review B*, 47(3):1651, 1993.

[KT90] H Knörrer and E Trubowitz. A directional compactification of the complex bloch variety. *Commentarii Mathematici Helvetici*, 65(1):114–149, 1990.

[Kuc16] P. Kuchment. An overview of periodic elliptic operators. *Bull. Amer. Math. Soc.*, 53(3):343–414, 2016.

[Kue96] Konrad Horst Wilhelm Kuefner. Invariant regions for quasilinear reaction-diffusion systems and applications to a two population model. *Nonlinear Differential Equations and Applications NoDEA*, 3(4):421–444, 1996.

[Kun] Jan Kuneš. 4 wannier functions and construction of model hamiltonians.

[Lad16] Mark Ladd. *Bonding, structure and solid-state chemistry*. Oxford University Press, 2016.

[LAS08] Nicola Gigli Luigi Ambrosio and Giuseppe Savaré. *Gradient Flows in Metric Spaces and in the Spaces of Probability Measures*. Birkhäuser Basel, 2008.

[Lep17] Thomas Lepoutre. *Contributions en dynamique de populations*. Habilitation à diriger des recherches, Université Claude Bernard (Lyon 1), 2017.

[LL01] E.H. Lieb and M. Loss. *Analysis*, volume 14 of *Graduate studies in mathematics*. 2001.

[LM12] Jacques Louis Lions and Enrico Magenes. *Non-homogeneous boundary value problems and applications*, volume 1. Springer Science & Business Media, 2012.

[LM13] Matthias Liero and Alexander Mielke. Gradient structures and geodesic convexity for reaction–diffusion systems. *Phil. Trans. R. Soc. A*, 371(2005):20120346, 2013.

[LM14] Thomas Lepoutre and Salome Martinez. Steady state analysis for a relaxed cross diffusion model. *Discrete and Continuous Dynamical Systems-Series A*, 2:613–633, 2014.

[LM17] Thomas Lepoutre and Ayman Moussa. Entropic structure and duality for multiple species cross-diffusion systems. *Nonlinear Analysis*, 2017.

[LN06] Dung Le and Toan Trong Nguyen. Everywhere regularity of solutions to a class of strongly coupled degenerate parabolic systems. *Communications in Partial Differential Equations*, 31(2):307–324, 2006.

[LPR12] Thomas Lepoutre, Michel Pierre, and Guillaume Rolland. Global well-posedness of a conservative relaxed cross diffusion system. *SIAM Journal on Mathematical Analysis*, 44(3):1674–1693, 2012.

[LS67] Olga A Ladyzenskaja and Vsevolod A Solonnikov. Nn ural ceva, linear and quasilinear equations of parabolic type, translated from the russian by s. smith. translations of mathematical monographs, vol. 23. *American Mathematical Society, Providence, RI*, 63:64, 1967.

[LW15] Yuan Lou and Michael Winkler. Global existence and uniform boundedness of smooth solutions to a cross-diffusion system with equal diffusion rates. *Communications in Partial Differential Equations*, 40(10):1905–1941, 2015.

[LZ05] Yi Li and Chunshan Zhao. Global existence of solutions to a cross-diffusion system in higher dimensional domains. *Discrete Contin. Dyn. Syst*, 12(2):185–192, 2005.

[Mat10] Donald M Mattox. *Handbook of physical vapor deposition (PVD) processing*. William Andrew, 2010.

[MAT16] MATLAB. *Optimization Toolbox User's Guide (R2016b)*. The MathWorks Inc., 3 Apple Hill Drive Natick, MA 01760-2098, 2016.

[Max66] J Clerk Maxwell. On the dynamical theory of gases. *Proceedings of the Royal Society of London*, 15:167–171, 1866.

[McC97] Robert J McCann. A convexity principle for interacting gases. *Advances in mathematics*, 128(1):153–179, 1997.

[MI08] Koichi Momma and Fujio Izumi. Vesta: a three-dimensional visualization system for electronic and structural analysis. *Journal of Applied Crystallography*, 41(3):653–658, 2008.

[MM08] V. Mikhaelets and V. Molyboga. One-dimensional Schrödinger operators with singular periodic potentials. *Methods Funct. Anal. Topology*, 14(2):184–200, 2008.

[MMY+12] Nicola Marzari, Arash A Mostofi, Jonathan R Yates, Ivo Souza, and David Vanderbilt. Maximally localized wannier functions: Theory and applications. *Reviews of Modern Physics*, 84(4):1419, 2012.

[Mol16] Fabien Mollica. *Optimization of ultra-thin Cu (In, Ga) Se2 based solar cells with alternative back-contacts*. PhD thesis, Université Pierre et Marie Curie-Paris VI, 2016.

[Mou16] Ayman Moussa. Some variants of the classical aubin–lions lemma. *Journal of Evolution Equations*, 16(1):65–93, 2016.

[MSV03] Nicola Marzari, Ivo Souza, and David Vanderbilt. An introduction to maximally-localized wannier functions. *Psi-K newsletter*, 57:129, 2003.

[MV97] Nicola Marzari and David Vanderbilt. Maximally localized generalized wannier functions for composite energy bands. *Physical review B*, 56(20):12847, 1997.

[MYL+08] Arash A Mostofi, Jonathan R Yates, Young-Su Lee, Ivo Souza, David Vanderbilt, and Nicola Marzari. wannier90: A tool for obtaining maximally-localised wannier functions. *Computer physics communications*, 178(9):685–699, 2008.

[Nen83] G Nenciu. Existence of the exponentially localised wannier functions. *Communications in mathematical physics*, 91(1):81–85, 1983.

[Neu49] John Von Neumann. On rings of operators. reduction theory. *Annals of Mathematics*, 50(2):401–485, 1949.

[NW06] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, New York, 2nd edition, 2006.

[OW05] Felix Otto and Michael Westdickenberg. Eulerian calculus for the contraction in the wasserstein distance. *SIAM journal on mathematical analysis*, 37(4):1227–1255, 2005.

[Pai09] Kevin J Painter. Continuous models for cell migration in tissues and applications to cell sorting via differential chemotaxis. *Bulletin of Mathematical Biology*, 71(5):1117–1147, 2009.

[Pan] A. Pankov. Introduction to spectral theory of schrödinger operators. Notes.

[PBE96] John P Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized gradient approximation made simple. *Physical review letters*, 77(18):3865, 1996.

[PBMM02] Michel Posternak, Alfonso Baldereschi, Sandro Massidda, and Nicola Marzari. Maximally localized wannier functions in antiferromagnetic mno within the flapw formalism. *Physical Review B*, 65(18):184422, 2002.

[Pie10] Michel Pierre. Global existence in reaction-diffusion systems with control of mass: a survey. *Milan Journal of Mathematics*, 78(2):417–455, 2010.

[PP08] Jacobus W Portegies and Mark A Peletier. Well-posedness of a parabolic moving-boundary problem in the setting of wasserstein gradient flows. *arXiv preprint arXiv:0812.1269*, 2008.

[PP13] Gianluca Panati and Adriano Pisante. Bloch bundles, marzari-vanderbilt functional and maximally localized wannier functions. *Communications in Mathematical Physics*, 322(3):835–875, 2013.

[PS00] Michel Pierre and Didier Schmitt. Blowup in reaction-diffusion systems with dissipation of mass. *SIAM review*, 42(1):93–106, 2000.

[PT87] J. Pöschel and E. Trubowiz. *Inverse spectral theory. Pure and applied mathematics*. Academic Press, 1987.

[PU+16] Würfel Peter, W Uli, et al. *Physics of solar cells: from basic principles to advanced concepts*. John Wiley & Sons, 2016.

[PWJ+14] Michael Powalla, Wolfram Witte, Philip Jackson, Stefan Paetel, Erwin Lotter, Roland Wuerz, Friedrich Kessler, Carsten Tschamber, Wolfram Hempel, Dimitrios Hariskos, et al. Cigs cells and modules with high efficiency on glass and flexible substrates. *IEEE Journal of Photovoltaics*, 4(1):440–446, 2014.

[Red89] Reinhard Redlinger. Invariant sets for strongly coupled reaction-diffusion systems under general boundary conditions. *Archive for Rational Mechanics and Analysis*, 108(4):281–291, 1989.

[RS78a] M. Reed and B. Simon. *Methods of modern mathematical physics. I: Functional Analysis*. Elsevier, 1978.

[RS78b] M. Reed and B. Simon. *Methods of modern mathematical physics. IV: Analysis of operators*. Elsevier, 1978.

[RS78c] Michael Reed and Barry Simon. *IV: Analysis of Operators*, volume 4. Elsevier, 1978.

[Saa03] Yousef Saad. *Iterative methods for sparse linear systems*. SIAM, 2003.

[Sak13] Rei Sakuma. Symmetry-adapted wannier functions in the maximal localization procedure. *Physical Review B*, 87(23):235109, 2013.

[San17] Filippo Santambrogio. {Euclidean, metric, and Wasserstein} gradient flows: an overview. *Bulletin of Mathematical Sciences*, 7(1):87–154, 2017.

[SB94] B Sporkmann and H Bross. Calculation of wannier functions for fcc transition metals by fourier transformation of bloch functions. *Physical Review B*, 49(16):10869, 1994.

[SE05] VP Smirnov and RA Evarestov. Symmetry analysis for localized function generation and chemical bonding in crystals: Sr zr o 3 and mgo as examples. *Physical Review B*, 72(7):075138, 2005.

[SJ95] Jana Stará and Oldrich John. Some (new) counterexamples of parabolic systems. *Commentationes Mathematicae Universitatis Carolinae*, 36(3):503–510, 1995.

[SKT79] Nanako Shigesada, Kohkichi Kawasaki, and Ei Teramoto. Spatial segregation of interacting species. *Journal of Theoretical Biology*, 79(1):83–99, 1979.

[SMV01] Ivo Souza, Nicola Marzari, and David Vanderbilt. Maximally localized wannier functions for entangled energy bands. *Physical Review B*, 65(3):035109, 2001.

[Ste71] Josef Stefan. ueber das gleichgewicht und die bewegung, insbesondere die diffusion von gasgemengen. *Sitzber. Akad. Wiss. Wien*, 63:63–124, 1871.

[SU01] VP Smirnov and DE Usvyat. Variational method for the generation of localized wannier functions on the basis of bloch functions. *Physical Review B*, 64(24):245108, 2001.

[Tem08] V. N. Temlyakov. Greedy approximation. *Acta Numerica*, 17:235–409, 2008.

[THJ05] Kristian Sommer Thygesen, Lars Bruno Hansen, and Karsten Wedel Jacobsen. Partly occupied wannier functions. *Physical review letters*, 94(2):026405, 2005.

[Tru67] N.S. Trudinger. On Harnack type inequalities and their application to quasilinear elliptic equations. *Commun. Appl. Math.*, 20(4):721–747, 1967.

[TZ11] Vladimir N. Temlyakov and Pavel Zheltov. On performance of greedy algorithms. *Journal of Approximation Theory*, 163(9):1134–1145, 2011.

[VBC79] J Von Boehm and J-L Calais. Variational procedure for symmetry-adapted wannier functions. *Journal of Physics C: Solid State Physics*, 12(18):3661, 1979.

[Vel15] O. Veliev. *Multidimensional periodic Schrödinger operator. Perturbation theory and applications.* Academic Press, 2015.

[VT08] Phan Van Tuoc. On global existence of solutions to a cross-diffusion system. *Journal of Mathematical Analysis and Applications*, 343(2):826–834, 2008.

[Wan37] Gregory H Wannier. The structure of electronic excitation levels in insulating crystals. *Physical Review*, 52(3):191, 1937.

[Wan05] Yi Wang. The global existence of solutions for a cross-diffusion system. *Acta Mathematicae Applicatae Sinica (English Series)*, 21(3):519–528, 2005.

[Wen13] Zijuan Wen. Turing instability and stationary patterns in a predator-prey systems with nonlinear cross-diffusions. *Boundary Value Problems*, 2013(1):155, 2013.

[Wie92] Michael Wiegner. Global solutions to a class of strongly coupled parabolic systems. *Mathematische Annalen*, 292(1):711–727, 1992.

[Wil70] Wilbert Wils. Direct integrals of hilbert spaces i. *Mathematica Scandinavica*, 26(1):73–88, 1970.

[W.R92] P.E.Sacks W.Rundell. Reconstruction techniques for classical inverse sturm-liouville problems. *Mathematics of computation*, 58:161–183, 1992.

[WS80] H. Wolkowicz and G.P.H. Styan. Bounds on eigenvalues using traces. *Linear Algebra Appl.*, 29:471–506, 1980.

[Yag93] Atsushi Yagi. Global solution to some quasilinear parabolic system in population dynamics. *Nonlinear Analysis: Theory, Methods & Applications*, 21(8):603–630, 1993.

[ZJ15] Nicola Zamponi and Ansgar Juengel. Analysis of degenerate cross-diffusion population models with volume filling. In *Annales de l'Institut Henri Poincare (C) Non Linear Analysis*. Elsevier, 2015.

[ZM15] Jonathan Zinsl and Daniel Matthes. Transport distances and geodesic convexity for systems of degenerate diffusion equations. *Calculus of Variations and Partial Differential Equations*, 54(4):3397–3438, 2015.