MASTER

Analysis of a diffusion-reaction system modelling formation processes of solar cells, image deblurring methods and methods for extracting interdiffusion coefficents

van der Heide, O.

Award date:
2015

Technische Universiteit
Eindhoven
University of Technology

Department of Mathematics and Computer Science
Centre for Analysis, Scientific computing and Applications

Analysis of a Diffusion-Reaction System Modelling Formation Processes of Solar Cells, Image Deblurring Methods and Methods for Extracting Interdiffusion Coefficients

Master Thesis
Industrial and Applied Mathematics

Oscar van der Heide

Supervisors:

prof.dr. I.S. Pop (Eindhoven University of Technology)
dr.habil A. Muntean (Eindhoven University of Technology)
dr.ing. J. Emmelkamp (TNO)

Eindhoven, July 2015

Abstract

In this thesis we explore several mathematical tools beneficial to modelling the formation process of CIGS-based thin film solar cells. More specifically, we present methods for deblurring experimentally obtained concentration profiles and methods for extracting interdiffusion coefficients from these (deblurred) concentration profiles. We also discuss part of a model that is being developed at TNO to describe the formation of absorber layers in CIGS-based solar cells and we present a method to numerically solve the underlying set of non-linear reaction-diffusion. The discussed model is subjected to a rigorous mathematical analysis where, under certain assumptions, existence and uniqueness of solutions to the model is shown. This thesis also includes a chapter in which the (quantum) physical principles are explained that govern the working of basic solar cells.

Part of the work presented in this thesis was performed during an internship at TNO/Solliance in Eindhoven.

Keywords: Image Deblurring, Identification of Interdiffusion Coefficients, Non-Linear Reaction-Diffusion Equations, Finite Volume Scheme, Method of Rothe.

Mathematics Subject Classication (2010): 35K55, 35K57, 46N20, 46N40, 65M08, 65M12.

Contents

1 Introduction

Solar cells are electrical devices that convert light energy into electrical energy that can be used to power our electrical devices. The most common material used as absorber layer for solar cells is crystalline silicon. A promising alternative to crystalline silicon absorber layers is to use a crystalline structure consisting of copper, indium, gallium and selenide ($Cu(In_xGa_{1-x})Se_2$), commonly referred to as CIGS. Whereas crystalline silicon absorber layers are typically required to have a thickness of around $200\mu m$ to achieve good efficiencies, CIGS absorber layers typically have a thickness of 1 to $2\mu m$ and they are considered *thin film* solar cells. Much less material is needed to produce CIGS-based solar cells meaning that CIGS-based solar cells have the potential to be cheaper than the traditional silicon-based solar cells. On top of that, thin film solar cells like the CIGS-based solar cells have the potential to be flexible.

A common method to produce CIGS-based solar cells with such high efficiencies is through a process called co-evaporation. In this process, a soda-lime glass subtrate with a conductive coating of molybdenum (back contact) is put into a vacuum environment. Copper, indium, gallium and selenium vapor are released and adhere to the substrate to form a layer of CIGS. While this process results in solar cells with high efficiencies, it is not suitable for large scale production.

On the other hand, there is the so called *two-step process*. In the *first step* a precursor is prepared by depositing layers of copper, indium and gallium onto the molybdenum coated soda-lime glass substrate. In the *second step*, the precursor is mounted in a furnace where it is *selenized.* That is, the precursor is thermally annealed in a nitrogen environment in which vapor of selenium is released. The selenium is absorbed by the precursor stack and through diffusion and reaction a layer of $Cu(In_xGa_{1-x})Se_2$ (CIGS) forms eventually. The two-step process is more suitable for large scale production but at the moment the process is not fully understood.

Part of the research at TNO/Solliance is focussed at better understanding CIGS-based solar cells and to design methods to effeciently produce CIGS-based solar cells on a large scale. Among other things, a model is being developed describing the physical phenomena that govern the formation of the CIGS absorber layers in the two-step process. This model 'starts' at the moment the precursor has been produced in step one of the process. It 'ends' after the selenization step has been completed. In the end the model will be used to simulate the formation of CIGS-absorber layers and hopefully the real production process can be taylored to produce better CIGS-based solar cells with the help of these simulations.

The model requires several parameters as input. Among others, diffusion coefficients describing the mobility of the different components in the system and reaction rates for the chemical reactions that take place to form new binary/ternary/quaterny phases are needed. Little data is available in the literature or databases on such multicomponent, multiphase systems. Hence methods are needed to obtain these parameters from experimental data obtained from measurements performed at TNO/Solliance. One of the objectives at TNO/Solliance was to develop such methods.

Furthermore, the current method used to numerically solve the equations in the model turns out to be relatively slow. Especially the the calculation of the diffusion processes take up too much time. A second goal objective at TNO/Solliance was to develop a robust numerical method that is fast and flexible in the sense that it can easily be adapted to include new components or phases when needed.

The general outline of this work is as follows:

- In Chapter 2 we discuss solid-state diffusion in more detail and present methods to extract concentration-dependent diffusion coefficients based on experimental concentration-depth

profiles. MATLAB scripts are developed that implement these methods and the results are discussed.

- At TNO/Solliance, concentration-depth profiles of CIGS layers were generated using so called Cross-section Energy Dispersive X-ray Spectroscopy measurements. The resulting profiles appear to be heavily blurred though. In Chapter 3 present methods to deblur the profiles. All the methods described in this chapter are implemented in MATLAB.
- Chapter 4 contains part of a diffusion-reaction model that is being developed by TNO/Solliance. We refer to this part of the model as the *precursor model*. The precursor model describes the physical and chemical processes within precursor after the first step of the two-step process but before the second step. We present a numerical method to solve the system of equations of the precursor model that improves the currently used numerical method at TNO/Solliance.
- In Chapter 6 the methods developed in the previous chapters are tested with data from measurements performed at TNO/Solliance.
- The final chapter, Chapter 7, contains the conclusions and suggestions.
- The Appendix covers a textbook-like introduction into the working principles of basic solar cells. We start by reviewing concepts from high school physics but then we quickly dive into the realm quantum physics and semiconductor devices. The physical concepts discussed here are also relevant to understanding the working principles of several of the measurement devices used at TNO/Solliance.

The MATLAB scripts developed during this graduation project and digital drawings made for this thesis are available upon request at oscarvanderheide@gmail.com.

2 Identification of Diffusion Coefficients

The CIGS absorber layers form during a diffusion-reaction process. To model properly this process one needs to know the relevant diffusion coefficients. In this chapter we present methods for deriving concentration dependent diffusion coefficients in multicomponent solids from experimental data. But, before we start discussing these methods, we have a look at diffusion in solids in more detail in Section 2.1. We see that there are several types of diffusion coefficients and this is related to the fact that fluxes can be expressed in different ways. As it turns out, the so called interdiffusion coefficients are the ones we will be able to derive from measurements. In Section 2.2 we describe the so called Boltzmann-Matano method for deriving interdiffusion from concentration profiles for solids consisting of two different atomic components only. We also discuss some refinements of this method. In Section 2.3 we discuss (the problems) with multicomponent diffusion and partially extend the methods discussed in Section 2.2 to systems with three components.

The main sources used in writing this chapter are the books *Kinetics of Materials* by Balluffi et al [3] and *Thermodynamics, Diffusion and the Kirkendall Effect in Solids* by Paul et al [43].

2.1 Diffusion in Crystalline Solids

2.1.1 Diffusion Mechanisms in Crystalline Solids

Consider a crystalline solid. We think of such a solid as a large three dimensional lattice. The lattice sites are occupied by the different components making up the solid. Different components may be constrained to different lattice sites. For example, in an *interstitional* solid, the atoms of one component are much smaller than the atoms of other components and the former may occupy the spaces in between the latter. In *substitutional* solids all lattice sites are equivalent in this respect. In this chapter we only focus our attention on substitutional alloys.

Atoms diffuse through a crystalline solid by jumping between lattice sites. When an atom leaves a lattice site, it must be replaced in order to preserve the amount of lattice sites within the solid (but it does not necessarily have to be replaced be another atom, it can also be replaced by a vacanacy as will be explained in the next paragraph). It was believed at first that diffusion of atoms in a substitional alloy is mainly facilitated by a *direct exchange mechanism* or a *ring mechanism* [3]. Under a direct exchange mechanism an atom diffuses within the lattice by swapping places with another atom. A ring mechanism is similar but involves more than two - say N - atoms. Atom 1 jumps into the site of atom 2, atom 2 jumps into the site of atom 3, all the way up to atom N which jumps into the site of atom 1. One can imagine that the energy barrier to diffusion under these two mechanisms is high.

Through the revolutionary work of Ernest Kirkendall it was discovered that the dominant diffusion mechanism in substitutional alloys is in fact a *vacancy mechanism* [31]. This mechanism can be explained as follows. Thermodynamic considerations dictate that solids at temperatures above absolute zero will always have some unoccupied lattice sites. We refer to unoccupied lattice sites as *vacancies*. Neighbouring atoms can jump into such vacancies. Compared with direct exchange and ring mechnisms the energy barrier for diffusion by a vacancy mechanism is much lower. When an atom jumps into a vacancy, it leaves a behind vacancy itself. We could interpret this as a vacancy diffusing through the crystalline solid. In fact, we could view the vacancies as one of the components making up the solid.

2.1.2 Fick's Law and Onsager's Transport Equations

Now that we know how atoms can diffuse through a crystalline solid, let us look at the *driving forces* behind diffusion. Consider a solid consisting of n different interacting atomic components. We refer to this setup as an n-component system. Let C_i denote the concentration of component i ($[C_i] = [\text{mol}/m^3]$). The flux F_i ($[F_i] = \text{mol}/(m^2 s)$) of component i across some unit section within the solid is defined as

$$F_i := v_i C_i,$$

where v_i is velocity of component i relative to the section across which we are considering the flux. The velocity v_i should be interpreted as a mean atomic velocity. It is important to note that v_i, and hence F_i, depend on how the section across which we are measuring the flux moves relative to the solid. For example, think of a situation where the section is fixed with respect to the laboratory in which we are doing flux measurements. If you push a solid across this section, then surely the above definition will give rise to fluxes - even if there is no diffusion process going on within the solid! That is, F_i possibly includes a general *bulk flux* due to *bulk velocity* as well. And we are not interested in this bulk flux - we are only interested in the flux due to diffusion processes. Therefore we need to define the sections across which we are measuring fluxes in such a way that the bulk flux is eliminated. This is by no means a trivial task and, unfortunately, there is no single correct way to do this. We will come back on this issue in Section 2.1.4. For now, assume that we have chosen our sections in such a way that each F_i represents a diffusion flux. According to Fick's law [22], the flux F_i then varies linearly with its own concentration gradient. That is,

$$F_i = -D_i \nabla C_i,$$

where D_i is a symmetric positive definite second order diffusion tensor (with units $[D_i] = m^2/s$). It is important to note that the components of D_i are in general not constants. They may very well depend on the thermodynamic quantities like concentrations C_i of the different components, the pressure P and on the temperature θ. In this chapter we will generally assume P and θ to be constant though. Only in the last section we will discuss the temperature dependence of diffusion coefficients.

Fick wasn't the only one who proposed a linear relationship between the flux of a component and some *thermodynamic driving force.*[1] For example, Ohm's law states that the current - that is, the flux of charged particles - varies linearly with a gradient in electric potential. Similarly, Fourier's law states that the heat flux in a system varies linearly with a temperature gradients. Lars Onsager formulated a thermodynamical theory for systems undergoing irreversible (that is, entropy producing) processes, generalizing all of the above into what is now called Onsager's transport equations [40, 41]. More specifically, the theory states that the material flux of component i relative to some chosen reference frame can be written as

$$F_i = \sum_j L_{ij} X_j.$$

Here X_j are *thermodynamic forces* and L_{ij} are the corresponding *transport coefficients*. Examples of thermodynamic forces are temperature, pressure or electrical potential gradients and, for us the most important forces, gradients in *chemical potentials* of the different species. In what follows we restrict ourselves to systems in which the gradients in chemical potentials are the only driving forces. Let's denote the chemical potentials by μ_i. Then the transport equations reduce to

$$F_i = \sum_{j=1}^{n} L_{ij} \nabla \mu_j.$$

[1] Strictly speaking, the concentration gradients are not forces because they have the wrong units. But, as we will see in equation (2.2), concentration gradients can be related to the gradient of chemical potentials, which do have the correct units (and this is compensated by the coefficients in front of them having different units).

Note that chemical potential gradients are independent of the choice of sections we use to express the fluxes F_i. Because - of course - the fluxes F_i do depend on the choice of these sections, it follows that the transport coefficients L_{ij} depend on them as well.

One question that may arise at this point is how the Onsager Transport Equations and Fick's Law are related to one another. They both claim to describe the same process but at first sight they look a bit different. As it turns out, by making several assumptions and an appropriate choice of reference frame we can recover Fick's law from the Onsager transport equations. We will do this for binary systems first. In later sections we will look at systems with more than two atomic components. But, before we do all of this, it is useful to look at some general thermodynamic relations first.

2.1.3 Thermodynamic Relations

Consider a system of n different atomic components. Let the vacancies be the $(n+1)$-th component. Let N_i be the molar fraction of component i (note that N_i is dimensionless). Assuming that the atomic fraction of vacancies is neglible compared to the atomic fraction of the other components, we can say that

$$\sum_{i=1}^{n} N_i = 1.$$

The molar volume V_{mol} at a given composition in the solid is then given by

$$V_{\text{mol}} = \sum_{i=1}^{n} N_i V_i,$$

where V_i is the partial molar volume of component i (with $[V_i] = m^3/\text{mol}$). We assume that all the V_i are *constants* and are equal to the molar volume of the pure component i. In that case, the total volume of the solid does not change under a diffusion process. The reader is referred to the books by Balluffi et al [3] and De Groot and Mazur [16] for a proof of this statement.

The molar fractions of the components can be related to concentrations by

$$C_i = \frac{N_i}{V_{\text{mol}}} = \frac{N_i}{\sum_{i=1}^{n} N_i V_i}. \tag{2.1}$$

Using the above relations we see that

$$\sum_{i=1}^{n} C_i = \frac{\sum_{i=1}^{n} N_i}{V_{\text{mol}}} = \frac{1}{V_{\text{mol}}}$$

and

$$\sum_{i=1}^{n} C_i V_i = \sum_{i=1}^{n} \frac{N_i V_i}{V_{\text{mol}}} = 1.$$

Following Balluffi et al [3], we assume the chemical potentials and concentrations of the components to be related by

$$\mu_i = \mu_i^0 + k\theta \ln\left(\gamma_i V_{\text{mol}} C_i\right).$$

Here k is the so called Boltzmann-constant, θ is the temperature, γ_i is the so called activity-coefficient of component i and μ_i^0 is a constant. Furthermore note that gradients $\nabla\mu_i$ and ∇C_i

are then related by

$$\begin{aligned}
\nabla\mu_i &= \nabla\mu_i^0 + k\theta\nabla\left[\ln\left(\gamma_i V_{\text{mol}} C_i\right)\right] \\
&= k\theta\nabla\left[\ln\left(\gamma_i\right) + \ln\left(V_{\text{mol}}\right) + \ln\left(C_i\right)\right] \\
&= k\theta\left[\nabla\ln\left(\gamma_i\right) + \nabla\ln\left(V_{\text{mol}}\right) + \nabla\ln\left(C_i\right)\right] \\
&= k\theta\left[\frac{\partial\ln\left(\gamma_i\right)}{\partial\ln\left(C_i\right)}\nabla\ln\left(C_i\right) + \frac{\partial\ln\left(V_{\text{mol}}\right)}{\partial\ln\left(C_i\right)}\nabla\ln\left(C_i\right) + \nabla\ln\left(C_i\right)\right] \\
&= k\theta\left[\frac{\partial\ln\left(\gamma_i\right)}{\partial\ln\left(C_i\right)} + \frac{\partial\ln\left(V_{\text{mol}}\right)}{\partial\ln\left(C_i\right)} + 1\right]\nabla\ln\left(C_i\right) \\
&= k\theta\left[\frac{\partial\ln\left(\gamma_i\right)}{\partial\ln\left(C_i\right)} + \frac{\partial\ln\left(V_{\text{mol}}\right)}{\partial\ln\left(C_i\right)} + 1\right]\frac{1}{C_i}\nabla C_i.
\end{aligned} \tag{2.2}$$

Finally, there is the so called *Gibbs-Duhem relation* (which can be found, for example, in the article by Brady[5]). In our situation of constant temperature and pressure this relation can be expressed as

$$\sum_{i=1}^{n+1} C_i\nabla\mu_i = 0. \tag{2.3}$$

The Gibbs-Duhem relation is a way of stating that in a system of $n+1$ components there are only n independent chemical potential gradients. Now we will also make the general assumption that our solids have sufficient defects (edge dislocations, grain boundaries, etcetera) that can act as sources or sinks for vacancies to keep the chemical potential μ_V constant. But then the gradient $\nabla\mu_V$ is zero and equation (2.3) reduces to

$$\sum_{i=1}^{n} C_i\nabla\mu_i = 0, \tag{2.4}$$

leading to the conclusion that for the n different atomic components there only $n-1$ independent chemical potential gradients.

2.1.4 Reference Frames for Diffusion Couples

We need to be careful in how we decide to measure fluxes because we want to eliminate the influences of general bulk flow. Unfortunately there is no unique way to do this. Before we make specific choices in this respect it is convenient to note that in this chapter we will only be working with so called *diffusion couples* [17]. A diffusion couple consists of two solid beams that are brought in contact with each other. The beams may consist of different atomic components but, at least initially, there are no concentration gradients present within each beam. As a result of this assumption, there will only be concentration gradients in the direction normal to the initial contact plane. Assuming diagonal diffusion tensors, Fick's law tells us that the diffusion fluxes point in the direction normal to the initial contact plane as well. In other words, the diffusion problem can be considered as one-dimensional in this case and the diffusion tensors reduce to scalar quantities. With this particular setup, the sections we use to express the fluxes will always be parallel to the initial contact plane. Following Brady [5] and De Groot and Mazur [16], we discuss three different ways of defining the sections:

- First of all, we may try to express fluxes in terms of sections that are at a fixed distance relative to one of the ends of the diffusion couple. That is, we choose one of the ends as the origin and then $F_i(x)$ gives the flux across a section a distance x away from the origin. This surely eliminates the influence of some external forces - like someone pushing

against the solid - acting on the solid as a whole because the movement of the sections themselves cancel against the movement of the solid. Because this setting coincides with how most experiments are performed in real laboraties we will refer to this setup as having measured the fluxes using *laboratory-fixed section*. We will denote such fluxes by $\check{F}_i$. The corresponding diffusion coefficients will be denoted by $\check{D}_{ij}$.
Problems arise with laboratory-fixed sections in situations where the solid happens to shrink or expand during the diffusion process. In particular, the fluxes may then depend on which end of the diffusion couple is fixed in the laboratory.

- As a second choice, it may seem more natural to attach sections to specific lattice planes within the (crystalline) solid. To illustrate this, imagine yourself sitting at a lattice plane, counting all of the atoms passing by. If somehow the lattice plane is able to move relative to the ends of the diffusion couple - like when volume changes occur - then so will our section. Fluxes defined in this manner are called *instrinsic fluxes* and they will be denoted by $\hat{J}_i$ [9]. The corresponding diffusion coefficients are called *intrinsic diffusion coefficients* and they will be denoted by $\hat{D}_{ij}$. Even though the intrinsic fluxes seem to measure diffusion in its most pure form, measuring the intrinsic diffusion coefficients turns out to be problematic because it requires keeping track of specific lattice planes within the solid.

- Thirdly, we could try to be more specific about the bulk velocity itself. To this end, let v represent the bulk velocity. Given the velocities v_i of the separate components with respect to some section, what should the bulk velocity be (with respect to the same section)? Surely v would have to be some weighted average of the velocities of the components, i.e.

$$v = \sum_i \chi_i v_i,$$

where the $\chi_i \in [0,1]$ are dimensionless quantities with the property that $\sum_{i=1}^n \chi_i = 1$. Whatever we choose for the weights χ_i, we then define the *diffusion flux* of component i to be $C_i(v_i - v)$. That is, we look at how the flux of i differs from the flux of the bulk. Because v and v_i are expressed in terms of the same sections, the effects of these specific sections cancel out. In other words, the expression $C_i(v_i - v)$ is independent of how we decide to measure the v_i. We will refer to $C_i(v_i - v)$ as the *interdiffusion flux* of component i and denote it by $\tilde{F}_i$. The corresponding *interdiffusion coefficients* will be denoted by $\tilde{D}_{ij}$.
An important property of interdiffusion fluxes is that they are not all independent. Indeed, note that

$$\sum_{i=1}^n \frac{\chi_i}{C_i}\tilde{F}_i = \sum_{i=1}^n \chi_i(v_i - v) = \sum_{i=1}^n \chi_i v_i - v\sum_{i=1}^n \chi_i = v - v = 0. \tag{2.5}$$

As a result, there are only $n-1$ independent interdiffusion fluxes.
As for the weights χ_i, several choices could be made. For example we could take $\chi_i := N_i$, i.e. the atomic fractions of the components. Alternatively - and this is the choice we will settle for - we could work volume fractions $\chi_i := V_i C_i$. In the latter case, note that equation (2.5) reduces to

$$\sum_{i=1}^n \frac{\chi_i}{C_i}\tilde{F}_i = \sum_{i=1}^n V_i \tilde{F}_i = 0.$$

That is, the fluxes are defined in terms of sections across which there is no net volume flux. We refer to this setup as using *volume-fixed sections*.

At this point, the interdiffusion fluxes seem a bit mysterious and it may be difficult to relate them to real experiments. Luckily, as it turns out, if we make the assumption that each of the partial molar volumes V_i is constant, then it can be shown that the total volume of the diffusion couple remains fixed. In that case the fluxes $\check{F}_i$ and $\tilde{F}_i$ happen to coincide and hence also the

diffusion coefficients $\check{D}_{ij}$ and $\tilde{D}_{ij}$ coincide (see Chapter 3 of the textbook by Balluffi et al [3] for a derivation). We get the best of both worlds: a practicle way to measure fluxes $\tilde{F}_i = \check{F}_i$ and additional relation $\sum_{i=1}^{n} V_i \tilde{F}_i = 0$ between them.

Now we look at what can be said about Onsager's transport equations for diffusion couples in the case of lattice-fixed and volume-fixed sections. We will work with *binary* diffusion couples only. In such a system we have two different atomic components, say A and B. We will also treat the vacancies in the solid as a separate component denoted V.

2.1.5 Onsager's Transport Equations for Binary Intrinsic Diffusion

If we attach our sections to the lattice planes inside of the diffusion couple, the Onsager's transport equations for the intrinsic fluxes in the binary diffusion couple can be expressed as

$$\begin{pmatrix} \hat{F}_A \\ \hat{F}_B \\ \hat{F}_V \end{pmatrix} = \begin{pmatrix} \hat{L}_{AA} & \hat{L}_{AB} & \hat{L}_{AV} \\ \hat{L}_{BA} & \hat{L}_{BB} & \hat{L}_{BV} \\ \hat{L}_{VA} & \hat{L}_{VB} & \hat{L}_{VV} \end{pmatrix} \begin{pmatrix} \nabla\mu_A \\ \nabla\mu_B \\ \nabla\mu_V \end{pmatrix}.$$

Now remember that we assumed $\nabla\mu_V = 0$, i.e. that there are sufficient sources and sinks for vacancies to keep the chemical potential constant through the solid. Then we see that expression for the intrinsic diffusion fluxes reduce to

$$\begin{pmatrix} \hat{F}_A \\ \hat{F}_B \\ \hat{F}_V \end{pmatrix} = \begin{pmatrix} \hat{L}_{AA} & \hat{L}_{AB} \\ \hat{L}_{BA} & \hat{L}_{BB} \\ \hat{L}_{VA} & \hat{L}_{VB} \end{pmatrix} \begin{pmatrix} \nabla\mu_A \\ \nabla\mu_B \end{pmatrix}. \tag{2.6}$$

Now suppose an atom of type A crosses our lattice plane. When this movement is mediated by a vacancy mechanism, a vacancy crosses the lattice plane in the opposite direction. Similarly, in the rare case that the movement of the A atom is mediated by a direct exchange or ring mechanism, an atom of type B must cross the lattice plane in the opposite direction (we don't count the exchange of atoms of the same type as diffusion). In other words, the instrinsic fluxes must satisfy

$$\hat{F}_A = -\hat{F}_B - \hat{F}_V.$$

Alternatively, we could say that the total flux of A and B atoms must be opposed by the flux of vacancies because

$$\hat{F}_A + \hat{F}_B = -\hat{F}_V. \tag{2.7}$$

Using this relation between the intrinsic fluxes, it follows from equation 2.6 that transport coefficients for the vacancies can be expressed in terms of the transport coefficients for the components A and B:

$$\begin{aligned} \hat{L}_{VA} &= -(\hat{L}_{AA} + \hat{L}_{AB}), \\ \hat{L}_{VB} &= -(\hat{L}_{BA} + \hat{L}_{BB}). \end{aligned}$$

Instead of the original nine coefficient, we see that at this point only four coefficients are needed to describe the system. Using the Gibbs-Duhem relation (see equation (2.4)), which in this case states that

$$C_A\nabla\mu_A + C_B\nabla\mu_B = 0,$$

we can further rewrite the expressions for the fluxes as

$$\begin{aligned} \hat{F}_A &= \left(\hat{L}_{AA} - \frac{C_A}{C_B}\hat{L}_{AB}\right)\nabla\mu_A, \\ \hat{F}_B &= \left(\hat{L}_{BB} - \frac{C_B}{C_A}\hat{L}_{BA}\right)\nabla\mu_B. \end{aligned}$$

Following the book of Balluffi et al [3] we can rewrite the chemical potential gradients in terms of the corresponding concentration gradients and find that

$$\begin{aligned}\hat{F}_A &= \left[kT\left(\frac{\hat{L}_{AA}}{C_A}-\frac{\hat{L}_{AB}}{C_B}\right)\left(1+\frac{\partial \ln \gamma_A}{\partial \ln C_A}+\frac{\partial \ln V_{\text{mol}}}{\partial \ln C_A}\right)\right]\nabla C_A,\\ \hat{F}_B &= \left[kT\left(\frac{\hat{L}_{BB}}{C_B}-\frac{\hat{L}_{BA}}{C_A}\right)\left(1+\frac{\partial \ln \gamma_B}{\partial \ln C_B}+\frac{\partial \ln V_{\text{mol}}}{\partial \ln C_B}\right)\right]\nabla C_B.\end{aligned}$$

Here γ_i is the so called activity coefficient for component i and V_{mol} is the molar volume, but they are not important at this point. What's important to note here is that if we denote the terms in brackets by $-\hat{D}_A$ and $-\hat{D}_B$ respectively then we see that Fick's law has been recovered!

There is a problem though. The intrinsic diffusion fluxes $\hat{F}_A$ and $\hat{F}_B$ do not necessarily cancel out. And if they don't, we know from equation 2.7 that a net flux of vacancies arises within the material. These vacancies can be annihilited (or created) by defects like edge dislocations within the diffusion couple. When this happens it is possible for lattice planes in the diffusion couple to move relative to the ends of the couple. In other words, our reference frame can move within the diffusion couple. This behaviour was first observed by Ernest Kirkendall by placing so called inert markers in the couple that are fixed to a lattice plane [31]. When the diffusion process started, the inert markers were observed to move relative to the ends of the diffusion couple. It was definite proof that the dominant diffusion mechanism is in fact a vacancy mechanism. If diffusion only occurs by a direct exchange or ring mechanism, then it would follow that $\hat{F}_A = -\hat{F}_B$ and there would be no such thing as a Kirkendall effect.

Now we know that lattice planes can move within the diffusion couple. And different lattice plane possibly move with different speeds. This makes a lattice-fixed frame impractical for describing the diffusion process. Furthermore, we see that each of the intrinsic fluxes $\hat{F}_A$ and $\hat{F}_B$ have their own diffusion coefficient. But there is really only *one* diffusion process going on: the process which describes the mixing of atoms of components A and B. Why would we need two different coefficients to describe this mixing?

2.1.6 Onsager's Transport Equations for Binary Interdiffusion

Instead of working with lattice-fixed sections, let's have a look at what happens when we consider volume-fixed sections instead. In a completely similar way as for the lattice-fixed sections we find that

$$\tilde{F}_A = -\tilde{D}_A \nabla C_A, \quad \text{and} \quad \tilde{F}_B = -\tilde{D}_B \nabla C_B.$$

This time, we also know from equation (2.5) that

$$V_A \tilde{F}_A + V_B \tilde{F}_B = 0.$$

But then we see that

$$\begin{aligned}-\tilde{D}_B \nabla C_B &= \tilde{F}_B\\ &= -\frac{V_A}{V_B}\tilde{F}_A\\ &= \frac{V_A}{V_B}\tilde{D}_A \nabla C_A\\ &= \frac{V_A}{V_B}\tilde{D}_A \nabla\left(\frac{1-V_B C_B}{V_A}\right)\\ &= -\tilde{D}_A \nabla C_B.\end{aligned}$$

In other words, the diffusion coefficients $\check{D}_A$ and $\check{D}_B$ are the same and we can simply define $\check{D} := \check{D}_A = \check{D}_B$. That's nice! If we assume constant partial molar volumes V_i then we also know that $\tilde{D}$ is the correct diffusion coefficient to be used in a laboratory-fixed setting.

Note that we forget all about the flux of vacancies at this point. But that's no problem since we assume the vacancies to be in chemical equilibrium throughout the diffusion couple anyways. It is really only the atomic components that we wish to describe.

2.1.7 Darken's Equations

At this point it should also be noted that the interdiffusion and intrinsic diffusion coefficients are related in the binary case by *Darken's second equation* [51, 10]:

$$\begin{aligned}\tilde{D} &= C_A V_A \hat{D}_B + C_B V_B \hat{D}_A \\ &= C_A V_A \hat{D}_B + (1 - C_A V_A)\hat{D}_A \\ &= N_A \hat{D}_B + (1 - N_A)\hat{D}_A.\end{aligned}$$

Note that both the intrinsic diffusion coefficients $\hat{D}_A$ and $\hat{D}_B$ - and hence $\tilde{D}$ - will generally be concentration dependent. Since $V_A C_A + V_B C_B = 1$, i.e. the concentrations are related by one another, we can either say that $\tilde{D} = \tilde{D}(C_A)$ or $\tilde{D} = \tilde{D}(C_B)$.

We conclude that in a binary diffusion couple there are two intrinsic diffusion coefficients and one interdiffusion coefficient. This behaviour generalizes to couples consisting of more than two components. In Section 2.3 we will find that in general a couple of n components requires $n(n-1)$ intrinsic diffusion coefficients and $(n-1)^2$ interdiffusion coefficients. But first, let's have a look at how one can extract the concentration dependent interdiffusion coefficient $\tilde{D}$ for binary systems from experimental measurements.

2.2 Extracting Interdiffusion Coefficients in Binary Systems

2.2.1 Boltzmann-Matano Method

Suppose we create a diffusion couple like before by joining together two solid beams. Each beam consists entirely out of atoms of type A and B. Initially the concentrations of both the A and B atoms are uniform throughout both beams separately. Let us say that the concentration of atomic component i in the left beam is given by C_i^L and in the right beam it is given by C_i^R. Furthermore let's say that the initial interface between the two beams is positioned at $x = 0$ in a laboratory-fixed reference frame. We make the *very important assumption* that the left and right ends of the diffusion couple remain unchanged during the process. That is, $C_i(-\infty, t) = C_i^L$ and $C_i(\infty, t) = C_i^R$ at all times $t \geq 0$. In other words, we assume we are dealing with a semi-infinite diffusion couple. We also make the assumption of constant partial molar volumes V_i again so that the fluxes and diffusion coefficients are in fact interdiffusion fluxes and interdiffusion coefficients. Fick's second law (or alternatively, the *continuity equation*, see Chapter (4) for more details) then tells us that the time evolution of the concentration of the atomic component i is described by the partial differential equation

$$\begin{aligned}\frac{\partial C_i}{\partial t} &= -\frac{\partial \tilde{F}_i}{\partial x} \\ &= \frac{\partial}{\partial x}\left(\tilde{D}(C_i)\frac{\partial C_i}{\partial x}\right) \\ &= \frac{\partial \tilde{D}(C_i)}{\partial C_i}\left(\frac{\partial C_i}{\partial x}\right)^2 + \tilde{D}(C_i)\frac{\partial^2 C_i}{\partial x^2}.\end{aligned}$$

In what follows, we make the arbitrary choice to express everything in terms of the concentration C_B.

Our goal in this section will be to derive expressions for $\tilde{D}$ as a function of C_B that can be evaluated from concentration profiles. A concentration profile is understood to be a plot of the concentration C_B versus the distance x from the origin of the laboratory-fixed reference frame at some *fixed time* t^*. In other words, the concentration C_B and hence the flux $\tilde{F}_B$ are functions of x only with t acting more like a parameter. Unless necessary, we will not explicitly denote the dependence of C_B and $\tilde{F}_B$ on t.

The starting point for deriving expressions for $\tilde{D}$ is the so called Boltzmann-Matano method [4, 36]. The general idea of this method is as follows. Suppose that at position x^* (and time t^*) the concentration of component B is given by C_B^*. From the expression for the flux

$$\tilde{F}_B(x^*) = -\tilde{D}(C_B^*)\left(\frac{\partial C_i}{\partial x}\right)_{x^*},$$

it follows that

$$\tilde{D}(C_B^*) = -\frac{\tilde{F}_B(x^*)}{\left(\frac{\partial C_B}{\partial x}\right)_{x^*}}. \tag{2.8}$$

Now the concentration gradient is something we can 'easily' measure from concentration profiles (there are some issues, but more on that later). If somehow we could measure the flux at x^* from the concentration profile as well, we would be able to compute $\tilde{D}$ for the concentration C_B^*. To measure a flux, one would normally have to perform two measurements, one quickly after the other, and then somehow count how many moles have crossed particular a section in the time between the two measurements. In other words, there is a time and a space component involved. At first sight, a single concentration profile cannot reveal anything about this time component. However, as it turns out, time and space are not entirely independent of one another in diffusion processes. This is revealed - and this is the first step in the Boltzmann-Matano method - by introducing a new variable $\lambda = (x - x_M)/t^{1/2}$.[2] Here x_M refers to the location of the so called *Matano plane* but more on that later. Using the chain rule we see that

$$\frac{\partial}{\partial t} = \frac{\partial \lambda}{\partial t}\frac{d}{d\lambda} = -\frac{1}{2}\frac{x - x_M}{t^{3/2}}\frac{d}{d\lambda} = -\frac{1}{2}\frac{\lambda}{t}\frac{d}{d\lambda}, \tag{2.9}$$

$$\frac{\partial}{\partial x} = \frac{\partial \lambda}{\partial x}\frac{d}{d\lambda} = \frac{1}{t^{1/2}}\frac{d}{d\lambda}. \tag{2.10}$$

With these relations, the diffusion equation

$$\left(\frac{\partial C_B}{\partial t}\right)_{x,t} = \frac{\partial}{\partial x}\left(\tilde{D}(C_i(x,t))\left(\frac{\partial C_i}{\partial x}\right)_{x,t}\right)$$

transforms into

$$\begin{aligned} -\frac{1}{2}\frac{\lambda}{t}\frac{dC_B}{d\lambda}(\lambda) &= \frac{1}{t^{1/2}}\frac{d}{d\lambda}\left(\tilde{D}(C_B(\lambda))\frac{1}{t^{1/2}}\frac{dC_B}{d\lambda}(\lambda)\right) \\ &= \frac{1}{t}\frac{d}{d\lambda}\left(\tilde{D}(C_B(\lambda))\frac{dC_B}{d\lambda}(\lambda)\right). \end{aligned}$$

After multiplying both sides by t - and leaving out the explicit dependence of the variables on λ for shorter notations - we arrive at

$$-\frac{1}{2}\lambda\frac{dC_B}{d\lambda} = \frac{d}{d\lambda}\left(\tilde{D}(C_B)\frac{dC_B}{d\lambda}\right). \tag{2.11}$$

[2] That is, we start looking for a similarity solution. It should be stressed that similarity solution only exists for initial data where the concentration is constant on both sides of the origin - where in this case the origin is the location of the initial contact plane.

Note that we are dealing with an ordinary differential equation instead of a partial differential equation now! The next step is to integrate the whole equation with respect to λ. Before we do this, remember that the ends of the diffusion couple remain unchanged during the diffusion process. In terms of λ, we could say that $C_B(\lambda) \to C_B^L$ and $C_B(\lambda) \to C_B^R$ as $\lambda \to \infty$. Now let λ^* be some reference value for λ and define $C_B^* := C_B(\lambda^*)$. Then, if we integrate the left-hand side of equation 2.11 between $\lambda = -\infty$ and $\lambda = \lambda^*$ and doing a substitution of variables we see that [3]

$$\begin{aligned}\int_{-\infty}^{\lambda^*} \left[-\frac{1}{2}\lambda \frac{dC_B}{d\lambda}\right] d\lambda &= -\frac{1}{2}\int_{C_B(-\infty)}^{C_B(\lambda^*)} \lambda(C_B) dC_B \\ &= -\frac{1}{2}\int_{C_B^L}^{C_B^*} \lambda(C_B) dC_B.\end{aligned}$$

Note that it is only possible to perform the substituion of variables if C_B is a monotonically increasing or decreasing function (otherwise λ cannot be written as function of C_B).

Now on the other hand, integrating the right-hand side of equation 2.11 gives

$$\int_{-\infty}^{\lambda^*} \left[\frac{d}{d\lambda}\left(\tilde{D}(C_B)\frac{dC_B}{d\lambda}\right)\right] d\lambda = \left(\tilde{D}(C_B)\frac{dC_B}{d\lambda}\right)_{\lambda^*} - \left(\tilde{D}(C_B)\frac{dC_B}{d\lambda}\right)_{-\infty}.$$

Since the ends of the diffusion couple have not yet been affected at the specific time t^* at which the concentration profile is produced, the gradient of C_B vanishes as one approaches $-\infty$ and ∞. In particular,

$$\left(\tilde{D}(C_B)\frac{dC_B}{d\lambda}\right)_{-\infty} = 0.$$

After equating the two integrals again we find that

$$\tilde{D}(C_B^*)\left(\frac{dC_B}{d\lambda}\right)_{\lambda^*} = -\frac{1}{2}\int_{C_B^L}^{C_B^*} \lambda(C_B) dC_B.$$

and - assuming $\left(\frac{dC_B}{d\lambda}\right)_{\lambda^*}$ to be non-zero - we obtain an expression for $\tilde{D}(C_B^*)$:

$$\tilde{D}(C_B^*) = -\frac{1}{2\left(\frac{dC_B}{d\lambda}\right)_{\lambda^*}} \int_{C_B^L}^{C_B^*} \lambda(C_B) dC_B. \tag{2.12}$$

It would be more convenient to express the right-hand side in terms of the variables x and t again. To do this, remember that a concentration profile is assumed to be generated at some fixed time t^*. Then, for a given value of λ, the Boltzmann transformation tells us that x can be expressed as

$$x = \sqrt{t^*}\lambda + x_M.$$

[3] To make this a bit more precise, write $C_B = \phi(\lambda)$ and suppose ϕ can be inverted (that is, λ can be written as a function of C_B) and denote the inverse by f. Using integration by substitution we see that

$$\begin{aligned}-\frac{1}{2}\int_{-\infty}^{\lambda^*} \lambda \frac{dC_B}{d\lambda} d\lambda &= -\frac{1}{2}\int_{-\infty}^{\lambda^*} f(\phi(\lambda))\phi'(\lambda) d\lambda \\ &= -\frac{1}{2}\int_{\phi(-\infty)}^{\phi(\lambda^*)} f(C_B) dC_B \\ &= -\frac{1}{2}\int_{C_B^L}^{C_B^*} \lambda dC_B.\end{aligned}$$

Because t^* and x_M are constants, it follows that if λ is a function of C_B, then so is x. Hence the integral in equation (2.12) can be written as

$$\int_{C_B^L}^{C_B^*} \lambda(C_B) dC_B = \frac{1}{\sqrt{t^*}} \int_{C_B^L}^{C_B^*} x(C_B) - x_M dC_B.$$

Furthermore, using equation (2.10) we see that

$$\left(\frac{dC_B}{d\lambda}\right)_{\lambda^*} = \sqrt{t^*} \left(\frac{\partial C_B}{\partial x}\right)_{x^*}.$$

After substituting this into equation (2.12) we see that $\tilde{D}(C^*)$ can be expressed as:

$$\tilde{D}(C_B^*) = -\frac{1}{2t^* \left(\frac{dC_B}{dx}\right)_{x^*}} \int_{C_B^L}^{C^*} x(C_B) - x_M dC_B. \tag{2.13}$$

We see that the diffusion coefficient has been expressed in terms that, at least in principle, can be evaluated from a concentration profile. Comparing equation (2.13) with equation (2.8) reveals that the flux $\tilde{F}_B$ can be expressed as

$$\tilde{F}_B(x^*) = \frac{1}{2t^*} \int_{C_B^L}^{C^*} x(C_B) - x_M dC_B. \tag{2.14}$$

But what about the x_M? The value x_M refers to the position of the so called Matano plane and it determines a reference value for computing the integral.[4] The Matano plane is in fact the initial contact plane of the diffusion couple. As the diffusion process continues, it is possible for this Matano plane to move with the couple. At the particular time t^* at which the given concentration profile is generated it can be found by solving the equation

$$\int_{-\infty}^{x_M} \left(C_B(x) - C_B^L\right) dx = \int_{x_M}^{\infty} \left(C_B^R - C_B(x)\right) dx$$

for x_M. This procedure is illustrated in the left picture of figure (2.1). After having found the Matano plane, the integral in equation (2.13) can be evaluated, as illustrated in the right picture of figure (2.1).

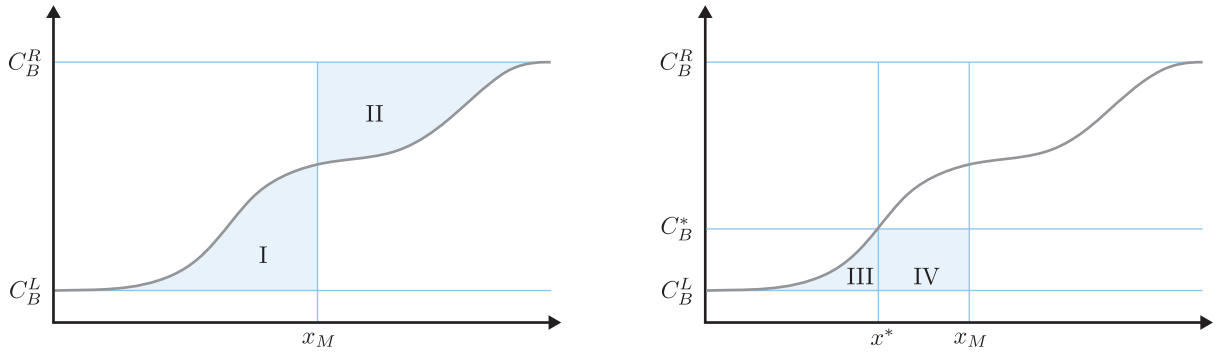

Figure 2.1:
(Left) To find the Matano plane we must find the position x_M for which I and II are equal. (Right) After having located the Matano plane, the integral $\int_{C_B^L}^{C^*} x - x_M dC_B$ corresponds to the sum of the areas III and IV.

Putting everything together, the Boltzmann-Matano method can be summarized in algorithmic form as follows:

[4] Normally, when integrating a function $f = f(x)$, we automatically take $f = 0$ as a reference value for evaluating position for the integral. That is, $\int f(x)dx = \int (f(x) - 0)dx$.

1. Given a concentration profile at a time t^*, locate the Matano plane by finding the value of x_M for which $\int_{-\infty}^{x_M} C_B(x) - C_B^L dx = \int_{x_M}^{\infty} C_B^R - C_B(x) dx$;

2. For $C_B^* \in [C_B^L, C_B^R]$:

   (a) Compute integral $\int_{C_B^L}^{C^*} x(C_B) - x_M dC_B$ from the concentration profile;

   (b) Compute the concentration gradient $\left(\frac{dC_B}{dx}\right)_{x^*}$ from the concentration profile;

   (c) Compute $\tilde{D}(C_B^*)$ using equation (2.13).

There are a few problems with the Boltzmann-Matano method though. First of all, to evaluate the integral $\int_{C_B^L}^{C^*} x(C_B) - x_M dC_B$, one basically needs to invert the graph of C_B as a function of x. But this may not be possible - for example in case C_B has local extrema. It should be noted that for diffusion couples with two different atomic components such local extrema should not occur. However, as we will discuss later, for diffusion couples with three or more components the so called cross-diffusion terms may give rise to uphill diffusion and hence to local extrema. Secondly, one needs to determine the position of the Matano plane. A fine grid is needed to determine the position accurately, and even then one introduces a numerical error into the solution. Thirdly, for materials in which the partial molar volumes V_i are not constant, the Matano plane may not be unique. Also, for systems with more than two components, different components may have different Matano planes. And - of course - using different Matano planes ultimately results in different diffusion coefficients. To overcome these issues, we use an improvement of the Boltzmann-Matano method originally proposed by Den Broeder [18].

2.2.2 Den Broeder Method

Suppose we have a concentration profile as depicted below.

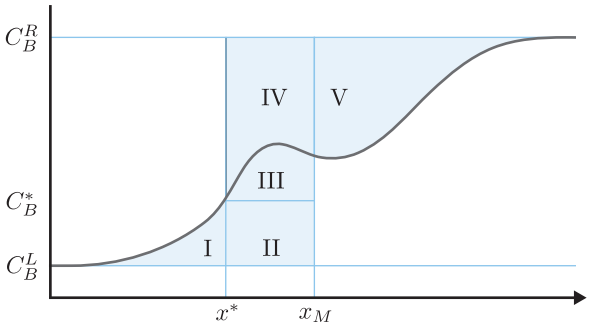

Figure 2.2: Illustration for Den Broeder method.

Note that

$$\int_{C_B^L}^{C^*} x(C_B) - x_M dC_B = I + II.$$

To find an expression for $I + II$ that does not make use of the location of the Matano plane, Den Broeder introduced the relative concentration Y_B defined as

$$Y_B = \frac{C_B - C_B^L}{C_B^R - C_B^L}.$$

With this definition, we see that

$$\begin{aligned} Y_B^*(II + III + IV) &= Y_B(x_M - x^*)(C_B^R - C_B^L) \\ &= (x_M - x^*)(C_B^* - C_B^L) \\ &= II, \end{aligned}$$

and (trivially)

$$I = Y^* I + (1 - Y^*)I.$$

Furthermore, from the definition of the Matano plane we know that

$$I + II + III = V.$$

Putting all of this together we find that

$$\begin{aligned} I + II &= Y^* I + (1 - Y^*)I + II \\ &= Y^* I + (1 - Y^*)I + Y_B^*(II + III + IV) \\ &= Y^* I + (1 - Y^*)I + Y_B^*(IV + V - I) \\ &= (1 - Y^*)I + Y_B^*(IV + V) \\ &= (1 - Y^*) \int_{-\infty}^{x^*} \left(C_B(x) - C_B^L\right) dx + Y_B^* \int_{x^*}^{+\infty} \left(C_B^R - C_B(x)\right) dx. \end{aligned}$$

It follows that

$$\tilde{D}(C_B^*) = \frac{1}{2t^* \left(\frac{dC_B}{dx}\right)_{x^*}} \left[(1 - Y_B^*) \int_{-\infty}^{x^*} \left(C_B(x) - C_B^L\right) dx + Y_B^* \int_{x^*}^{\infty} \left(C_B^R - C_B(x)\right) dx \right]. \quad (2.15)$$

Comparing equation (2.15) with equation (2.8) reveals that the flux $\tilde{F}_B$ can be expressed as

$$\tilde{F}(x^*) = -\frac{1}{2t^*} \left[(1 - Y_B^*) \int_{-\infty}^{x^*} \left(C_B(x) - C_B^L\right) dx + Y_B^* \int_{x^*}^{\infty} \left(C_B^R - C_B(x)\right) dx \right]. \quad (2.16)$$

We see that it is no longer necessary to locate Matano plane. Furthermore, because we integrate with respect to x this time, it is no problem if C_B exhibits local extrema (like it does in the (artificial) picture). The Den Broeder method can be summarized in algorithmic form as follows:

1. Given a concentration profile at a time t^*, pick a concentration $C_B^* \in [C_B^L, C_B^R]$:
   (a) Compute the relative concentration $Y_B^* = \frac{C_B^* - C_B^L}{C_B^R - C_B^L}$.
   (b) Compute the integrals $\int_{-\infty}^{x^*} \left(C_B(x) - C_B^L\right) dx$ and $\int_{x^*}^{\infty} \left(C_B^R - C_B(x)\right) dx$ from the concentration profile;
   (c) Compute the concentration gradient $\left(\frac{dC_B}{dx}\right)_{x^*}$ from the concentration profile;
   (d) Compute $\tilde{D}(C_B^*)$ using equation (2.15).

The Den Broeder method as presented above only works when the partial molar volumes V_i are constant. When this is not the case the method needs to be - and in fact can be - modified. We will assume the partial molar volumes to be constant though.

2.2.3 Numerical Tests

We implemented both the Boltzmann-Matano method and the Den Broeder method in `MATLAB`. We generated concentration profiles using an interdiffusion coefficient that is either constant, linear in C_B, quadratic in C_B and finally some oscillating function of C_B.[5] As expected, the results for both methods are similar, assuming that the Boltzmann-Matano method can indeed be used (as mentioned before, for the Boltmann-Matano method the concentration profile needs to be inverted and that may not be possible). However, from a numerical point of view it could be argued that the Den Broeder method is better. The reason being that the locating the position of the Matano-plane introduces additional errors that are not present when using the Den Broeder method. To illustrate this, we have computed the fluxes for some (random) test case using both equation(2.14) and equation (2.16). The results are plotted in Figure 2.3. The difference in fluxes at the right end of the diffusion couple between both methods are due to the error in exactly locating the Matano plane. The error - and hence the difference between the methods - can be reduced by introducing more numerical grid points. But this comes at the cost of making the method more 'expensive'. We conclude that the Den Broeder method is superior to the Boltzmann-Matano method and therefore it will be our method of choice.

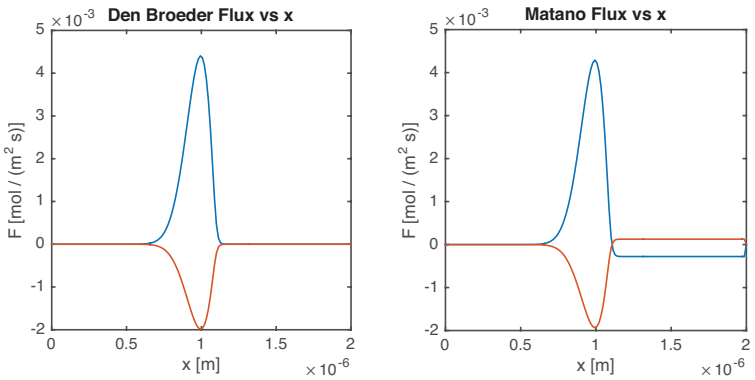

Figure 2.3: Fluxes for the two atomic components computed using the Den Broeder method (left) and the Boltzmann-Matano method (right).

The results obtained with the Den Broeder method are presented in figures (2.4) to (2.7). Each time, on the left a concentration profile is shown at some time t^*. The initial conditions are shown as dashed lines. The profile at time t^* is generated using the numerical scheme to be discussed in Chapter 4. The computed interdiffusion coefficients as function of concentration are presented are shown in logplots.

[5] The numerical method used to generate the profiles is the finite volume scheme together with explicit Euler time integration to be discussed in Chapter 4.

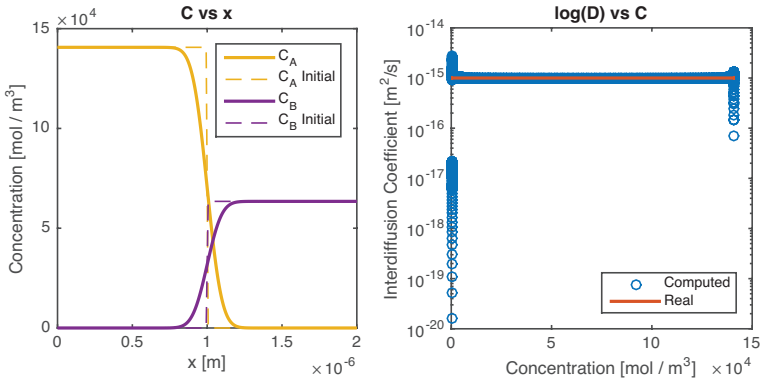

Figure 2.4: Den Broeder method for constant interdiffusion coefficient. The purple and yellow lines represent the concentrations of components A and B respectively. The solid red lines represent the real diffusion coefficient. The blue o's represent the diffusion coefficient as recovered by the Den Broeder method.

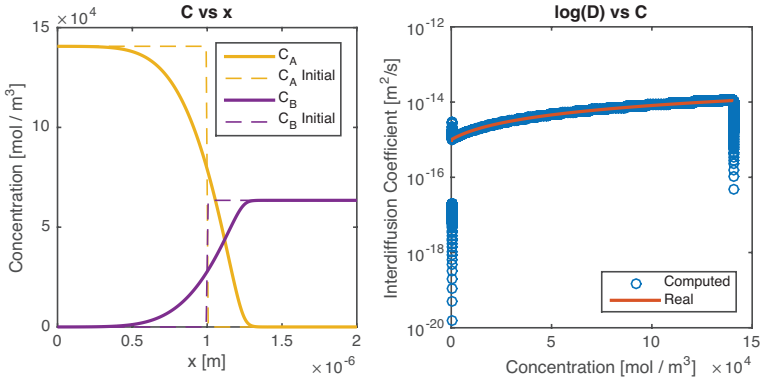

Figure 2.5: Den Broeder method for interdiffusion coefficient depending linearly on concentration.

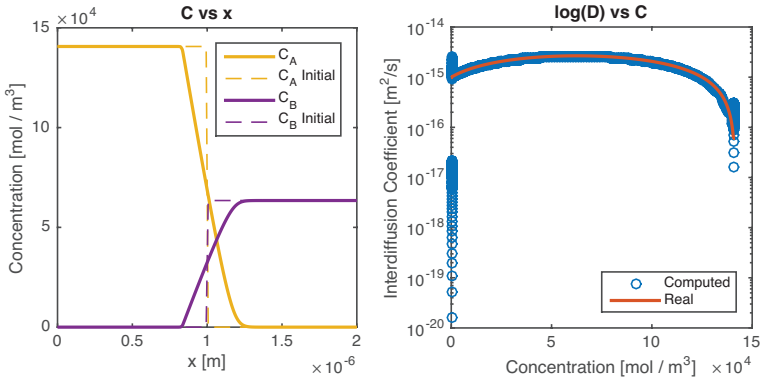

Figure 2.6: Den Broeder method for interdiffusion coefficient depending quadratically on concentration.

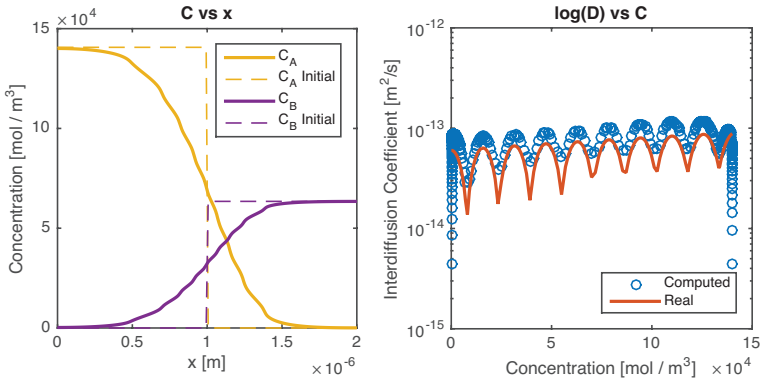

Figure 2.7: Den Broeder method for linearly increasing interdiffusion coefficient with oscillations.

We see that the Den Broeder method does a pretty good job of recovering the concentration dependent interdiffusion coefficients. Only at the left and right ends we see that the method produces weird results. This is related to the fact that the Den Broeder method (and the Boltzmann-Matano method as well) requires division by the concentration gradient. But the concentration gradient vanishes near the ends of the diffusion couple because the ends remain unaffected. This makes the calculation at the ends unstable. This is not a serious problem as we can easily extrapolate the solution. For the oscillating case this is not so straightforward but honestly we don't expect to see such behaviour in real experiments. More importantly though, what happens when the concentration gradient vanishes within the interdiffusion zone?

2.2.4 Phases with Narrow Homogeneity Range

When preparing diffusion couples of metallic elements it is common for new binary phases to form within the diffusion zone. Often these phases have no or a very narrow homogeneity range. That is, there is no or a very small concentration gradient within these phases. This makes both the Boltzmann-Matano and the Den Broeder method unstable. To deal with this issue, Wagner introduced the concepts of *integrated interdiffusion coefficients* and *average interdiffusion coefficients* [57]. These concepts and their use are explained below.

Suppose a new phase grows at the initial interface between A and B. Let's refer to this phase as the $\beta-$phase. Suppose that the left end of the phase is located at x_{β_L} and the right end at x_{β_R}. Let the molar fractions at these locations be denoted by $N_B^{\beta_L}$ and $N_B^{\beta_R}$ respectively. In other words, $N_B^{\beta_L}$ denotes the molar fraction of component B at the left end of the $\beta-$phase. Similarly, $N_B^{\beta_R}$ denotes the molar fraction of component B at the right end of the $\beta-$phase Futhermore, define $\Delta N_B^\beta := N_B^{\beta_R} - N_B^{\beta_L}$. Then the integrated interdiffusion coefficient $\tilde{D}_{\text{int}}^\beta$ is defined as

$$\tilde{D}_{\text{int}}^\beta = \int_{N_B^{\beta_L}}^{N_B^{\beta_R}} \tilde{D}(N_B) dN_B.$$

Assuming $\tilde{D}$ to be constant and equal to some value $\tilde{D}^\beta$ over the narrow composition range, we can write

$$\tilde{D}_{\text{int}}^\beta = \tilde{D}^\beta \Delta N_B^\beta.$$

The question is now how to compute this integrated interdiffusion coefficient. To this end, note that Fick's law, assuming constant molar volume V_{mol} in the phase of interest, tells us that the diffusive flux $\tilde{F}_B$ can be written as

$$\tilde{F}_B = -\tilde{D}\frac{\partial C_B}{\partial x} = -\frac{\tilde{D}}{V_{\text{mol}}}\frac{\partial N_B}{\partial x}.$$

But then we see that

$$\tilde{D}^{\beta}_{\text{int}} = \int_{N_B^{\beta_L}}^{N_B^{\beta_R}} \tilde{D}(N_B) dN_B = -V_{\text{mol}} \int_{x_{\beta_L}}^{x_{\beta_R}} \tilde{F}_B(x) dx.$$

On the other hand, remember from equation (2.16) that

$$\tilde{F}(x^*) = -\frac{1}{2t^*}\left[(1-Y_B^*)\int_{-\infty}^{x^*}\left(C_B(x)-C_B^L\right)dx + Y_B^*\int_{x^*}^{\infty}\left(C_B^R - C_B(x)\right)dx\right]. \tag{2.17}$$

It follows that

$$\begin{aligned} & \tilde{D}^{\beta}_{\text{int}} \\ = \; & -V_{\text{mol}} \int_{x_{\beta_1}}^{x_{\beta_2}} \tilde{F}_B(x) dx \\ = \; & -\frac{V_{\text{mol}}}{2t^*}\int_{x_{\beta_1}}^{x_{\beta_2}}\left[(1-Y_B(x))\int_{-\infty}^{x}\left(C_B(y)-C_B^L\right)dy + Y_B(x)\int_{x}^{\infty}\left(C_B^R - C_B(y)\right)dy\right] \end{aligned} \tag{2.18}$$

Note that the integrated interdiffusion coefficient does not require the evalutation of a concentration gradient. Given the integrated interdifusion coefficient, we can compute the average interdiffusion coefficient using

$$\tilde{D}^{\beta}_{\text{average}} = \frac{\tilde{D}^{\beta}_{\text{int}}}{\Delta N_B^{\beta}}. \tag{2.19}$$

The suggested use of the concept of average interdiffusion coefficients is the following. Suppose it is known that an intermetallic compound grows diffusion zone. Take $[x_{\beta_L}, x_{\beta_R}]$ to be the interval in which this new phase has grown.[6] Then the average interdiffusion coefficient over this range could be interpreted as the interdiffusion coefficient in the new phase.

In phases where there is no composition range, i.e. line compounds, ΔN_B^{β} is zero and the average interdiffusion coefficient cannot be computed. In that case the integrated interdiffusion coefficient can still be related to the growth rate of this new phase, as discussed by Wagner [57]. We will assume that we do not have to deal with line compounds though.

The above method can be summarized in algorithmic form as follows.

1. Given a concentration profile at a time t^*, pick a range $[x_{\beta_L}, x_{\beta_R}]$
   (a) Evaluate the fluxes over the range $[x_{\beta_L}, x_{\beta_R}]$ using equation (2.17);
   (b) Integrate the fluxes over $[x_{\beta_L}, x_{\beta_R}]$ and use this to compute $\tilde{D}^{\beta}_{\text{int}}$ from equation (2.18)
   (c) Compute $\tilde{D}^{\beta}_{\text{average}}$ using equation (2.19).

The method is tested with an artificially generated concentration profile again. In figure (2.8) a profile is shown that resembles the concentration profile with a pure $A-$phase (left end of the couple), a B-phase (right end of the couple) and some new intermetallic phase that started growing in between. In the middle and right plots of this figure the results of the Den Broeder method are shown. The $\log(\tilde{D})$ vs C_B plot in the middle looks pretty good: it captures the peak in the diffusion coefficient pretty good. But the plot of $\log(\tilde{D})$ vs x shows some weird behaviour. Near the left end of the couple it shows no results due to the concentration gradient being

[6] Note that it is assumed that new phases grow in layers over the complete height of the diffusion couple. For systems of two components it can be shown using the so called *Gibbs phase rule* that this is indeed what happens. For systems with three or more components, this behaviour is not guaranteed [54].

(close to) zero there (so that MATLAB returning NaN when trying to divide by the concentration gradient). Near the right end of the couple we see that the method is a few orders of magnitude off.

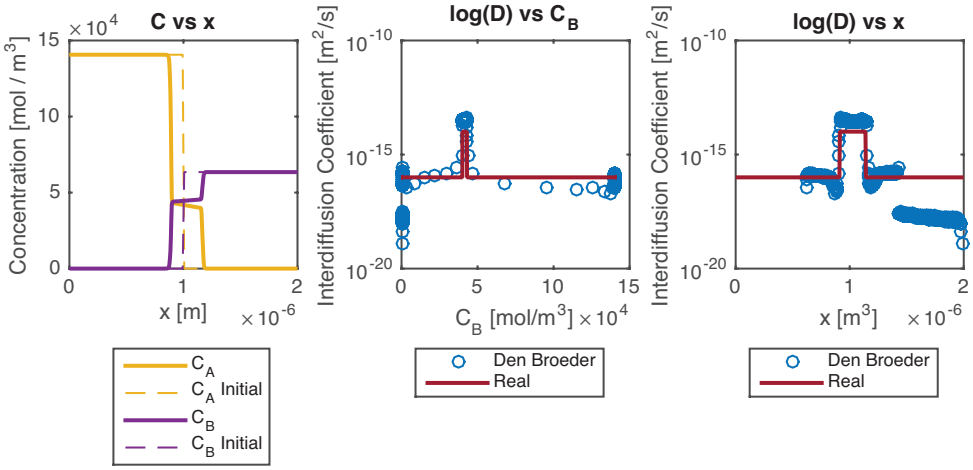

Figure 2.8: Den Broeder method for interdiffusion coefficient with a single peak.

In figure (2.9) the same concentration profile is shown. This time, we tried to compute the average interdiffusion coefficient in the three phases separately. We see that the $\log(\tilde{D})$ vs C_B plot is quite similar but the $\log(\tilde{D})$ vs x plot is much better. The issue of 'dividing by a zero-concentration gradient' is resolved. Note that the method is not spot-on (it also depends on how you choose the concentration ranges over which the average interdiffusion coefficient is computed) but - at least for this test case - the errors are within an order of magnitude.

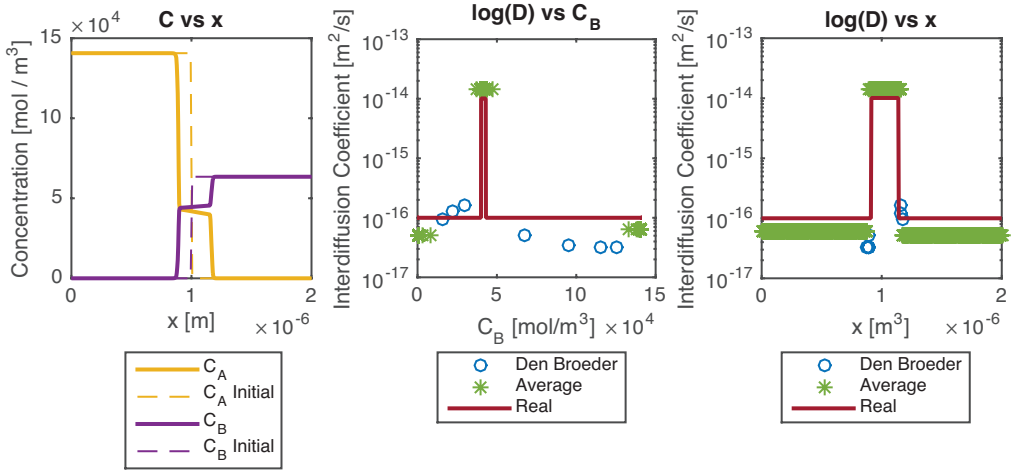

Figure 2.9: Average interdiffusion coefficients

On a side note, it is in itself interesting to note that having a diffusion coefficient with a single peak over a small concentration range can give rise to a concentration profile that resembles one in which an intermetallic phase grows at the initial contact plane. If we run the diffusion process for a longer time, this new phase seems to grow in thickness. Furthermore, using an interdiffusion coefficient that has several peaks gives rise to a concentration profile that resembles the situation where several intermetallic phases grow at the initial contact plane. In other words, it seems as if the growth of new phases can be encoded into the interdiffusion coefficient. Other ways of modelling this behaviour is to keep track of boundaries between different phases explicitly (similar to the so called *Stefan Problem*), or, to specifiy the chemical reactions that are responsible for the growth of the new phases and incorporate them separately into the model (see Chapter 4).

2.3 Interdiffusion in Multicomponent Systems

2.3.1 Onsager's Transport Equations

Now suppose that the beams making up our diffusion couple contain more than 2, say n, atomic components. In a completely similar fashion as before we could work out the Onsager transport equations using both lattice-fixed sections and volume-fixed (or, assuming constant partial molar volumes, laboratory-fixed) sections. Using volume-fixed sections the transport equations read as

$$\tilde{F}_i = \sum_{j=1}^{n} \tilde{L}_{ij} \nabla \mu_i.$$

Using the Gibbs-Duhem relation we know there are only $n-1$ independent chemical potential gradients. Assume component n to be the *dependent component*. Furthermore, assuming Vegard's law to be obeyed we know we are dealing with a volume-fixed frame in which the relation

$$\sum_{i=1}^{n} V_i \tilde{F}_i = 0$$

is satisfied. This shows there are only $n-1$ independent fluxes. Following De Groot and Mazur [16] and Brady [5] one can use these relations to show that

$$\tilde{F}_i = \sum_{j=1}^{n-1} \tilde{L}_{ij} \left[\sum_{k=1}^{n-1} \left(\delta_{jk} + \frac{v_k}{v_n} \right) \nabla \mu_k \right], \quad i \in \{1, \ldots, n-1\},$$

with δ_{ij} being the Kronecker delta function. Like before we can rewrite the chemical potential gradients in terms of concentration gradients. In Balluffi et al [3] it is shown how one arrives at

$$\tilde{J}_i = -\sum_{j=1}^{n-1} \tilde{D}_{ij}^{n} \nabla C_i,$$

with

$$\tilde{D}_{ij}^{n} = -\sum_{k=1}^{n-1} \sum_{s=1}^{n-1} \tilde{L}_{is} \left(\delta_{ks} - \frac{v_k}{v_n} \right) \frac{\partial \mu_k}{\partial C_j}. \tag{2.20}$$

We write $\tilde{D}_{ij}^{n}$ to emphasize that the $n-$th component has been chosen as the dependent component. From the above expressions we see that each interdiffusion flux requires $n-1$ interdiffusion coefficients. There are n different atomic and hence n different fluxes in the system. However, since we are working in a volume-fixed reference frame, we know that there are only $n-1$ *independent* interdiffusion fluxes and hence the system can be described by using $(n-1)^2$ interdiffusion coefficients. Note here that we could do a similar analysis for intrinsic fluxes, only then to arrive at the conclusion that we have n independent intrinsic fluxes and hence $(n-1)n$ intrinsic diffusion coefficients. For the case $n = 2$ we have already seen this in Section 2.1 of this chapter.

Compared with Fick's law, which only 'requires' n intrinsic or $n-1$ interdiffusion coefficients, we see that the Onsager formalism gives rise to so called *cross-diffusion terms*. That is, in Fick's law the concentration of a component only evolves under the influence of this components' own concentration gradient. In the Onsager formalism, it can also evolve under the concentration gradient of other components. This can give rise to behaviour like uphill diffusion which is indeed observed in literature for multicomponent systems [13]. Only if we assume that $D_{ij}^{n} = 0$ for $i \neq j$ does the Onsager formalism reduce to Fick's law.

2.3.2 Problems with Interdiffusion Coefficients for Multicomponent Systems

As just discussed, we need $(n-1)^2$ interdiffusion coefficients to describe diffusion in an n-component system. In fact, not all of them are independent. This is related to the result that the transport coefficient matrix $\tilde{L}$ is symmetric. That is, the components $\tilde{L}_{ij}$ satisfy $\tilde{L}_{ij} = \tilde{L}_{ji}$ and as a result there are only $\frac{1}{2}(n-1)n$ independent transport coefficients. Using these relations together with equation 2.20 it can be shown that only $\frac{1}{2}(n-1)n$ independent interdiffusion coefficients exist. However, the relations between the $\tilde{D}_{ij}$ are not as simple as $\tilde{D}_{ij} = \tilde{D}_{ji}$. In general the relations between the interdiffusion coefficients cannot be evaluated (they require derivatives $\partial\mu_i/\partial C_j$ to be known, as mentioned by Brady [5]) and we still need to determine the $(n-1)^2$ interdiffusion coefficients independently.

To be able to calculate the $(n-1)^2$ interdiffusion coefficients from measurements, we need at least $(n-1)^2$ equations. But where do we get the equations from? Remember from the Boltzmann-Matano method that in a two component system the interdiffusion flux $\tilde{F}_B$ could be expressed as

$$\tilde{F}_B = \frac{1}{2t^*}\int_{C_B^-}^{C_B^*} x(C_B) - x_M dC_B,$$

see equation (2.14). Similarly, using the Den Broeder method we found $\tilde{F}_B$ to be

$$\tilde{F}_B(x^*) \;=\; -\frac{1}{2t^*}\left[(1-Y_B^*)\int_{-\infty}^{x^*}\left(C_B(x)-C_B^L\right)dx + Y_B^*\int_{x^*}^{\infty}\left(C_B^R - C_B(x)\right)dx\right],$$

see equation (2.16)

According to Dayananda and Kim [13], the above two expressions for the interdiffusion flux should hold for any component in a multicomponent system. That is, for each atomic component i the interdiffusion flux $\tilde{F}_i$ is given by

$$\begin{aligned}\tilde{F}_i(x^*) &= \frac{1}{2t^*}\int_{C_i^-}^{C_i^*} x(C_i) - x_M dC_i \quad \text{(Boltzmann-Matano)}\\ &= -\frac{1}{2t^*}\left[(1-Y_i^*)\int_{-\infty}^{x^*}\left(C_i(x)-C_i^L\right)dx + Y_i^*\int_{x^*}^{\infty}\left(C_i^R - C_i(x)\right)dx\right] \quad \text{(Den Broeder)}.\end{aligned}$$

Each flux gives rise to an expression that can be evaluated from experimental measurements using either of the two methods (but as discussed before we prefer the Den Broeder method). But that's not enough information yet to solve for the interdiffusion coefficients. This can be illustrated using a ternary system. Consider a system consisting of atomic components A, B and C. Choose C as the dependent component. From the general discussion above, we know that the system has two independent interdiffusion fluxes $\tilde{F}_A$ and $\tilde{F}_B$ and we need to determine four interdiffusion coefficients that are related to the fluxes by

$$\begin{pmatrix}\tilde{F}_A\\ \tilde{F}_B\end{pmatrix} = -\begin{pmatrix}\tilde{D}_{AA}^C & \tilde{D}_{AB}^C\\ \tilde{D}_{BA}^C & \tilde{D}_{BB}^C\end{pmatrix}\begin{pmatrix}\nabla C_A\\ \nabla C_B\end{pmatrix}. \tag{2.21}$$

As we just saw, experimental measurements allow us to evaluate the interdiffusion fluxes (and also the concentration gradients). But that gives us two equations for four unknowns. To overcome this issue, we need another diffusion couple of which the *diffusion path* in the phase diagram crosses the diffusion path of the original couple.

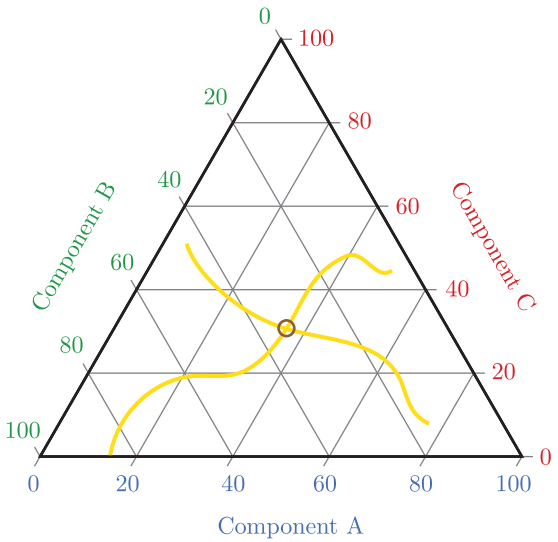

Figure 2.10: Two possible diffusion paths (yellow) in a ternary phase diagram. A diffusion path represents the composition range present within a diffusion couple after the diffusion process has started. The ends of the yellow lines represent the (unaffected) composition at the left and right ends of the diffusion couples. The two diffusion paths intersect in the point indicated by the circle.

At the composition where the paths intersect, we get four equations - two from each diffusion couple - that can be solved for the four unknown interdiffusion coefficients. But even if we are able to produce two diffusion couples whose diffusion paths intersect, we will only find the diffusion coefficient at the concentrations / compositions at the intersection point. To get a general idea of the concentration dependence of the interdiffusion coefficients, a lot of different diffusion couples need to be prepared. This is experimentally challenging. And for systems of even more components the situation gets worse. For example, in a system with four components, we need nine interdiffusion coefficients. A single diffusion couple can give three expressions. Hence we need three different couples whose diffusion paths intersect. This is practically impossible to achieve experimentally.

On top of the above, there is another issue with (inter)diffusion coefficients in multicomponent systems. Remember how the superscript in $\tilde{D}_{ij}^{n}$ refers to the choice of the dependent component? As it turns out, the interdiffusion coefficients are not necessarily the same for different choices of dependent components. That is, D_{ij}^{k} is not necessarily equal to D_{ij}^{l} for $k \neq l$. In Chapter 9.1 of Paul et al [43] relations between D_{ij}^{k} for different k in the ternary case are presented. A possible guide in choosing which component to consider as dependent component is that the resulting matrix of interdiffusion coefficients should be positive definite [30] This requirement should be interpreted as a generalization of the physical requirement that the interdiffusion coefficient in a two component system should always be positive.

Taking the problems with multicomponent diffusion discussed in this subsection into account, one can understand why there are no databases of (inter)diffusion coefficients for systems with three or more components. Luckily, some methods have been proposed in literature that can partially deal with the discussed problems. We explore these methods in the next subsection for ternary systems, i.e. systems consisting of three components. The reason is that, as explained in Chapter 6, we are interested in deriving interdiffusion coefficients for a system consisting of copper, indium and gallium.

2.3.3 Average Interdiffusion Coefficients for Ternary Systems

Instead of trying to compute interdiffusion coefficients at specific concentrations, Dayananda and Sohn [15] proposed to compute *average interdiffusion coefficients* over concentration ranges as follows. Let $[x_L, x_R]$ be some interval in the spatial domain. Define average interdiffusion coefficients $\bar{\tilde{D}}^n_{ij}$ as

$$\bar{\tilde{D}}^n_{ij} = \frac{\int_{C_j(x_L)}^{C_j(x_R)} \tilde{D}^n_{ij} dC_j}{\int_{C_j(x_L)}^{C_j(x_R)} dC_j}.$$

Note that the $\bar{\tilde{D}}^n_{ij}$ are constants over their specified concentration ranges. From now on we will assume a ternary system again. By performing a clever integration one can find additional independent equations for the average interdiffusion coefficients that can be evaluated using a single concentration profile. Indeed, if we multiply 2.21 by $(x - x_M)^p$ for some integer p over and subsequently integrate over interval $[x_L, x_R]$, we find for $i \in \{A, B\}$ that

$$\begin{aligned}\int_{x_L}^{x_R} \tilde{F}_i(x)(x - x_M)^p dx &= -\bar{\tilde{D}}^C_{iA} \int_{x_L}^{x_R} \frac{\partial C_A}{\partial x}(x)(x - x_M)^p dx - \bar{\tilde{D}}^C_{iB} \int_{x_L}^{x_R} \frac{\partial C_B}{\partial x}(x)(x - x_M)^p dx \\ &= -\bar{\tilde{D}}^C_{iA} \int_{C_A(x_L)}^{C_A(x_R)} (x(C_A) - x_M)^p dC_A - \bar{\tilde{D}}^C_{iB} \int_{C_B(x_L)}^{C_B(x_R)} (x(C_B) - x_M)^p dC_B.\end{aligned}$$

For $p = 0$ this becomes

$$\int_{x_L}^{x_R} \tilde{F}_i(x) dx = -\bar{\tilde{D}}^C_{iA} \left[C_A(x_R) - C_A(x_L)\right] - \bar{\tilde{D}}^C_{iB} \left[C_B(x_R) - C_B(x_L)\right].$$

On the other hand, for $p = 1$ we find

$$\begin{aligned}\int_{x_L}^{x_R} \tilde{F}_i(x)(x - x_M) dx &= -\bar{\tilde{D}}^C_{iA} \int_{C_A(x_L)}^{C_A(x_R)} (x(C_A) - x_M) dC_A \\ &\quad -\bar{\tilde{D}}^C_{iB} \int_{C_B(x_L)}^{C_B(x_R)} (x(C_B) - x_M) dC_A \qquad (2.22)\end{aligned}$$

If we look closely at the integrals on the right-hand side, and remembering the Boltzmann-Matano method (see equation (2.14)), we see that they are related to the fluxes by

$$\int_{C_i(x_L)}^{C_i(x_R)} (x(C_i) - x_M) dC_i = 2t^* \left[\tilde{F}_i(x_R) - \tilde{F}_i(x_L)\right].$$

It follows that

$$\int_{x_L}^{x_R} \tilde{F}_i(x)(x - x_M) dx = -2t^* \bar{\tilde{D}}^C_{iA} \left[\tilde{F}_A(x_R) - \tilde{F}_A(x_L)\right] - 2t^* \bar{\tilde{D}}^C_{iB} \left[\tilde{F}_B(x_R) - \tilde{F}_B(x_L)\right]. \quad (2.23)$$

Putting (2.22) and (2.23) together, we arrive at a systems of four equations for the four unknowns $\bar{\tilde{D}}^C_{ij},\ i, j \in \{A, B\}$. If we introduce the notation

$$\Delta C_i = \left[C_i(x_R) - C_i(x_L)\right], \quad \Delta \tilde{F}_i = \left[\tilde{F}_i(x_R) - \tilde{F}_i(x_L)\right],$$

then this system of equations can be written in matrix-form as:

$$-\begin{pmatrix} \Delta C_A & \Delta C_B & 0 & 0 \\ 0 & 0 & \Delta C_A & \Delta C_B \\ 2t^* \Delta \tilde{F}_A & 2t^* \Delta \tilde{F}_B & 0 & 0 \\ 0 & 0 & 2t^* \Delta \tilde{F}_A & 2t^* \Delta \tilde{F}_B \end{pmatrix} \begin{pmatrix} \bar{\tilde{D}}^C_{AA} \\ \bar{\tilde{D}}^C_{AB} \\ \bar{\tilde{D}}^C_{BA} \\ \bar{\tilde{D}}^C_{BB} \end{pmatrix} = \begin{pmatrix} \int_{x_L}^{x_R} \tilde{F}_A(x) dx \\ \int_{x_L}^{x_R} \tilde{F}_B(x) dx \\ \int_{x_L}^{x_R} \tilde{F}_A(x)(x - x_M) dx \\ \int_{x_L}^{x_R} \tilde{F}_B(x)(x - x_M) dx \end{pmatrix}. \quad (2.24)$$

The determinant of the matrix on the left-hand side is given by

$$-\left(2t^*\Delta C_B\Delta\tilde{F}_A - 2t^*\Delta C_A\Delta\tilde{F}_B\right)^2,$$

and we assume it to be non-zero so that the system (2.24) has a unique solution.

One may be tempted to look for 'real' interdiffusion coefficients at specific concentrations by making the interval $[x_L, x_R]$ smaller and smaller. This approach was proposed by ˇCermák et al [6]. Later it was shown by Cheng et al [8] that this method is not at all reliable because the matrix in (2.24) becomes increasingly ill-conditioned as $x_R \to x_L$. In general, we will partition the into two or three different

The method for extracting average interdiffusion coefficients can be summarized in algorithmic form as follows.

1. Given a concentration profile at a time t^*, locate the Matano plane by finding the value of x_M for which $\int_{-\infty}^{x_M} C_i(x) - C_i^L dx = \int_{x_M}^{\infty} C_i^R - C_i(x)dx$ (the location of the Matano plane should be independent of i under the assumption of constant partial molar volumes);

2. Choose a dependent component and label it C. The independent components will be labeled A and B respectively;

3. Pick a range $[x_L, x_R]$;

   (a) Evaluate the fluxes over the range $[x_L, x_R]$ using equation 2.17;

   (b) Use the fluxes to evaluate the integrals in the right-hand side of equation (2.24);

   (c) Set up the matrix in equation (2.24).

   (d) Solve equation (2.24) for the average interdiffusion coefficients $\tilde{\bar{D}}^C_{AA}, \tilde{\bar{D}}^C_{AB}, \tilde{\bar{D}}^C_{BA}, \tilde{\bar{D}}^C_{BB}$;

   (e) Check whether the matrix

   $$\begin{pmatrix} \tilde{\bar{D}}^C_{AA} & \tilde{\bar{D}}^C_{AB} \\ \tilde{\bar{D}}^C_{BA} & \tilde{\bar{D}}^C_{BB} \end{pmatrix}$$

   is positive-definite. If not, choose different range $[x_L, x_R]$ or choose a different component as dependent component and try again.

2.3.4 Numerical Tests

We implemented the algorithm for extracting average interdiffusion coefficients in `MATLAB`. In Figures 2.11 and 2.12 two test results are presented. The results are generated as follows. We assume to be working with a diffusion couple for which initially the concentrations at both sides of the initial contact plane are uniform (we do not show the initial conditions in the plots). In the first test case a concentration profile at some time t^* is generated with four constant interdiffusion coefficients. In the second test case the interdiffusion coefficients are given a linear dependence on one of the concentrations. In both cases we computed the average interdiffusion coefficients over the region left and right of the Matano plane respectively.

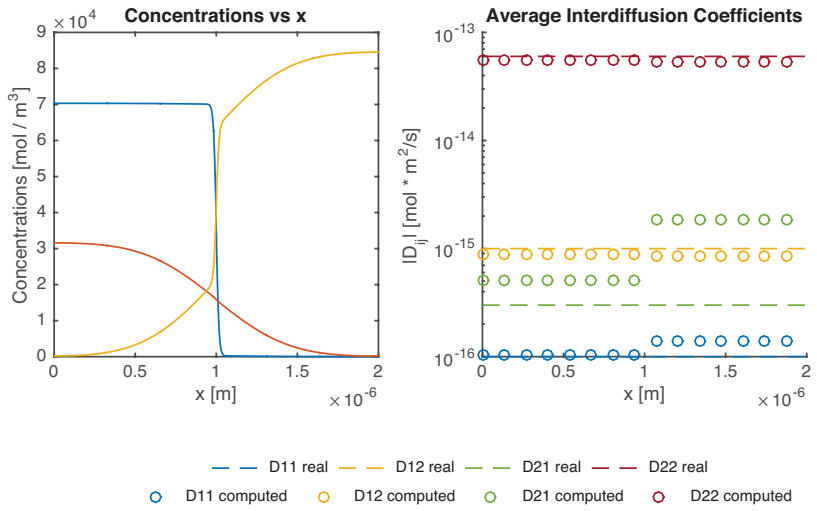

Figure 2.11: (Left) Concentration profile with dependent component shown in red. (Right) Computed average interdiffusion coefficients for ternary system with constant interdiffusion coefficients.

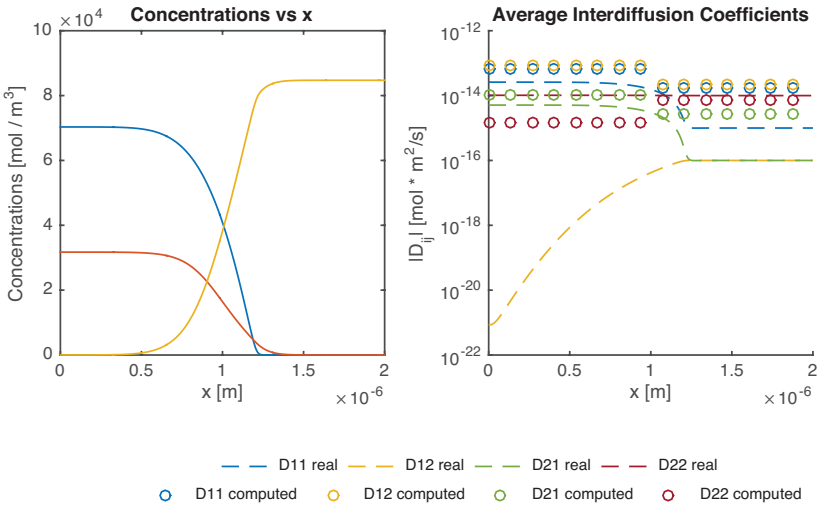

Figure 2.12: (Left) Concentration profile with dependent component shown in red. (Right) Computed average interdiffusion coefficients for ternary system with interdiffusion coefficients depending linearly on one of the concentrations.

In Figure 2.11, we see that the three out of four of the constant interdiffusion coefficients are recovered nicely. For the linear case the results as presented in Figure 2.12 seem worse. At first we thought there must have been programming errors because Dayananda and Sohn seemed to arrive at good results.[7] But then it would be strange to arrive at good results for constant

[7]It should be noted here that Dayananda and Sohn [15] use a different method to recover the concentration profiles from the computed average interdiffusion coefficients. They make use of the fact that a ternary system with constant diffusion coefficients has an exact (similarity) solution. To recover the concentration profile, they divide the measured concentration profile into ranges, compute the average interdiffusion coefficients over these ranges and then they use the 'exact' solution in each of these ranges and patch them together. As boundary conditions for the exact (similarity) solutions they use the values known from the measured concentration profiles. See Chapter 3 of the Phd thesis *Analysis of interdiffusion and diffusion paths in multicomponent systems* by Day [11] for more details as well. But then, if the number of ranges used increases (i.e. as the intervals $[x_L, x_R]$ become smaller), the result is going to resemble the measured concentration profile better and better simply because more exact values (the boundary points) are known (it is as if one is doing numerical computations using more and more grid points). On the other hand, as explained in the paper by Cheng et al [8], the computed average

coefficients because exactly the same method is used. We decided to 'turn a blind eye' to the bad results and we tried to use the computed coefficients to recover the concentration profile (as shown on the let in Figure) from the initial concentrations by using the numerical scheme to be discussed in Chapter 4.[8] Not surprisingly, for the case of constant interdiffusion coefficients we are able to recover the concentration profile to good accuracy (results not shown). However, and this was quite surprising, Figure 2.13 reveals that for the case of linear interdiffusion coefficients we also obtain good results. Even though the computed interdiffusion coefficients appeared to be pretty bad.

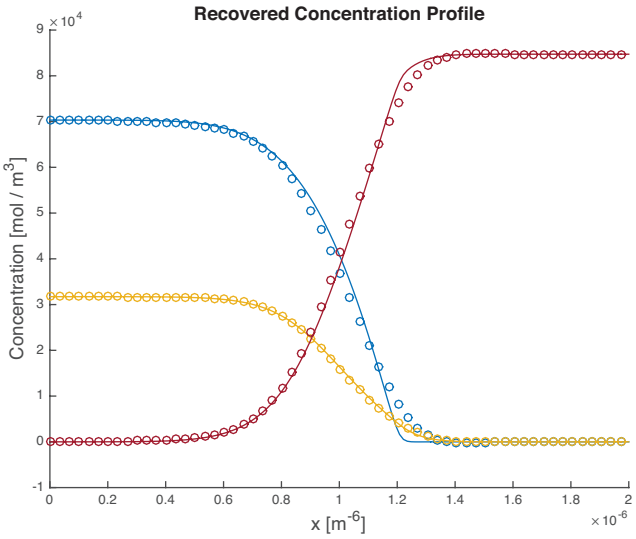

Figure 2.13: The concentration profile used to compute average interdiffusion coefficients is shown in solid lines. The circles represent the profile that has been recovered using the average interdiffusion coefficients.

A possible explanation for the observation that seemingly 'bad' interdiffusion coefficient may still recover the concentration profile to good accuracy is that the interdiffusion coefficients may not be *unique* in the sense that different sets of interdiffusion coefficients may give rise to the same concentration profile at time t^*.

It should also be noted that to recover the concentration profile shown in Figure 2.13, we fit interdiffusion coefficients depending linearly on one of the concentrations to the computed average interdiffusion coefficients. But, of course, we already knew beforehand that the diffusion coefficients had a linear dependence on one of the concentrations. In real situations, it may not be obvious how one should do the fit. And since diffusion is a sensitive process, doing a 'wrong' fit may lead to unstable results.

We conclude from the test results that although theoretically it is possible to recover (average) interdiffusion coefficients in ternary systems, the results should be treated with great care. Results obtained from a single measurement may not be reliable. If possible, a single diffusion couple

interdiffusion coefficients become less reliable as the intervals $[x_L, x_R]$ become smaller. So, while the method from Dayananda and Sohn may show good results if the correct interdiffusion coefficients are found, it may also show good results if 'bad' interdiffusion coefficients are found but one uses enough ranges over which average interdiffusion coefficients are computed. The method we use to test the computed coefficients is a finite volume scheme paired with Euler forward time integration, starting from the initial concentrations. If the computed coefficients are bad, then the 'recoverd' concentration profile is expected to be bad as well.

[8] We only computed average interdiffusion coefficients over certain concentration ranges. To use these computed coefficients in the method to be discussed in Chapter 4 we used a (second order) polynomial interpolation to obtain interdiffusion coefficients as a function of one of the concentrations.

should be measured at different times and the resulting average interdiffusion coefficients should be compared. If they coincide within reasonable margins of error, this could be an indication that the 'correct' coefficients have been computed. It would be even better if multiple diffusion couples could be prepared with crossing diffusion paths so that the diffusion coefficients at some specific concentrations can be computed using the 'normal' Boltzmann-Matano / Den Broeder as explained in Subsection 2.3.2. For thin films the latter may be practically impossible to achieve though.

2.4 Temperature Dependence

So far we have assumed the temperature to be constant at all times during the experiments. If temperature is allowed to vary, it is generally assumed that diffusion coefficients D have a temperature dependence that follows the so called Arrhenius formula [2]:

$$D(T) = \mathring{D} \exp\left(-\frac{Q}{RT}\right).$$

Here $\mathring{D}$ is a temperature-independent coefficient called the *pre-exponential factor*, Q is called the *activation energy* and is also assumed to be independent of temperature. Finally R is the so called *universal gas constant.* The activation energy is related to the energy required for a single atom to make a jump in the crystal lattice. The higher the activation energy, the more difficult it is for atoms to jump and hence the diffusion coefficient will be lower. We also see that the diffusion coefficient increases as T increases. That should not be too surprising. The atoms within a crystal lattice will vibrate around their equilibrium lattice positions with higher energies as temperature increases, making it easier for them to jump within the lattice.

To find $\mathring{D}$ and Q, note that

$$\begin{aligned} \log\left[D(T)\right] &= \log\left[\mathring{D}\exp\left(-\frac{Q}{RT}\right)\right] \\ &= \log\left[\mathring{D}\right] - \frac{Q}{RT}. \end{aligned}$$

It follows that in a plot of $\log\left[D(T)\right]$ versus $1/T$ the slope of the graph is given by $-Q/R$ and the value at which the graph intersects the y-axis gives us $\log\left[\mathring{D}\right]$. Note that at least two data points are needed - that is, measurements at at least two different temperatures need to be performed - in order to compute the slope and the intersection with the y-axis.

In general D and hence $\mathring{D}$ may be concentration dependent. In that case we would have to fix concentrations first and then perform the above steps to find $\mathring{D}$ at these specific concentrations. A similar thing applies to Q.

3 Image Deblurring Methods

An important tool in analyzing CIGS samples produced at TNO/Solliance is *cross-section Energy Dispersive X-Ray Spectroscopy,* hereinafter referred to as 'EDX'. EDX measurements can be used to produce profiles atomic fraction versus depth in a sample. The method can be explained as follows. After a substrate - think of a piece of glass precursor - has been produced it is cut into smaller samples. Then high energy electrons are fired at the cut plane of the samples in a vacuum environment. These electrons kick out electrons of the atoms present in the sample, basically creating electron vacancies in the atoms. The electron vacancies in the atoms are subsequently filled by other electrons within the atoms and in doing so, X-rays are emitted. The emitted X-rays are different for different kind of atoms. By localizing the emitted X-ray at different positions from the cut plane it is possible to generate a profile of atomic fractions versus depth. By making certain assumptions on the molar volume within the samples - see Chapter 2.1.3 - the atomic fraction profiles can be converted into concentration profiles.

There is a problem with the EDX measurements though. We are interested in creating atomic fraction profiles of layers that have a depth on micrometer level. In order to get clear images on this length scale the measurement device must have a resolution on the nanometer scale. That is, it must be able to distinguish between things that are only a few nanometers apart. Unfortunately, as it turns out, the resolution of the EDX measurement device is insufficient to produce clear profiles on the length scale we are interested in. As a result, the measured profiles appear blurred. For instance, when sharp interfaces between different components are expected, the profiles would suggest that the components have diffused into one another. To get a better understanding of the different processes that lead to the formation of CIGS layers - and also to be able to obtain concentration profiles that can be used for extracting interdiffusion coefficients - it is necessary to deblur the measured profiles. In this chapter we will formulate the above problem in a more mathematical language and work on deblurring methods.

The main source used in writing this chapter is the book *Computational Methods for Inverse Problems* by Vogel [55].

3.1 Deblurring as an Inverse Problem

3.1.1 Inverting Ill-Conditioned Matrices

Let Ω be a domain in $\mathbb{R}^d$ for some $d \in \mathbb{N}$. A black-and-white image on Ω can be represented by a function $f : \Omega \to [0, 1]$ that gives the intensity of the image at point $x \in \Omega$. For example, for a black-and-white image we could have $f(0)$ represent a black dot, $f(1)$ a white dot and $f(x)$ represents a shade of gray for $x \in (0, 1)$. Now suppose that some 'true' image t is measured with a measurement device. The measurement device outputs a 'measured' image m. Besides introducing noise - to be represented by η - the measurement device may also blur the image. If we let B be the corresponding blurring operator then the measured image m can be represented as

$$m = B(t) + \eta.$$

(This should be read as: measured image = Blurring operator applied to true image + noise). Theoretically speaking, the real image can be obtained from the measured image by inverting the blurring operator:

$$t = B^{-1}(m - \eta). \tag{3.1}$$

Note that we are dealing with a so called inverse problem here: we use the output (m) to recover the input (t). Of course, to compute t as in (3.1) both the blurring operator B and the noise η

need to be known then. In this section we will assume B to be known. In reality, of course, we do not know B and we have to guess it. This is the subject of Section 3.3.

Now, given B, one could try to approximate t by computing $B^{-1}(m)$. If the noise is not too bad, we should at least get close to t, right? Unfortunately, this procedure leads to terrible results - as we will later see in figure 3.2. To explain why this is the case we start by formulating everything in a discrete setting. For simplicity assume Ω is a one-dimensional interval that can be covered by equidistant gridpoints $x_1, \ldots, x_n$ for some $n \in \mathbb{N}$. Define vectors $\boldsymbol{t}, \boldsymbol{m}, \boldsymbol{\eta} \in \mathbb{R}^n$ by

$$\begin{aligned} \boldsymbol{t}_i &:= t(x_i), \\ \boldsymbol{m}_i &:= m(x_i), \\ \boldsymbol{\eta}_i &:= \eta(x_i). \end{aligned}$$

Furthermore, we assume the blurring operator B to be linear. Then its discrete counterpart $\boldsymbol{B}$ can be represented by an $n \times n$ matrix. To obtain $\boldsymbol{t}$ from $\boldsymbol{m}$ we would have to invert the matrix $\boldsymbol{B}$. Of course we know how to invert matrices. But the problem is, to obtain a good approximation to t, we need n to be large. And typically deblurring operators become increasingly *ill-conditioned* as n becomes large. That is, small perturbations in the data - like noise - get amplified and cause the numerical solution to blow. Even the tiny errors resulting from finite arithmetic in computers may cause blow-up in this respect.

The above behaviour can be better understood by considering the *singular value decomposition* of $\boldsymbol{B}$. That is, decompose $\boldsymbol{B}$ as $\boldsymbol{B} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^T$, where $\boldsymbol{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n]$ and $\boldsymbol{V} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n]$ are orthogonal matrices ($\boldsymbol{U}^{-1} = \boldsymbol{U}^T$ and $\boldsymbol{V}^{-1} = \boldsymbol{V}^T$) and $\boldsymbol{\Lambda} = \operatorname{diag}\{\sigma_1, \ldots, \sigma_n\}$ is a diagonal matrix whose entries are the singular values of $\boldsymbol{B}$. Assume that the singular values are ordered $\sigma_1 \geq \ldots \geq \sigma_n$. Then we see that

$$\begin{aligned} \boldsymbol{B}^{-1}(\boldsymbol{m}) &= \boldsymbol{B}^{-1}\left(\boldsymbol{B}(\boldsymbol{t}) + \boldsymbol{\eta}\right) \\ &= \boldsymbol{t} + \boldsymbol{B}^{-1}\boldsymbol{\eta} \\ &= \boldsymbol{t} + \left(\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^T\right)^{-1}\boldsymbol{\eta} \\ &= \boldsymbol{t} + \boldsymbol{V}\boldsymbol{\Lambda}^{-1}\boldsymbol{U}^T\boldsymbol{\eta} \\ &= \boldsymbol{t} + \sum_{i=1}^{n} \sigma_i^{-1} \left\langle \boldsymbol{u}_i, \boldsymbol{\eta} \right\rangle \boldsymbol{v}_i. \end{aligned}$$

We see that, for small σ_i, even a little bit of noise in the direction of $\boldsymbol{u}_i$ results in a huge amount noise in the direction of $\boldsymbol{v}_i$. And usually for blurring operators the singular value $\sigma_n \to 0$ as $n \to \infty$.[9]

3.1.2 Tikhonov Filter

To deal with this issue, a filter could be imposed that excludes the influence of singular values σ_i that are smaller than some chosen threshold $\alpha > 0$. That is, instead of computing an

[9] Hansen [25] mentions how this statement is difficult / impossible to prove in general. As an example, think of a diffusion process being responsible for blurring. For simplicity assume the diffusion coefficient to be constant and equal to 1. Using a finite difference approximation to the spatial derivatives we end up with an $n \times n$ Toeplitz matrix. It can be shown that the eigenvalues of this matrix are given by

$$\lambda_i = 2\left(1 - \cos\left(\frac{\pi i}{n+1}\right)\right).$$

Because the matrix is real and symmetric the eigenvalues coincide with the singular values. We see that $\lambda_n = \sigma_n \to 0$ as $n \to \infty$.

approximation to $\boldsymbol{t}$ using

$$\boldsymbol{B}^{-1}\boldsymbol{m} = \boldsymbol{V}\boldsymbol{\Lambda}^{-1}\boldsymbol{U}^T\boldsymbol{m} = \sum_{i=1}^{n} \sigma_i^{-1} \langle \boldsymbol{u}_i, \boldsymbol{m} \rangle \boldsymbol{v}_i,$$

one computes an approximation $\boldsymbol{t}_\alpha$ to $\boldsymbol{t}$ as a truncated sum

$$\boldsymbol{t}_\alpha = \sum_{\sigma_i \geq \alpha} \sigma_i^{-1} \langle \boldsymbol{u}_i, \boldsymbol{m} \rangle \boldsymbol{v}_i.$$

Following Yagle and Vogel [59, 55] at least three issues can be identified with this method of deblurring. First of all, it is required to compute the singular value decomposition of $\boldsymbol{B}$. For one-dimensional images this is not really a problem. For higher dimensional images this may be too expensive, especially in applications where real-time image enhancement is required. Secondly, truncating the sum may give rise to ringing artifacts comparable to the *Gibbs phenomenon* seen in Fourier series. Thirdly, it is not possible to add constraints - like requiring the solution approximation to be positive everywhere - to the problem.

Luckily, all of the above issues can be dealt with. To deal with the second issue, a filter could be imposed that gradually kills the effect of small singular values rather than sharply cutting them off. For example, consider the Tikhonov filter function $\omega_\alpha, \alpha > 0$ defined as

$$\omega_\alpha(\sigma) = \frac{\sigma}{\sigma + \alpha}.$$

We see that $\omega_\alpha(\sigma) \approx 1$ for large σ while

$$\lim_{\sigma \downarrow 0} \frac{\omega_\alpha(\sigma^2)}{\sigma} = \lim_{\sigma \downarrow 0} \frac{\sigma}{\sigma^2 + \alpha} = 0.$$

Now define

$$\boldsymbol{\Lambda}_\alpha := \operatorname{diag}\left[\omega_\alpha(\sigma_1^2)\sigma_1^{-1}, \ldots, \omega_\alpha(\sigma_n^2)\sigma_n^{-1}\right],$$

and approximate $\boldsymbol{t}$ by

$$\begin{aligned} \boldsymbol{t}_\alpha &:= (\boldsymbol{V}\boldsymbol{\Lambda}_\alpha\boldsymbol{U}^T)\boldsymbol{m}, \\ &= \sum_{i=1}^{n} \omega_\alpha(\sigma_i^2)\sigma_i^{-1} \langle \boldsymbol{u}_i, \boldsymbol{m} \rangle \boldsymbol{v}_i \\ &= \sum_{i=1}^{n} \frac{\sigma_i}{\sigma_i^2 + \alpha} \langle \boldsymbol{u}_i, \boldsymbol{m} \rangle \boldsymbol{v}_i. \end{aligned}$$

The Tikhonov filter functions in such a way that the effect of large singular values is left intact while the effect of small singular values is gradually mitigated. Now it is interesting to note that

$$\begin{aligned} \left(\boldsymbol{B}^T\boldsymbol{B} + \alpha\boldsymbol{I}\right)^{-1}\boldsymbol{B}^T\boldsymbol{m} &= \left(\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{U}^T\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^T + \alpha\boldsymbol{I}\right)^{-1}\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{U}^T\boldsymbol{m} \\ &= \left(\boldsymbol{V}\left(\boldsymbol{\Lambda}^2 + \alpha\boldsymbol{I}\right)\boldsymbol{V}^{-1}\right)^{-1}\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{U}^T\boldsymbol{m} \\ &= \boldsymbol{V}\left(\boldsymbol{\Lambda}^2 + \alpha\boldsymbol{I}\right)^{-1}\boldsymbol{V}^{-1}\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{U}^T\boldsymbol{m} \\ &= \boldsymbol{V}\left(\boldsymbol{\Lambda}^2 + \alpha\boldsymbol{I}\right)^{-1}\boldsymbol{\Lambda}\boldsymbol{U}^T\boldsymbol{m} \\ &= \sum_{i=1}^{n} \frac{\sigma_i}{\sigma_i^2 + \alpha} \langle \boldsymbol{u}_i, \boldsymbol{m} \rangle \boldsymbol{v}_i \\ &= \boldsymbol{t}_\alpha. \end{aligned}$$

In other words, $\boldsymbol{t}_\alpha$ can be computed as $\boldsymbol{t}_\alpha = \left(\boldsymbol{B}^T\boldsymbol{B} + \alpha\boldsymbol{I}\right)^{-1}\boldsymbol{B}^T\boldsymbol{m}$. We see that it is in fact not necessary to compute the singular value decomposition in order to compute $\boldsymbol{t}_\alpha$.

But keep in mind that the result so far is specific to the choice of the filter function. For filter functions other than the Tikhonov filter function ω_α we still have not solved the issue of needing to compute the singular value decomposition of $\boldsymbol{B}$. In order to do so, it is convenient to reformulate the inverse problem as a *minimization problem*.

3.2 Deblurring as a Minimization Problem

3.2.1 Tikhonov Functional

Suppose we want to minimize a function $h : \mathbb{R} \to \mathbb{R}$. Assuming the function to be *convex*, all we have to do is find its critical points. That is, we have to find the points where the derivative h' is equal to zero. As it turns out, in our situation there is a convex *functional* $J_\alpha : \mathbb{R}^n \to \mathbb{R}$ - let's call it the *Tikhonov functional* - defined as

$$J_\alpha(\boldsymbol{f}) = \frac{1}{2}\|\boldsymbol{B}\boldsymbol{f} - \boldsymbol{m}\|_2^2 + \frac{\alpha}{2}\|\boldsymbol{f}\|_2^2,$$

whose gradient at $\boldsymbol{f} \in \mathbb{R}^n$ is given by

$$\boldsymbol{B}^T\left(\boldsymbol{B}\boldsymbol{f} - \boldsymbol{m}\right) + \alpha\boldsymbol{f}. \tag{3.2}$$

This fact will be derived further below in Subsection 3.2.6. For now, let's just work with it. Setting the gradient equal to zero and solving for $\boldsymbol{f}$ gives

$$\boldsymbol{f} = \left(\boldsymbol{B}^T\boldsymbol{B} + \alpha\boldsymbol{I}\right)^{-1}\boldsymbol{B}^T\boldsymbol{m}. \tag{3.3}$$

We see that the right-hand side coincides with the expression we found earlier for $\boldsymbol{t}_\alpha$: the solution approximation to the true image $\boldsymbol{t}$ obtained under the Tikhonov filter. By the above arguments, we can say that $\boldsymbol{t}_\alpha$ is a minimizer of the functional J_α. Hence the problem of deblurring $\boldsymbol{m}$ using the Tikhonov filter can be reformulated as: find $\boldsymbol{t}_\alpha$ such that

$$\boldsymbol{t}_\alpha = \operatorname{argmin}_{\boldsymbol{f}\in\mathbb{R}^n} J_\alpha(\boldsymbol{f}) = \operatorname{argmin}_{\boldsymbol{f}\in\mathbb{R}^n} \frac{1}{2}\|\boldsymbol{B}\boldsymbol{f} - \boldsymbol{m}\|_2^2 + \frac{\alpha}{2}\|\boldsymbol{f}\|_2^2. \tag{3.4}$$

Thinking of the deblurring problem as a minimization problem has several advantages. First of all it is not necessary to compute singular value decompositions to solve the minimization problem. Furthermore, we might impose conditions on the solution approximations by minimizing over different domains. For example, if we want to find positive solution approximations only then we should minimize over $\{\boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{x}_i \geq 0,\ 1 \leq i \leq n\}$ instead of the whole $\mathbb{R}^n$.

In the minimization problem (3.4), the first term of J_α (i.e. $\frac{1}{2}\|\boldsymbol{B}\boldsymbol{x} - \boldsymbol{m}\|_2^2$) is 'responsible' for making sure that the solution approximation fits the measured data. The second term (i.e. $\frac{\alpha}{2}\|\boldsymbol{f}\|_2^2$) comes from the Tikhonov filter function and adds a penalty to solution approximations having large Euclidean norms. The larger the α, the larger the penalty. This is how the method prevents blow-up caused by noise in the measured data. As a result, the method may be biased towards solutions that are 'dragged down' (as can be seen in the last plot of Figure 3.2).

3.2.2 Generalized Tikhonov Penalty Functional

Alternatively, we might penalize on the 2−norm of the *derivative* of $\boldsymbol{f}$ instead of penalizing on the 2−norm of $\boldsymbol{f}$ itself. That way we only penalize oscillations in the solutions without

introducing a bias for smaller normed solutions (i.e. without dragging down the solution). Since noise is highly oscillatory blow-up should be prevented. To make this method more precise for a one-dimensional setting (the concenration profiles we want to deblur have only one spatial dimension), let $\boldsymbol{L}$ be the so called first-difference matrix defined as

$$\boldsymbol{L} = \begin{pmatrix} -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{pmatrix} \in \mathbb{R}^{(n-1)\times n}. \tag{3.5}$$

Under the assumption that $\boldsymbol{f}$ is sampled from some continuous image f on equidistant gridpoints with a distance Δx between the gridpoints, we define the discrete derivative of $\boldsymbol{f}$ as $\boldsymbol{Df}$, where

$$\boldsymbol{D} := \frac{1}{\Delta x}\boldsymbol{L}$$

is the *first-derivative matrix*. The corresponding minimization problem can then be formulated as: find $\boldsymbol{t}_\alpha$ such that

$$\boldsymbol{t}_\alpha = \text{argmin}_{\boldsymbol{f}\in\mathbb{R}^n} J_\alpha(\boldsymbol{f}) = \text{argmin}_{\boldsymbol{f}\in\mathbb{R}^n} \frac{1}{2}\|\boldsymbol{Bf} - \boldsymbol{m}\|_2^2 + \frac{\alpha}{2}\|\boldsymbol{Df}\|_2^2.$$

One could also try to penalize on both $\|\boldsymbol{f}\|_2^2$ and $\|\boldsymbol{Df}\|_2^2$. Furthermore, note that $\left(\boldsymbol{D}^T\boldsymbol{D}\right)\boldsymbol{f}$ can be interpreted as the discrete second derivative of $\boldsymbol{f}$. Then, if we let $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2) \in \mathbb{R}^3$, we define a *generalized Tikhonov penalty functional* $\text{Tikhonov}_{\boldsymbol{\alpha}} : \mathbb{R}^n \to \mathbb{R}$ as

$$\text{Tikhonov}_{\boldsymbol{\alpha}}(\boldsymbol{f}) = \alpha_0\|\boldsymbol{f}\|_2^2 + \alpha_1\|\boldsymbol{Df}\|_2^2 + \alpha_2\|\left(\boldsymbol{D}^T\boldsymbol{D}\right)\boldsymbol{f}\|_2^2.$$

We could generalize this even further by including third and higher order derivatives as well. We will stop at the second derivative though.

3.2.3 Total Variation Penalty Functional

A problem with including derivatives in the penalty functional as in $\text{Tikhonov}_{\boldsymbol{\alpha}}$) is that sharp edges may not be recovered in the deblurring process. Indeed, if we consider a piecewise constant signal, then the derivatives blow up at the jumps and the generalized Tikhonov functional will prevent us from recovering these sharp edges. As explained in the introduction to this chapter, we want to deblur concentration profiles of components that diffuse in one another. However, it is expected that - at least at room temperature - some of the components will hardly participate in the diffusion process. As a result there may be sharp edges in the concentration profile. Furthermore, newly formed line compounds (as discussed in Chapter (2)) may give rise to sharp edges as well. Hence the Tikhonov filter may not be an appropriate filter in our case. Luckily, now that the problem has been reformulated as a minimization problem, it is easy to introduce other types of filters as well by working with other types of penalty functionals.

One possible choice would be to penalize on the *total variation*, as proposed by Rudin et al [50]. The total variation of a (smooth enough) function $f : \Omega \to \mathbb{R}$ is defined as

$$\text{TV}(f) := \int_\Omega |\nabla f(x)| dx = \|\nabla f\|_1.$$

Note that having oscillations significantly adds to the total variation of a function. Since noise is usually highly oscillatory, one can image how penalizing total variation surpresses noise.

For the discrete counterpart we define the total variation as

$$TV(\boldsymbol{f}) := \|\boldsymbol{Lf}\|_1,$$

where $\boldsymbol{L}$ was defined in (3.5). Note that

$$\|\boldsymbol{L}\boldsymbol{f}\|_1 = \sum_{i=1}^{n-1} |\boldsymbol{f}_{i+1} - \boldsymbol{f}_i|$$

and from the latter expression one observes that sharp edges or discontinuities in the solution pose no problem.

3.2.4 Tikhonov and Total Variation Combined

A problem with Total Variation deblurring is that the resulting images may appear too 'blocky'. It may tend to create sharp edges where in reality there should be smooth edges (see Figure 3.5). Perhaps combining the Total Variation penalty functional together with the generalized Tikhonov penalty functional gives us enough flexibility to reproduce images that have both sharp edges and smooth edges. To this end, we combine both the generalized Tikhonov penalty functional and the Total Variation penalty functional in the functional $J_{\boldsymbol{\alpha},\beta} : \mathbb{R}^n \to \mathbb{R}$ defined as:

$$J_{\boldsymbol{\alpha},\beta}(\boldsymbol{f}) := \frac{1}{2}\|\boldsymbol{B}\boldsymbol{f} - \boldsymbol{m}\|_2^2 + \frac{1}{2}\text{Tikhonov}_{\boldsymbol{\alpha}}(\boldsymbol{f}) + \beta\text{TV}(\boldsymbol{f}).$$

The corresponding minimization problem can then be formulated as: find $\boldsymbol{t}$ such that

$$\boldsymbol{t} = \text{argmin}_{\boldsymbol{f}\in\mathbb{R}^n} J_{\boldsymbol{\alpha},\beta}(\boldsymbol{f}). \tag{3.6}$$

3.2.5 Smooth Approximation to Euclidean Norm

To solve the minimization problem (3.6), we compute the gradient of $J_{\alpha_0,\alpha_1,\alpha_2,\beta}$ and set it equal to zero. But there's one issue here: the Euclidean norm that is used in the Total Variation penalty functional is not differentiable at the origin and this messes up the differentiability of $J_{\alpha_0,\alpha_1,\alpha_2,\beta}$. To overcome this issue we will with a smooth approximation to the Euclidean norm. As suggested in Chapter 8 of the book by Vogel [55] we let $\gamma > 0$ be some small parameter and work with the approximation

$$\sqrt{x^2 + \gamma^2} \approx \sqrt{x^2} = |x|.$$

To approxiate the total variation of a vector, we first define a function $\psi_\gamma : \mathbb{R} \to \mathbb{R}$ by

$$\psi_\gamma(x^2) = 2\sqrt{x^2 + \gamma^2}.$$

Now define TV_γ by

$$TV_\gamma(\boldsymbol{f}) \quad := \quad \frac{1}{2}\sum_{i=1}^{n-1} \psi_\gamma\left([\boldsymbol{D}\boldsymbol{f}]_i^2\right)\Delta x,$$

where $\boldsymbol{D}$ is the first-derivative matrix defined earlier. Then $\mathrm{TV}_\gamma(\boldsymbol{f})$ smoothly approximates $\mathrm{TV}(\boldsymbol{f})$ since

$$\begin{aligned} TV_\gamma(\boldsymbol{f}) &= \frac{1}{2}\sum_{i=1}^{n-1}\psi_\gamma\left([\boldsymbol{D}\boldsymbol{f}]_i^2\right)\Delta x \\ &= \sum_{i=1}^{n-1}\left(\sqrt{\left(\frac{\boldsymbol{f}_{i+1}-\boldsymbol{f}_i}{\Delta x}\right)^2+\gamma^2}\right)\Delta x \\ &= \sum_{i=1}^{n-1}\sqrt{\left(\boldsymbol{f}_{i+1}-\boldsymbol{f}_i\right)^2+(\gamma\Delta x)^2} \\ &\approx \sum_{i=1}^{n-1}\sqrt{\left(\boldsymbol{f}_{i+1}-\boldsymbol{f}_i\right)^2} \\ &= \sum_{i=1}^{n-1}\left|\boldsymbol{f}_{i+1}-\boldsymbol{f}_i\right| \\ &= \mathrm{TV}(\boldsymbol{f}). \end{aligned}$$

Our new functional to be to be minimized becomes:

$$J_{\boldsymbol{\alpha},\beta,\gamma}(\boldsymbol{f}) := \frac{1}{2}\|\boldsymbol{B}\boldsymbol{f}-\boldsymbol{m}\|_2^2+\frac{1}{2}\mathrm{Tikhonov}_{\boldsymbol{\alpha}}(\boldsymbol{f})+\beta\mathrm{TV}_\gamma(\boldsymbol{f}). \tag{3.7}$$

We will refer to the first term in $J_{\boldsymbol{\alpha},\beta,\gamma}$ as the *Data-Fitting term*, the second as the *Tikhonov term* and the third as the *Total Variation term*. In order to minimize $J_{\boldsymbol{\alpha},\beta,\gamma}$ we compute its gradient first. Let's do that.

3.2.6 The Gradient of $J_{\boldsymbol{\alpha},\beta,\gamma}$

Computing the gradient of $J_{\boldsymbol{\alpha},\beta,\gamma}$ at some vector $\boldsymbol{f}$ boils down to computing the directional derivative of $J_{\boldsymbol{\alpha},\beta,\gamma}$ at $\boldsymbol{f}$ in an arbitrary direction $\boldsymbol{g}$ and expressing the result in an appropriate form (inner product of 'something' with $\boldsymbol{g}$, where something will then be the gradient of $J_{\boldsymbol{\alpha},\beta,\gamma}$ at $\boldsymbol{f}$). See Chapter 2 of Vogel [55] for the underlying mathematical details. Therefore, we proceed by computing the directional derivatives of each of the terms of $J_{\boldsymbol{\alpha},\beta,\gamma}$ separately. Then we combine them and derive an expression for the gradient of $J_{\boldsymbol{\alpha},\beta,\gamma}$ at some arbitrary vector $\boldsymbol{f}$.

For the Data-Fitting term, the directional derivative at $\boldsymbol{f}$ in the direction of $\boldsymbol{g}$ is given by

$$\begin{aligned} \frac{d}{d\tau}\frac{1}{2}\|\boldsymbol{B}(\boldsymbol{f}+\tau\boldsymbol{g})-\boldsymbol{m}\|_2^2 \;|_{\tau=0} &= \frac{d}{d\tau}\frac{1}{2}\left\langle\boldsymbol{B}(\boldsymbol{f}+\tau\boldsymbol{g})-\boldsymbol{m},\boldsymbol{B}(\boldsymbol{f}+\tau\boldsymbol{g})-\boldsymbol{m}\right\rangle_2|_{\tau=0} \\ &= \frac{1}{2}\left[\langle\boldsymbol{B}\boldsymbol{f},\boldsymbol{B}\boldsymbol{g}\rangle_2+\langle\boldsymbol{B}\boldsymbol{g},\boldsymbol{B}\boldsymbol{f}\rangle_2-\langle\boldsymbol{B}\boldsymbol{g},\boldsymbol{m}\rangle_2-\langle\boldsymbol{m},\boldsymbol{B}\boldsymbol{g}\rangle_2\right] \\ &= \langle\boldsymbol{B}\boldsymbol{f},\boldsymbol{B}\boldsymbol{g}\rangle_2-\langle\boldsymbol{m},\boldsymbol{B}\boldsymbol{g}\rangle_2 \\ &= \langle\boldsymbol{B}\boldsymbol{f}-\boldsymbol{m},\boldsymbol{B}\boldsymbol{g}\rangle_2 \\ &= \left\langle\boldsymbol{B}^T(\boldsymbol{B}\boldsymbol{f}-\boldsymbol{m}),\boldsymbol{g}\right\rangle_2. \end{aligned} \tag{3.8}$$

For the Tikhonov term, note that if $\boldsymbol{M}$ is some matrix, then

$$\begin{aligned} \frac{d}{d\tau}\frac{\alpha}{2}\|\boldsymbol{M}(\boldsymbol{f}+\tau\boldsymbol{g})\|_2^2\,|_{\tau=0} &= \frac{d}{d\tau}\frac{\alpha}{2}\left\langle\boldsymbol{M}\boldsymbol{f}+\tau\boldsymbol{M}\boldsymbol{g},\boldsymbol{M}\boldsymbol{f}+\tau\boldsymbol{M}\boldsymbol{g}\right\rangle_2|_{\tau=0} \\ &= \frac{\alpha}{2}\left[\langle\boldsymbol{M}\boldsymbol{g},\boldsymbol{M}\boldsymbol{f}\rangle_2+\langle\boldsymbol{M}\boldsymbol{f},\boldsymbol{M}\boldsymbol{g}\rangle_2\right] \\ &= \left\langle\alpha\boldsymbol{M}^T\boldsymbol{M}\boldsymbol{f},\boldsymbol{g}\right\rangle_2. \end{aligned}$$

It follows that the directional derivative at $\boldsymbol{f}$ in the direction of $\boldsymbol{g}$ of the Tikhonov term is given by

$$\frac{d}{d\tau}\text{Tikhonov}_{\boldsymbol{\alpha}}(\boldsymbol{f}+\tau\boldsymbol{g}) = \left\langle \alpha_0\boldsymbol{f} + \alpha_1\boldsymbol{D}^T\boldsymbol{D}\boldsymbol{f} + \alpha_2(\boldsymbol{D}^T\boldsymbol{D})^T(\boldsymbol{D}^T\boldsymbol{D})\boldsymbol{f}, \boldsymbol{g}\right\rangle_2 \tag{3.9}$$

For the Total Variation term, we use the differentiability of ψ together with the chain rule to find that

$$\begin{aligned}\frac{d}{d\tau}\beta\text{TV}_\gamma(\boldsymbol{f}+\tau\boldsymbol{g})\mid_{\tau=0} &= \frac{\beta}{2}\sum_{i=1}^{n}\frac{d}{d\tau}\psi_\gamma\left([\boldsymbol{D}\boldsymbol{f}+\tau\boldsymbol{D}\boldsymbol{g}]_i^2\right)\Delta x\mid_{\tau=0}\\ &= \frac{\beta}{2}\sum_{i=1}^{n}\psi'_\gamma\left([\boldsymbol{D}\boldsymbol{f}]_i^2\right)[\boldsymbol{D}\boldsymbol{f}]_i\,[\boldsymbol{D}\boldsymbol{g}]_i\,\Delta x\end{aligned}$$

If we define a matrix $\boldsymbol{\Psi}'_{\boldsymbol{f}}$ as

$$\boldsymbol{\Psi}'_{\boldsymbol{f}} := \begin{pmatrix}\psi'_\gamma\left([\boldsymbol{D}\boldsymbol{f}]_i^2\right) & & \\ & \ddots & \\ & & \psi'_\gamma\left([\boldsymbol{D}\boldsymbol{f}]_i^2\right)\end{pmatrix}$$

then the expression for the directional derivative can be more conveniently written as

$$\begin{aligned}\frac{d}{d\tau}\text{TV}_\gamma(\boldsymbol{f}+\tau\boldsymbol{g})\mid_{\tau=0} &= \beta\Delta x\,(\boldsymbol{D}\boldsymbol{g})^T\,\boldsymbol{\Psi}'_{\boldsymbol{f}}\boldsymbol{D}\boldsymbol{f}\Delta x\\ &= \beta\left\langle\Delta x\boldsymbol{\Psi}'_{\boldsymbol{f}}\boldsymbol{D}\boldsymbol{f}, \boldsymbol{D}\boldsymbol{g}\right\rangle_2\\ &= \left\langle\beta\left(\Delta x\boldsymbol{D}^T\boldsymbol{\Psi}'_{\boldsymbol{f}}\boldsymbol{D}\right)\boldsymbol{f}, \boldsymbol{g}\right\rangle_2.\end{aligned} \tag{3.10}$$

Putting (3.8), (3.9) and (3.10) together we find that

$$\begin{aligned}&\frac{d}{d\tau}J_{\boldsymbol{\alpha},\beta,\gamma}(\boldsymbol{f}+\tau\boldsymbol{g})\\ =\ &\left\langle\boldsymbol{B}^T(\boldsymbol{B}\boldsymbol{f}-\boldsymbol{m}) + \alpha_0\boldsymbol{f} + \alpha_1\boldsymbol{D}^T\boldsymbol{D}\boldsymbol{f} + \alpha_2(\boldsymbol{D}^T\boldsymbol{D})^T(\boldsymbol{D}^T\boldsymbol{D})\boldsymbol{f} + \beta\left(\Delta x\boldsymbol{D}^T\boldsymbol{\Psi}'_{\boldsymbol{f}}\boldsymbol{D}\right)\boldsymbol{f}, \boldsymbol{g}\right\rangle_2.\end{aligned}$$

The term in the left slot of the inner product is called the gradient of $J_{\boldsymbol{\alpha},\beta,\gamma}$ at $\boldsymbol{f}$ - to be denoted by $\text{Grad}J_{\boldsymbol{\alpha},\beta,\gamma}(\boldsymbol{f})$ (again, see Chapter 2 of Vogel [55] for more mathematical details). That is,

$$\text{Grad}J_{\boldsymbol{\alpha},\beta,\gamma}(\boldsymbol{f}) = \boldsymbol{B}^T(\boldsymbol{B}\boldsymbol{f}-\boldsymbol{m}) + \alpha_0\boldsymbol{f} + \alpha_1\boldsymbol{D}^T\boldsymbol{D}\boldsymbol{f} + \alpha_2(\boldsymbol{D}^T\boldsymbol{D})^T(\boldsymbol{D}^T\boldsymbol{D})\boldsymbol{f} + \beta\left(\Delta x\boldsymbol{D}^T\boldsymbol{\Psi}'_{\boldsymbol{f}}\boldsymbol{D}\right)\boldsymbol{f}.$$

Note that the earlier claim about the gradient of J_α at $\boldsymbol{f}$ is seen to be true when setting α_1, α_2 and β equal to zero (see equation (3.2)).

3.2.7 Lagged Diffusivity

If $J_{\boldsymbol{\alpha},\beta,\gamma}$ attains its minimum at $\boldsymbol{f}$, then $\text{Grad}J_{\boldsymbol{\alpha},\beta,\gamma}(\boldsymbol{f}) = \boldsymbol{0}$. As a result, assuming $J_{\boldsymbol{\alpha},\beta,\gamma}$ to be convex, we only need to look for the zeros $\text{Grad}J_{\boldsymbol{\alpha},\beta,\gamma}$ in order to minimize $J_{\boldsymbol{\alpha},\beta,\gamma}$. Let's try to do that:

$$\begin{aligned}&\text{Grad}J_{\boldsymbol{\alpha},\beta,\gamma}(\boldsymbol{f}) = 0\\ \Leftrightarrow\ &\boldsymbol{B}^T(\boldsymbol{B}\boldsymbol{f}-\boldsymbol{m}) + \alpha_0\boldsymbol{f} + \alpha_1\boldsymbol{D}^T\boldsymbol{D}\boldsymbol{f} + \alpha_2(\boldsymbol{D}^T\boldsymbol{D})^T(\boldsymbol{D}^T\boldsymbol{D})\boldsymbol{f} + \beta\left(\Delta x\boldsymbol{D}^T\boldsymbol{\Psi}'_{\boldsymbol{f}}\boldsymbol{D}\right)\boldsymbol{f} = 0\\ \Leftrightarrow\ &\left[\boldsymbol{B}^T\boldsymbol{B} + \alpha_0\boldsymbol{I} + \alpha_1\boldsymbol{D}^T\boldsymbol{D} + \alpha_2(\boldsymbol{D}^T\boldsymbol{D})^T(\boldsymbol{D}^T\boldsymbol{D}) + \beta\left(\Delta x\boldsymbol{D}^T\boldsymbol{\Psi}'_{\boldsymbol{f}}\boldsymbol{D}\right)\right]\boldsymbol{f} = \boldsymbol{B}^T\boldsymbol{m}\\ \Leftrightarrow\ &\boldsymbol{f} = \left[\boldsymbol{B}^T\boldsymbol{B} + \alpha_0\boldsymbol{I} + \alpha_1\boldsymbol{D}^T\boldsymbol{D} + \alpha_2(\boldsymbol{D}^T\boldsymbol{D})^T(\boldsymbol{D}^T\boldsymbol{D}) + \beta\left(\Delta x\boldsymbol{D}^T\boldsymbol{\Psi}'_{\boldsymbol{f}}\boldsymbol{D}\right)\right]^{-1}\boldsymbol{B}^T\boldsymbol{m}\end{aligned} \tag{3.11}$$

Note that we cannot directly use the expression on the right-hand side to compute $\boldsymbol{f}$ since the right-hand side itself depends on $\boldsymbol{f}$ (in a non-linear fashion) due to the matrix $\boldsymbol{\Psi}'_{\boldsymbol{f}}$. However, if we set $\beta = 0$ for now (no Total Variation penalty), we see that we can directly compute the minmizer $\boldsymbol{f}$ as

$$\boldsymbol{f} = \left[\boldsymbol{B}^T\boldsymbol{B} + \alpha_0\boldsymbol{I} + \alpha_1\boldsymbol{D}^T\boldsymbol{D} + \alpha_2(\boldsymbol{D}^T\boldsymbol{D})^T(\boldsymbol{D}^T\boldsymbol{D})\right]^{-1}\boldsymbol{B}^T\boldsymbol{m}. \tag{3.12}$$

Note that setting α_1 and α_2 equal to zero as well yields an expression obtained earlier in equation (3.3).

As for the general case, note that even though equation (3.11) cannot be used directly, it can be manipulated easily into an iterative scheme for finding the minimizer. Indeed, suppose that $\boldsymbol{t}^k$ is an approximation to the minimizer (i.e. $\boldsymbol{t}^k$ is an approximation to the true image $\boldsymbol{t}$), then we compute a new solution approximation $\boldsymbol{t}^{k+1}$ as

$$\begin{aligned}
& \boldsymbol{t}^{k+1} \\
= \;& \left[\boldsymbol{B}^T\boldsymbol{B} + \alpha_0\boldsymbol{I} + \alpha_1\boldsymbol{D}^T\boldsymbol{D} + \alpha_2(\boldsymbol{D}^T\boldsymbol{D})^T(\boldsymbol{D}^T\boldsymbol{D}) + \beta\left(\Delta x\boldsymbol{D}^T\boldsymbol{\Psi}'_{\boldsymbol{t}^k}\boldsymbol{D}\right)\right]^{-1}\boldsymbol{B}^T\boldsymbol{m} \\
= \;& \boldsymbol{t}_k - \left[\boldsymbol{B}^T\boldsymbol{B} + \alpha_0\boldsymbol{I} + \alpha_1\boldsymbol{D}^T\boldsymbol{D} + \alpha_2(\boldsymbol{D}^T\boldsymbol{D})^T(\boldsymbol{D}^T\boldsymbol{D}) + \beta\left(\Delta x\boldsymbol{D}^T\boldsymbol{\Psi}'_{\boldsymbol{t}^k}\boldsymbol{D}\right)\right]^{-1}\mathrm{Grad}J_{\boldsymbol{\alpha},\beta,\gamma}(\boldsymbol{t}_k).
\end{aligned}$$

This method is called the *method of Lagged Diffusivity* (the 'diffusion coefficient' $\boldsymbol{\Psi}'_{\boldsymbol{t}^k}$ lags behind because it is evaluated using the solution approximation at the old time step) [56]. We continue iterating until the gradient is sufficiently close to zero or when there is no longer any real difference between $\boldsymbol{t}^{k+1}$ and $\boldsymbol{t}^k$.

3.2.8 Existence, Uniqueness and Convergence

So far we have ignored two very important issues. The first issue is related to the question whether the functional $J_{\boldsymbol{\alpha},\beta,\gamma}$ has a minimizer to begin with. And if that's the case, is it perhaps a local minimum or, preferably, a global minimum? And if there is a global minimum, is it unique? Proving these kind of results require advanced mathematical techqniues that are outside the scope of this work. In the case of either a pure Tikhonov problem (i.e. $\beta = 0$) or a pure Total Variation problem ($\boldsymbol{\alpha} = (0, 0, 0)$) such results can be found in the book by Vogel [55]. That does not guarantee anything about the combined method though.

Furthermore, we have written down an iterative scheme for finding a minimizer of $J_{\boldsymbol{\alpha},\beta,\gamma}$. Even if $J_{\boldsymbol{\alpha},\beta,\gamma}$ admits a unique minimizer, how do we know that the iterative scheme will find this solution? Again, the discussion of this issue is outside the scope of this work. For results in this respect, the reader is referred to an article by Vogel and Oman and an article by Chan and Mulet[7, 56].

3.2.9 Newton-Raphson Method

As an alternative to the method of Lagged Diffusivity, we could use the well-known Newton-Raphson method to iteratively find the vectors at which the gradient of $J_{\boldsymbol{\alpha},\beta,\gamma}$ vanishes. Say we are trying to minimize a convex function $h : \mathbb{R} \to \mathbb{R}$ by finding the zeros of its derivative h'. Given some initial approximation x_0 to a zero of h', the Newton-Raphson method tells us to compute an update x_1 as:

$$x_1 = x_0 - \frac{h'(x_0)}{h''(x_0)}. \tag{3.13}$$

We see that the second derivative of h is needed. In the case of our functional $J_{\boldsymbol{\alpha},\beta,\gamma}$ the equivalent of a second derivative is the *Hessian matrix*. We compute it by working out the second directional derivatives of $J_{\boldsymbol{\alpha},\beta,\gamma}$. The computations are similar to what has been done previously in the case

of the gradient (see Section 3.2.6) and they won't be written down here. We only present the results. For the second directional derivative we find that

$$\begin{aligned} & \frac{\partial^2}{\partial\xi\partial\tau} J_{\boldsymbol{\alpha},\beta,\gamma}(\boldsymbol{f}+\tau\mathbf{g}+\xi\mathbf{h})\mid_{\tau,\xi=0} \\ = \; & \left\langle \left(\boldsymbol{B}^T\boldsymbol{B} + \alpha_0\boldsymbol{I} + \alpha_1\boldsymbol{D}^T\boldsymbol{D} + \alpha_2\left(\boldsymbol{D}^T\boldsymbol{D}\right)^T\left(\boldsymbol{D}^T\boldsymbol{D}\right) + \beta\Delta x\boldsymbol{D}^T\left(\boldsymbol{\Psi}'_{\boldsymbol{f}}+\boldsymbol{\Psi}''_{\boldsymbol{f}}\right)\boldsymbol{D}\right)\boldsymbol{g},\boldsymbol{h}\right\rangle_2, \end{aligned}$$

where the matrix $\boldsymbol{\Psi}''_{\boldsymbol{f}}$ is defined as

$$\boldsymbol{\Psi}''_{\boldsymbol{f}} := \begin{pmatrix} 2\psi''\left([\boldsymbol{D}\boldsymbol{f}]_1^2\right)[\boldsymbol{D}\boldsymbol{f}]_1^2 & & \\ & \ddots & \\ & & 2\psi''\left([\boldsymbol{D}\boldsymbol{f}]_n^2\right)[\boldsymbol{D}\boldsymbol{f}]_n^2 \end{pmatrix}.$$

Then the Hessian of $J_{\boldsymbol{\alpha},\beta,\gamma}$ at $(\boldsymbol{f})$ - to be denoted by $\mathrm{Hess}J_{\boldsymbol{\alpha},\beta,\gamma}(\boldsymbol{f})$ - is given by the matrix

$$\mathrm{Hess}J_{\alpha,\beta,\gamma}(\boldsymbol{f}) = \boldsymbol{B}^T\boldsymbol{B} + \alpha_0\boldsymbol{I} + \alpha_1\boldsymbol{D}^T\boldsymbol{D} + \alpha_2\left(\boldsymbol{D}^T\boldsymbol{D}\right)^T\left(\boldsymbol{D}^T\boldsymbol{D}\right) + \beta\Delta x\boldsymbol{D}^T\left(\boldsymbol{\Psi}'_{\boldsymbol{f}}+\boldsymbol{\Psi}''_{\boldsymbol{f}}\right)\boldsymbol{D}.$$

Given a solution approximation $\boldsymbol{t}^k$, we could try to follow (3.13) and compute an update $\boldsymbol{t}^{k+1}$ as

$$\boldsymbol{t}^{k+1} = \boldsymbol{t}^k - \mathrm{Hess}J_{\boldsymbol{\alpha},\beta,\gamma}(\boldsymbol{t}^k)^{-1}\mathrm{Grad}J_{\boldsymbol{\alpha},\beta,\gamma}(\boldsymbol{f}).$$

Better convergence of the method is obtained when performing a so called *line search*. To this end, we first define the *search direction* S_k as

$$S_k := -\mathrm{Hess}J_{\boldsymbol{\alpha},\beta,\gamma}(\boldsymbol{t}^k)^{-1}\mathrm{Grad}J_{\boldsymbol{\alpha},\beta,\gamma}(\boldsymbol{f}).$$

Then we perform a line search along this search direction, i.e. we look for $\tau^* \geq 0$ that satisfies

$$\tau^* := \mathrm{argmin}_{\tau\geq 0} J_{\boldsymbol{\alpha},\beta,\gamma}(\boldsymbol{t}^k + \tau S_k)$$

and we set

$$\boldsymbol{t}^{k+1} = \boldsymbol{t}^k + \tau^* S_k.$$

The reason for doing the line search is that the method may not converge otherwise - even if there is a unique minimizer. Incorporating the line search is not a straightforward task though. See Chapter 3 of the book 'Numerical Optimization' by Nocedal and Wright [39] for details on how one can do this. A `MATLAB` script found on the Mathworks File Exchange Database that implements a line search algorithm is used.[10]

Note that the Newton-Raphson algorithm and the Lagged Diffusivity algorithm look very similar in the end: the term $\beta\Delta x\boldsymbol{D}^T\boldsymbol{\Psi}'_{\boldsymbol{t}^k}\boldsymbol{D}$ for the Lagged Diffusivity is replaced by the term $\beta\Delta x\boldsymbol{D}^T\left(\boldsymbol{\Psi}'_{\boldsymbol{t}^k}+\boldsymbol{\Psi}''_{\boldsymbol{t}^k}\right)\boldsymbol{D}$. We worked out both methods to increase our chances of converging.

3.3 Blurring Operator

3.3.1 Convolution

All this time we have assumed that the blurring matrix $\boldsymbol{B}$ is somehow known. In reality, we do not know $\boldsymbol{B}$ though (in fact, we don't even know if the blurring operator is linear and can be

[10] http://nl.mathworks.com/matlabcentral/fileexchange/44315-newton-method-with-line-search/content/LINESEARCH/line_search.m

represented as a matrix but let's not assume the situation to be that bad though). In this section we will try to describe possible candidates for the blurring operator $\boldsymbol{B}$. We start by remarking that a common way to blur an image t is to *convolve* it with some *response function* r. More specifically, assuming t and r to defined on the whole of $\mathbb{R}$, their *convolution product* $t \star r$ is defined as

$$(t \star r)(x) = \int_{\mathbb{R}} t(x-y)r(y)dy.$$

Note that for a fixed response function r the convolution with r is linear in t (and the other way around as well but we don't need that).

For our purposes it will be more convenient to work in a discrete setting again. Like before, place equidistant gridpoints x_i in our domain and define the vector $\boldsymbol{t}$ by $\boldsymbol{t}_i := f(x_i)$. Similarly, define the vector $\boldsymbol{r}$ by $\boldsymbol{r}_i := r(x_i)$. Then the discrete convolution product $(\boldsymbol{t} \star \boldsymbol{r})$ is defined as

$$(\boldsymbol{t} \star \boldsymbol{r})_i = \sum_{j=-\infty}^{\infty} \boldsymbol{t}_{i-j}\boldsymbol{r}_j.$$

The discrete convolution could be interpreted as follows: the i-th component of the convolved signal, i.e. $(\boldsymbol{t} \star \boldsymbol{r})_i$, is a combination of all components of $\boldsymbol{t}$. The value $\boldsymbol{r}_j$ gives the strength of the influence of $\boldsymbol{t}_{i-j}$ on $(\boldsymbol{t} \star \boldsymbol{r})_i$.

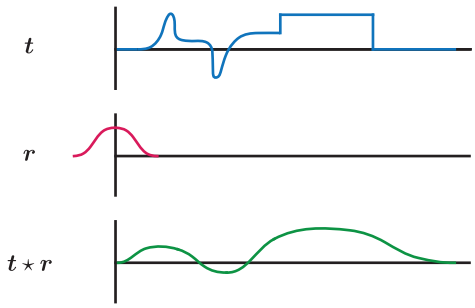

Figure 3.1: Convolution visualised

In our case, we assume each $\boldsymbol{t}_i$ to have influence only over a finite range. This can be achieved by giving the response function $\boldsymbol{r}$ finite support. That is, we assume $\boldsymbol{r}_i$ to be zero for $|i| > k$, where k is some positive integer. In that case the convolution product reduces to

$$(\boldsymbol{t} \star \boldsymbol{r})_i = \sum_{j=-k}^{k} \boldsymbol{t}_{i-j}\boldsymbol{r}_j.$$

In general, the response vector $\boldsymbol{r}$ depends on the measurement device that introduced blur (assuming that the blur is indeed the result of a (linear) convolution). In the case at hand there is no information available on the response function. Therefore we will try to make 'reasonable' guesses. First of all, we have no reason to expect the blur to be stronger in one direction than in others. That is, we expect the response function to be symmetric ($\boldsymbol{r}_i = \boldsymbol{r}_{-i}$). For example, we could give $\boldsymbol{r}$ a Gaussian shape. With this choice, the influence on $(\boldsymbol{t} \star \boldsymbol{r})_i$ is the strongest for $\boldsymbol{t}_i$ and it gradually weakens as $|i-j|$ increases. Another possible option would be to give $\boldsymbol{r}$ the shape of a square. Both choices have been implemented in `MATLAB` as we will see later.

Whatever we end up choosing for $\boldsymbol{r}$, it would be convenient to write the convolution operation in terms of some matrix $\boldsymbol{B}$ - the blurring operator - because then we can readily use the machinery developed in the previous section. Let's see how we can do that.

3.3.2 Boundary Conditions

Suppose we measure a blurred image b over n equidistant gridpoint $x_1, \ldots, x_n$ and define the vector $\boldsymbol{b} \in \mathbb{R}^n$ by $\boldsymbol{b}_i := b(x_i)$. Furthermore, forgetting about noise for now, we assume $\boldsymbol{b}$ to be the convolution product of some 'true' signal $\boldsymbol{t}$ with a finitely supported response function $\boldsymbol{r}$. In matrix-vector form the convolution can be expressed as

$$\begin{pmatrix} \boldsymbol{r}_k & \cdots & \boldsymbol{r}_0 & \cdots & \boldsymbol{r}_{-k} & & & & 0 \\ & \boldsymbol{r}_k & \cdots & \boldsymbol{r}_0 & \cdots & \boldsymbol{r}_{-k} & & & \\ & & \ddots & \ddots & \ddots & \ddots & \ddots & & \\ & & & \boldsymbol{r}_k & \cdots & \boldsymbol{r}_0 & \cdots & \boldsymbol{r}_{-k} & \\ 0 & & & & \boldsymbol{r}_k & \cdots & \boldsymbol{r}_0 & \cdots & \boldsymbol{r}_{-k} \end{pmatrix} \begin{pmatrix} \boldsymbol{t}_{-k+1} \\ \vdots \\ \boldsymbol{t}_1 \\ \vdots \\ \boldsymbol{t}_n \\ \vdots \\ \boldsymbol{t}_{n+k} \end{pmatrix} = \begin{pmatrix} \boldsymbol{b}_1 \\ \vdots \\ \boldsymbol{b}_n \end{pmatrix}. \tag{3.14}$$

We see that $\boldsymbol{b}$, a vector of length n, is determined by the vector $[\boldsymbol{t}_{-k+1}, \ldots, \boldsymbol{t}_{n+k}]$ of length $n + 2k$. To deal with this issue, we have to introduce *boundary conditions*. Before we do this, it is convenient to introduce the following notation:

$$\begin{aligned} \boldsymbol{t}_L &:= [\boldsymbol{t}_{-k+1}, \ldots, \boldsymbol{t}_0] \in \mathbb{R}^k, \\ \boldsymbol{t}_M &:= [\boldsymbol{t}_1, \ldots, \boldsymbol{t}_n] \in \mathbb{R}^k, \\ \boldsymbol{t}_R &:= [\boldsymbol{t}_{n+1}, \ldots, \boldsymbol{t}_{n+k}] \in \mathbb{R}^k, \end{aligned}$$

and

$$\boldsymbol{B}_L := \begin{pmatrix} \boldsymbol{r}_k & \cdots & \boldsymbol{r}_1 \\ 0 & \ddots & \vdots \\ \vdots & \ddots & \boldsymbol{r}_k \\ & \vdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{pmatrix} \in \mathbb{R}^{n \times k},$$

$$\boldsymbol{B}_M := \begin{pmatrix} \boldsymbol{r}_0 & \cdots & \boldsymbol{r}_{-k} & & 0 \\ \vdots & \ddots & \ddots & \ddots & \\ \boldsymbol{r}_k & \ddots & \ddots & \ddots & \boldsymbol{r}_{-k} \\ & \ddots & \ddots & \ddots & \vdots \\ 0 & & \boldsymbol{r}_k & \cdots & \boldsymbol{r}_0 \end{pmatrix} \in \mathbb{R}^{n \times n},$$

$$\boldsymbol{B}_R := \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \vdots & \\ \boldsymbol{r}_{-k} & \ddots & \vdots \\ \vdots & \ddots & 0 \\ \boldsymbol{r}_{-1} & \cdots & \boldsymbol{r}_{-k} \end{pmatrix} \in \mathbb{R}^{n \times k}.$$

With this notation, equation 3.14 can be expressed as

$$\boldsymbol{B}_L \boldsymbol{t}_L + \boldsymbol{B}_M \boldsymbol{t}_M + \boldsymbol{B}_R \boldsymbol{t}_R = \boldsymbol{b}. \tag{3.15}$$

Note that $\boldsymbol{t}_L$ and $\boldsymbol{t}_R$ correspond to the true image left and right of $\boldsymbol{b}$ while $\boldsymbol{t}_M$ is really the true version of $\boldsymbol{b}$. In the end, given $\boldsymbol{b}$, the vector $\boldsymbol{t}_M$ is what we want to compute. A naive, formal calculation shows that

$$\boldsymbol{t}_M = \boldsymbol{B}_M^{-1}\left(\boldsymbol{b} - \boldsymbol{B}_L\boldsymbol{t}_L - \boldsymbol{B}_R\boldsymbol{t}_R\right).$$

Of course this does not work, even under the assumption that $\boldsymbol{B}_M^{-1}$ can be computed and does not cause stability issues, simply because the $\boldsymbol{t}_L$ and $\boldsymbol{t}_R$ are unknown. This is an alternative way of stating that the problem in (3.14) is underdetermined. To deal with this issue we need to make assumptions on $\boldsymbol{t}_L$ and $\boldsymbol{t}_R$. Because $\boldsymbol{t}_L$ and $\boldsymbol{t}_R$ lie outside the boundaries of our domain of interest (we really only care about the true image $\boldsymbol{t}_M$) so we refer to these assumptions as boundary conditions. As we will see, different boundary conditions give rise to different kinds of blurring matrices. We will discuss two different options: Dirichlet boundary conditions and (anti)reflexive boundary conditions.

3.3.3 Dirichlet boundary conditions

Remember from the chapter on interdiffusion coefficients how it was important to assume that the ends of the diffusion couple remain unaffected? Suppose that $\boldsymbol{t}_M$ represents a concentration profile over a domain that covers the interdiffusion zone. Then left of the domain it is natural to assume the concentration to be the same as $\boldsymbol{t}_1$ while right of the domain it would make sense for the concentrations to be equal to $\boldsymbol{t}_n$. That is, we could make the assumption that

$$\begin{aligned}\boldsymbol{t}_L &= [\boldsymbol{t}_1, \ldots, \boldsymbol{t}_1],\\ \boldsymbol{t}_R &= [\boldsymbol{t}_n, \ldots, \boldsymbol{t}_n].\end{aligned}$$

Then we see that

$$\boldsymbol{B}_L\boldsymbol{t}_L = \begin{pmatrix} \boldsymbol{r}_k & \cdots & \boldsymbol{r}_1 \\ 0 & \ddots & \vdots \\ \vdots & \ddots & \boldsymbol{r}_k \\ \vdots & & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{t}_1 \\ \vdots \\ \boldsymbol{t}_1 \end{pmatrix} = \underbrace{\begin{pmatrix} \sum_{i=1}^k r_i & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ r_k & \vdots & & \vdots \\ 0 & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}}_{:=\boldsymbol{B}_L^{\text{Dirichlet}} \in \mathbb{R}^{n\times n}} \boldsymbol{t}_M.$$

In a similar fashion,

$$\boldsymbol{B}_R\boldsymbol{t}_R = \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \vdots & \\ \boldsymbol{r}_{-k} & \ddots & \vdots \\ \vdots & \ddots & 0 \\ \boldsymbol{r}_{-1} & \cdots & \boldsymbol{r}_{-k} \end{pmatrix} \begin{pmatrix} \boldsymbol{t}_k \\ \vdots \\ \boldsymbol{t}_k \end{pmatrix} = \underbrace{\begin{pmatrix} 0 & \cdots & 0 & 0 \\ \vdots & & \vdots & \vdots \\ \vdots & & \vdots & 0 \\ \vdots & & \vdots & r_{-k} \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 0 & \sum_{i=1}^k r_{-i} \end{pmatrix}}_{:=\boldsymbol{B}_L^{\text{Dirichlet}} \in \mathbb{R}^{n\times n}} \boldsymbol{t}_M$$

Now define

$$\boldsymbol{B}^{\text{Dirichlet}} := \boldsymbol{B}_L^{\text{Dirichlet}} + \boldsymbol{B}_M + \boldsymbol{B}_R^{\text{Dirichlet}}.$$

Then equation (3.15) reduces to

$$\boldsymbol{B}^{\text{Dirichlet}}\boldsymbol{t}_M = \boldsymbol{b}$$

and $\boldsymbol{B}^{\text{Dirichlet}}$ is our blurring operator.

3.3.4 Reflexive boundary conditions

For reflexive boundary conditions, we assume the image outside of the domain of interest to be a reflection of the image on the inside. More specifically, we assume

$$\boldsymbol{t}_0 = \boldsymbol{t}_1, \quad \boldsymbol{t}_{-1} = \boldsymbol{t}_2, \quad \ldots, \quad \boldsymbol{t}_{-k+1} = \boldsymbol{t}_k,$$

and

$$\boldsymbol{t}_{n+1} = \boldsymbol{t}_n, \quad \boldsymbol{t}_{n+2} = \boldsymbol{t}_{n-1}, \quad \ldots, \quad \boldsymbol{t}_{n+k} = \boldsymbol{t}_{n-k+1}.$$

Then

$$\begin{aligned} \boldsymbol{t}_L &= [\boldsymbol{t}_k, \ldots, \boldsymbol{t}_1], \\ \boldsymbol{t}_R &= [\boldsymbol{t}_n, \ldots, \boldsymbol{t}_{n-k+1}], \end{aligned}$$

and we see that

$$\boldsymbol{B}_L \boldsymbol{t}_L = \begin{pmatrix} \boldsymbol{r}_k & \cdots & \boldsymbol{r}_1 \\ 0 & \ddots & \vdots \\ \vdots & \ddots & \boldsymbol{r}_k \\ & \vdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{t}_k \\ \vdots \\ \boldsymbol{t}_1 \end{pmatrix} = \underbrace{\begin{pmatrix} 0 & \cdots & 0 & \boldsymbol{r}_k & \cdots & \boldsymbol{r}_1 \\ \vdots & & \vdots & 0 & \ddots & \vdots \\ & & & \vdots & \ddots & \boldsymbol{r}_k \\ & & & & \vdots & 0 \\ \vdots & & & & & \vdots \\ 0 & \cdots & & & \cdots & 0 \end{pmatrix}}_{:=\boldsymbol{B}_L^{\text{Reflexive}} \in \mathbb{R}^{n \times n}} \begin{pmatrix} \boldsymbol{t}_n \\ \vdots \\ \boldsymbol{t}_k \\ \vdots \\ \boldsymbol{t}_1 \end{pmatrix} = \boldsymbol{B}_L^{\text{Reflexive}} \boldsymbol{Y} \boldsymbol{t}_M,$$

where

$$\boldsymbol{Y} = \begin{pmatrix} 0 & & 1 \\ & \iddots & \\ 1 & & 0 \end{pmatrix} \in \mathbb{R}^{n \times n}.$$

is the $n \times n$ matrix that flips vectors upside-down. In a similar fashion we see that

$$\boldsymbol{B}_R \boldsymbol{t}_R = \underbrace{\begin{pmatrix} & & & & 0 \\ \boldsymbol{r}_{k-1} & & & & \\ \vdots & \ddots & & & \\ \boldsymbol{r}_{-1} & \cdots & \boldsymbol{r}_{-k} & 0 & 0 \end{pmatrix}}_{:=\boldsymbol{B}_R^{\text{Reflexive}} \in \mathbb{R}^{n \times n}} \boldsymbol{Y} \boldsymbol{t}_M.$$

Now define

$$\boldsymbol{B}^{\text{Reflexive}} := \boldsymbol{B}_L^{\text{Reflexive}} + \boldsymbol{B}_M + \boldsymbol{B}_R^{\text{Reflexive}}$$

to reduce equation 3.15 to

$$\boldsymbol{B}^{\text{Reflexive}} \boldsymbol{t}_M = \boldsymbol{b}.$$

Yet another possibility would be to have anti-reflexive boundary conditions. The corresponding blurring operator $\boldsymbol{B}^{\text{Anti-Reflexive}}$ is given by

$$\boldsymbol{B}^{\text{Anti-Reflexive}} := \left(2\boldsymbol{B}_L^{\text{Dirichlet}} - \boldsymbol{B}_L^{\text{Reflexive}}\right) + \boldsymbol{B}_M + \left(\boldsymbol{B}_R^{\text{Dirichlet}} - \boldsymbol{B}_R^{\text{Reflexive}}\right).$$

All three options have been implemented in `MATLAB`.

3.4 Constrained Optimization

In the test results that follow in the next section we see that sometimes the deblurring methods result in images that attain negative values. If the images we are trying to deblur correspond to concentration profiles or atomic fraction profiles then negative values make no physical sense. In order to avoid negative values, we could try to minimize the functional $J_{\boldsymbol{\alpha},\beta,\gamma}$ over the set $\mathbb{R}^{n,+} := \{\boldsymbol{f} \in \mathbb{R}^n \mid f_i \geq 0,\ 1 \leq i \leq n\}$ instead of over the whole of $\mathbb{R}^n$. Incorporating constraints such as non-negativity constraints as not a straightforward task: simply projecting each of the iterations $\boldsymbol{t}^k$ onto $\mathbb{R}^{n,+}$ may result in a method that simply does not convergence. In Vogel [55] methods for performing the contrained minimization are discussed for the case $\boldsymbol{\alpha} = (0, 0, 0)$ (i.e. pure Total Variation filter). These methods have been implemented in `MATLAB` and for the test cases considered they seem to be working for general $\boldsymbol{\alpha}$ (i.e. when including the Tikhonov filter) as well.

3.5 Test Results

The deblurring methods discussed in Sections 3.2, 3.3 and 3.4 been implemented in a `MATLAB` script. The script basically works as follows:

1. Load a blurred and noisy concentration profile $\boldsymbol{m}_i$.
2. Make a guess on the response function. Two options have been implemented: a Gaussian shaped and a square shaped response function. A parameter σ determines the 'width' of the shape.
3. Choose a type of boundary condition: Dirichlet, reflexive or anti-reflexive.
4. Choose parameters $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2)$ (Tikhonov) and β and γ (Total Variation).
5. Choose an iterative scheme: Lagged Diffusivity or Newton-Raphson.
6. Choose stopping criteria for the iterative scheme. For example, stop when both the norm of the update $\boldsymbol{t}^{k+1} - \boldsymbol{t}^k$ and / or the norm $\mathrm{Grad} J_{\boldsymbol{\alpha},\beta,\gamma}(\boldsymbol{t}^k)$ are smaller than some specified tolerance. We also want to set an upper limit to the amount of iterations to be performed.
7. To start the iterative procedure, we need to supply an initial guess $\boldsymbol{t}_0$. The most straightforward option is to set $\boldsymbol{t}_0 = \boldsymbol{m}$.
8. Now given $\boldsymbol{t}^k$, compute $\boldsymbol{t}^{k+1}$ using either the Lagged Diffusivity method or the Newton-Raphson method. Quit when stopping criteria are met.

We now present some test cases. To start we create a 'true' image $\boldsymbol{t}$, apply the chosen blurring operator $\boldsymbol{B}$ to it and add some random noise to obtain a 'measured' image $\boldsymbol{m}$. Then we try to recover the true image $\boldsymbol{t}$ using combinations of the Tikhonov and Total Variation methods. As for the true image, we use an image that contains both sharp and smooth edges. To blur the image, in each case a Gaussian shaped response function with reflexive boundary conditions is used. Whenever applicable, the parameter γ was set to 10^{-5}. We try blurring using the Tikhonov method first with $\boldsymbol{\alpha} = (\alpha_0, 0, 0)$ and we do not incorporate the non-negativity constraints yet.

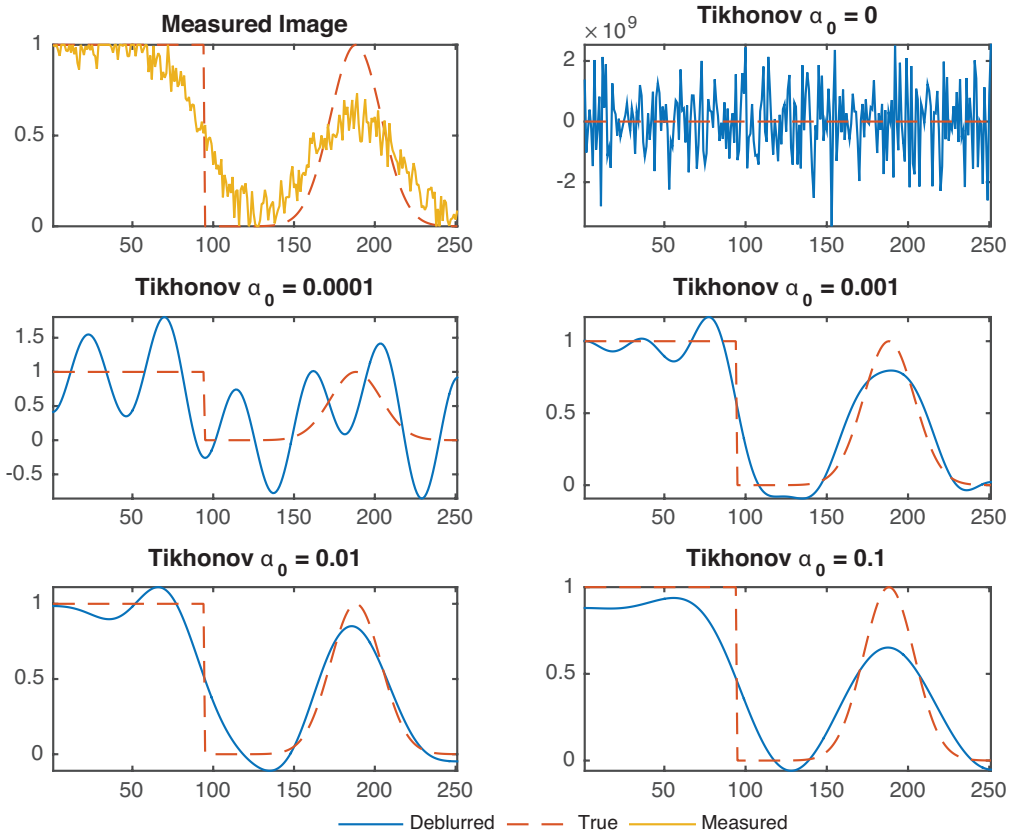

Figure 3.2: Deblurring with a Tikhonov filter for different values of α_0.

The upper left plot in the figure shows the real image and the measured image that is a blurred and noisy version of the real image. In the upper right plot we see the result of trying to deblur without applying any filter (corresponding to $\boldsymbol{\alpha} = (0, 0, 0)$ and $\beta = 0$)). As expected, the result is terrible. By increasing α_0 the results seem to get better. The plot for $\alpha_0 = 10^{-2}$ shows that working with the Tikhonov filter method allows one to obtain smooth edges nicely. The sharp edge is a problem though. Because sharp edges require high frequency components - which we kill using the Tikhonov filter - it is impossible to recover sharp edges. Moreover, the deblurred image gets 'dragged down' as α_0 becomes larger, as can be seen from the plot for $\alpha_0 = 0.1$. This behaviour can be explained from the expression (3.4), which shows that the norm of the image (interpreted as a vector) gets penalized more and more as α_0 increases. And images with lower values have smaller norms. We also see that the deblurred images attain negative values. This is unphysical and illustrates the need for a constrained algorithm that produces non-negative images only. Results obtained with the constrained algorithm are presented later.

For now, let us try deblurring using the Tikhonov method with $\boldsymbol{\alpha} = (0, \alpha_1, 0)$.

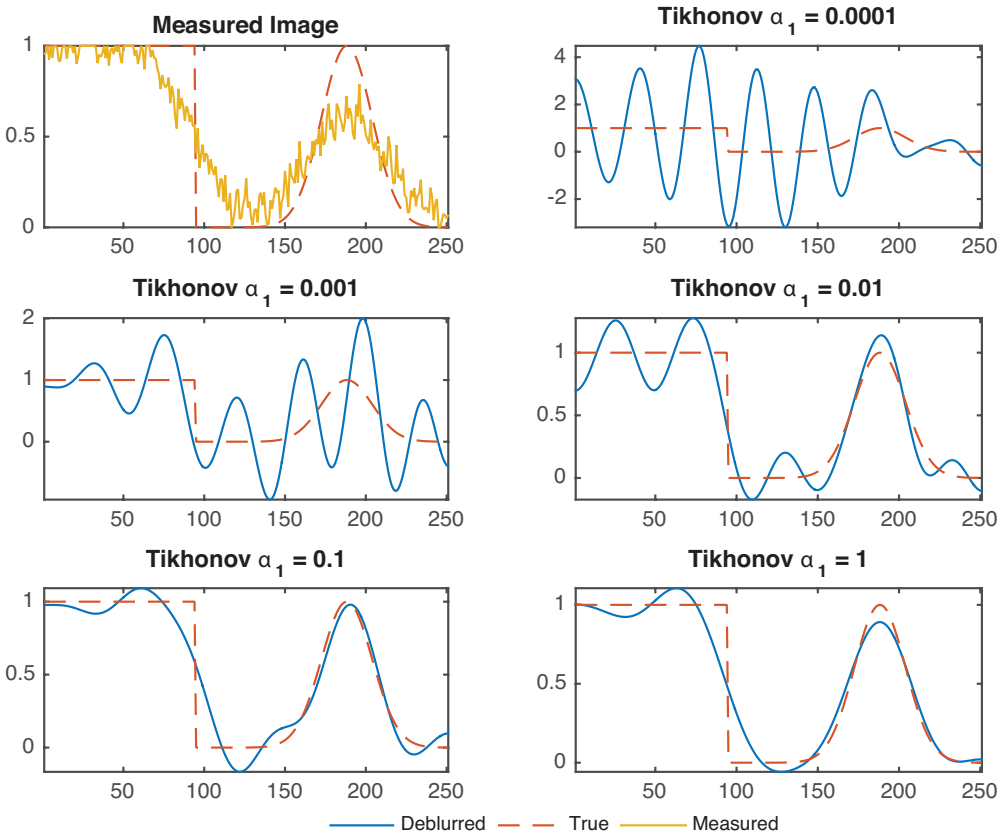

Figure 3.3: Deblurring with a Tikhonov filter for different values of α_1.

We see that the results are similar as the ones obtained for different values α_0. Smooth edges can be recoverd nicely while sharp edges are problematic. The solutions do not get dragged down this time though (because penalizing the norm of the gradient does not necessarily result in lower-valued images). The problem of negative solution values is still present.

Next, we try deblurring using the Tikhonov method with $\boldsymbol{\alpha} = (0, 0, \alpha_2)$.

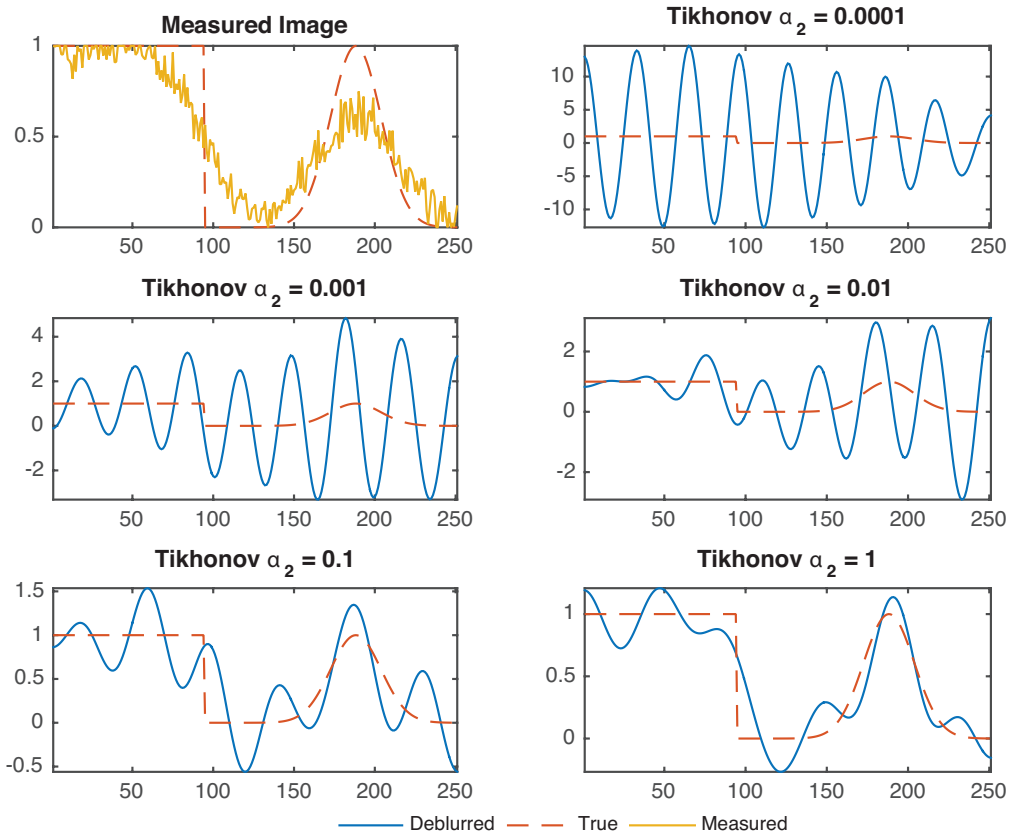

Figure 3.4: Deblurring with a Tikhonov filter for different values of α_2.

The plots appear similar to the ones obtained for α_1 (with the same type of problems associated to them), except that the solutions are even smoother than before.

Let us try the Total Variation method now.

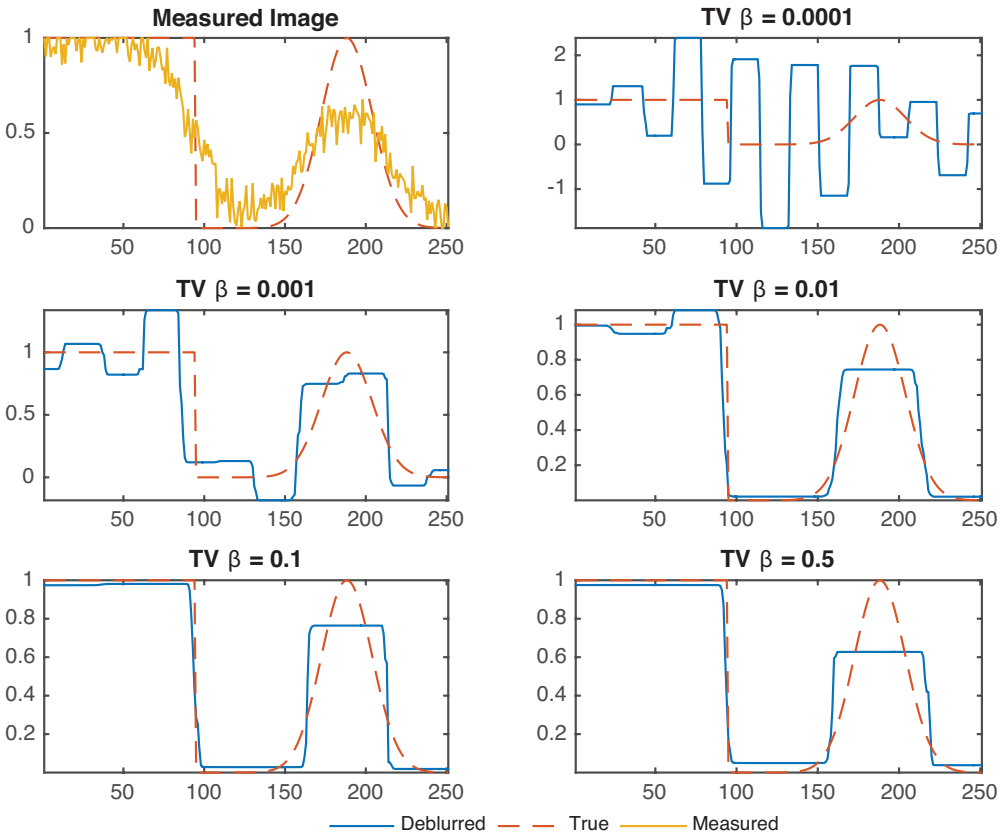

Figure 3.5: Deblurring with a Total Variation filter for different values of β.

The results are completely different from the ones obtained with the various Tikhonov filters. This time we see that the sharp edge can be recovered nicely while it is the smooth edge that is problematic. We see that the Total Variation method has the tendency to produce piecewise constant solutions. In the deblurred images there are still some small artifacts present. This can be explained by the fact that the Total Variation method requires the use of an iterative scheme to find solutions this time (for the pure Tikhonov filters we could compute the solutions directly using equation (3.12)). The iterative scheme needs stopping criteria as explained earlier. The small artifacts show that the method has not fully converged yet. By tightening the stopping criteria, the artifacts will be removed and the resulting deblurred images will indeed be piecewise constant.

It should be noted that in the above case we used the Lagged Diffusivity method. The results obtained with the Newton-Raphson method are similar.

Now we are going to try to combine the Tikhonov method and the Total Variation and hopefully get a 'best of both worlds' solution.

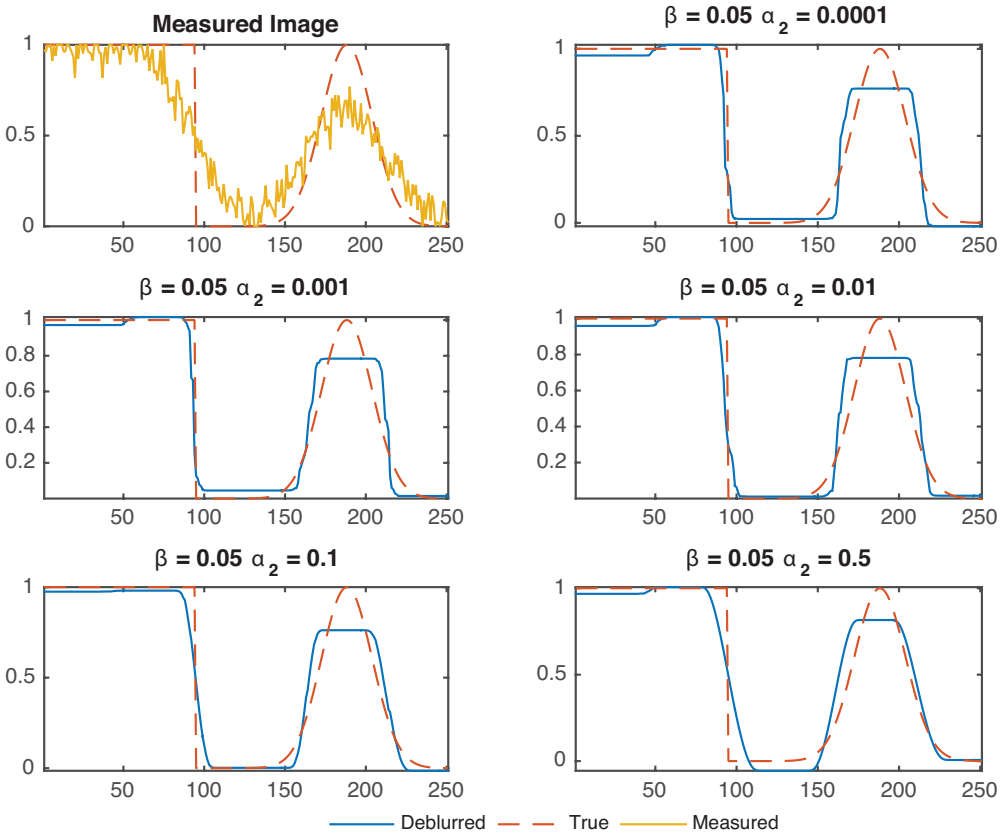

Figure 3.6: Deblurring with Tikhonov and Total Variation filters for different values of β and α_2.

We see, especially in the last two plots, that the resulting method produces images that are in between the two extremes of the pure Tikhonov and the pure Total Variation deblurring methods. The sharp edge is not as good as for the pure Total Variation deblurring, but the smooth edge is better. On the other hand, the smooth edge cannot be recovered as nicely as for a pure Tikhonov filter but the sharp edge is much better. By adjusting the weights β and α_2 (or more general, $\boldsymbol{\alpha}$) we can get more edgy images or more smooth images. It seems as if we can never recover *both* the sharp edge and the smooth edge perfectly using the above methods though. This is not too surprising: imagine that we do a Fourier decomposition of the image. The Tikhonov filters then surpress high frequency components from the image because that is where the noise can be found. But high frequency components are needed to produce sharp edegs. Hence 'the more Tikhonov, the less sharp edges'. On the other hand, the Total Variation filter tends to produce piecewise constant images, which is of course not how one would want to reproduce smooth edges. Hence 'the more Total Variation, the less smooth edges'. There will always be a tradeoff between these methods.

We note that, even though it is difficult to see in some of the plots, negative values may be attained. This is unwanted behaviour if one is trying to recover atomic fraction or concentration profiles. In the next series of plots we present results obtained under the constrained minimization algorithm mentioned in Section 3.4. The parameters used are the same as in 3.6. We see that the results are similar to the results from 3.6, except that the solutions are positive everywhere this time.

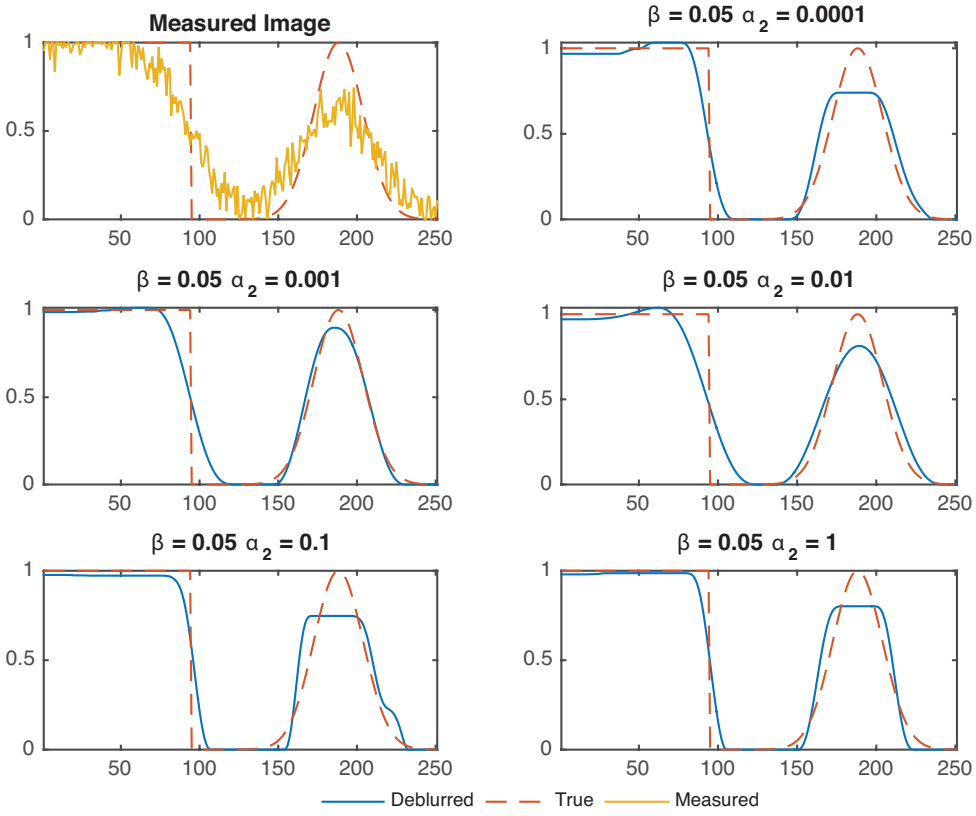

Figure 3.7: Deblurring under positivity constraints with Tikhonov and Total Variation filters for different values of β and α_2.

Finally, we mentioned that we also tried to deblur using so called Perona-Malik anisotropic diffusion [44]. While we were able to remove noise from images without introducing additional blur, we could not get the method to actually deblur the images. Hence, while the method in itself is interesting, it is not further discussed here and the results are not presented.

The above deblurring methods will be applied to real measurement data obtained by TNO/Solliance in Chapter 6.

4 Precursor Model and Numerical Implementation

In the introductory chapter we briefly mentioned the two-step process that is used to produce CIGS absorber layers at TNO/Solliance. Remember that in the first step of this process, layers of copper, indium and gallium are deposited on a soda-lime glass substrate coated with a layer of molybdenum. This deposition step is carried out either by sputtering or by electroplating. In the second step of the two-step process, the subtrate is loaded into a furnace for the selenization process.

Experiments performed at TNO/Solliance have shown that already during the storage of the substrate before selenization diffusion and chemical reactions take place. More specifically, it is observed that the layers of copper, indium and gallium diffuse into one another and reactions occur which lead to new intermetallic phases - even at room temperature! Because the precursors form the basis of what are to become CIGS solar cells, it is desired to have a model which describes the underlying physical and chemical processes. The goal of the model should be to describe the concentrations (in moles per unit volume) of the various components - including newly formed phases - throughout the precursor over time. The model is currently being developed at TNO/Solliance and in this chapter we will describe only a part this model. More specifically, we will focus on the physical and chemical processes that take place in the precursor stack during storage before the actual selenization step. We refer to this part of the model as the *precursor model*.

The reason for describing the precursor model in this work is that the current numerical implementation of the diffusion process is relatively slow. A faster method is desired. Since a numerical method is ultimately a way of solving equations, and the equations to be solved come from the model, we present the model first. Then we develop a numerical method to solve the equations.

As for the modelling part, we will start very general and along the way we will make observations and assumptions to taylor the model to the case at hand.

4.1 General Precursor Model

4.1.1 Continuity Equations

In the case at hand we are interested in the time evolution of the concentrations C_i of different components i inside a physical domain $\Omega \subset \mathbb{R}^3$. Each concentration C_i is a function of both space and time, i.e. $C = C(x,t)$, where $x = (x_1, x_2, x_3) \in \Omega$ and $t \in [0, \infty)$. Now think of a small control volume U contained in Ω. Assume this control volume to be stationary with respect to some *laboratory fixed reference frame*. The total amount of component i present in U is given by the integral $\int_U C(x,t)dx$. Now we stipulate that the amount of component i in U can only change due two processes: it either changes due to particles of component i leaving / entering U as a flux through the boundary ∂U, or, it changes due to sources / sinks of this component being present inside of U. That is,

$$\frac{d}{dt}\int_U C_i dx = -\int_{\partial U} \boldsymbol{F}_i \cdot \mathbf{n} dx + \int_U S_i dx,$$

where $\boldsymbol{F}_i$ and S_i are the flux and source functions respectively and $\boldsymbol{n}$ is the outward-pointing normal. Note that bold symbols refer to vector quantities. In general, the $\boldsymbol{F}_i$ and S_i depend on the concentration C_i, on x and on t and possibly on other factors as well. More on that later. For now, note that the surface integral can be rewritten to a volume integral using Gauss' divergence theorem. With this theorem the above equation becomes

$$\frac{d}{dt}\int_U C_i dx = -\int_U \mathrm{div}_x(\boldsymbol{F}_i) dx + \int_U S_i dx,$$

or, equivalently,

$$\int_V \left[\frac{\partial C_i}{\partial t} + \mathrm{div}_x(\boldsymbol{F}_i) - S_i \right] dx = 0.$$

This equation is generally referred to as the *continuity equation in integral form*. Because this equation is assumed to hold for all possible control volumes $U \subset \Omega$ and for all times (and assuming all functions under consideration to be smooth enough), it follows that the integrand must be identically zero. That is,

$$\frac{\partial C_i}{\partial t} = -\mathrm{div}_x(\boldsymbol{F}_i) + S_i \quad \text{in } \Omega \times [0, \infty).$$

This partial differential equation - the *continuity equation in differential form* - describes the dynamics of component i only but it can easily be generalized to multicomponent systems. To this end, define the vector $\boldsymbol{C} := [C_1, \ldots, C_n]$, where each C_i is a function for the concentration of component i and n is the number of different components in the system. Similarly, define vectors of functions $\boldsymbol{F} := [\boldsymbol{F}_1, \ldots, \boldsymbol{F}_n]$ and $\boldsymbol{S} := [S_1, \ldots, S_n]$. With this notation, the system of partial differential equations decsribing the dynamics of all the components can be written as

$$\frac{\partial \boldsymbol{C}}{\partial t} = -\mathrm{div}_x(\boldsymbol{F}) + \boldsymbol{S} \quad \text{in } \Omega \times [0, \infty).$$

Here the divergence operator is understood to be applied to each component of $\boldsymbol{F}$ separately. Note that this system of partial differential equations is very general and can describe the time evolution of basically any 'conserved' quantity. However, because it is so general, it is not saying much either! To better describe what is going on for our specific precursors, we have to make some choices.

Note the general trend that subscripts refer to the component under consideration and that bold symbols indicate the use of a vector. This trend will be continued throughout this chapter. Moreover, matrices will be denoted by bold symbols with bars.

4.1.2 Physical Domain of Interest

First of all, we have to define what our physical domain of interest is. The most obvious choice would be to say that the domain consists of the layers of molybdenum, copper, indium and gallium. The soda-lime glass substrate will not be included because it is inert and pretty much impermeable at the temperatures we will be dealing with - that is why glass was chosen as a substrate in the first place. Now let L, W and H be the length, width and height of the initial precursor stack together with the molybdenum (but without the soda-lime glass). As the components diffuse and react with each other, new phases form. Assuming constant partial molar volumes for the components the total volume of the system will not change, see Chapter 2.1.3. Then, at all times before the selenization step, our physical domain of interest will simply be the box $\Omega := [0, L] \times [0, W] \times [0, H]$. Typically, H is around $2\mu m$ while L and W are on the order of centimeters. It should be noted that initially the precursor is assumed to be uniform in the horizontal direction as depicted in Figure (4.1). Therefore, we only expect to see changes over in time in the vertical direction and as a result we will only be working with one spatial dimension (the vertical direction) in doing numerical simulations. To keep the model as general as possible we will keep on working with the domain Ω though.

During the selenization, the domain Ω changes because of the uptake of selenium but that is outside the scope of the present work.

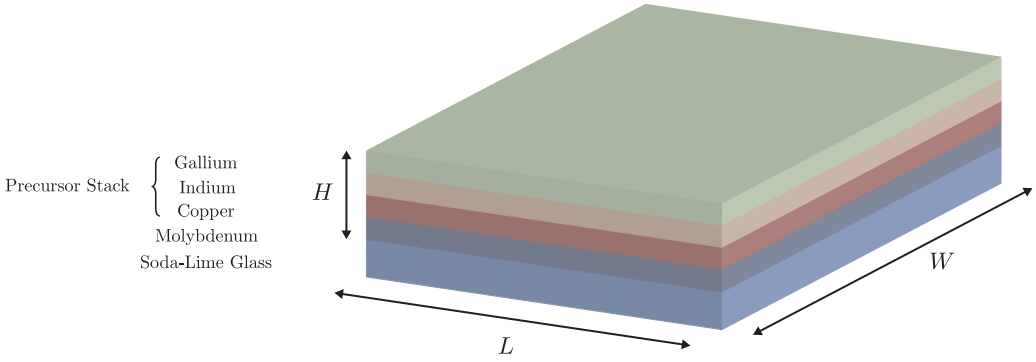

Figure 4.1: Graphical representation of the precursor in its initial state.

4.1.3 Components in the System

Next, we have to say what components we will be modelling. Obviously we have to include molybdenum, copper, indium, gallium and in the end $Cu(In_xGa_{1-x})Se_2$. However, X-ray diffraction measurements performed at TNO/Solliance and at other research institutes have shown that many intermediate binary and ternary phases form within the precursor. This proces already starts before the selenization step, when precursors are stored at room temperature. Among others, the phases $Cu_2Ga, Cu_9Ga_4, CuIn, Cu_1In_2$ and $Cu_{11}In_9$ have been observed to form within the precursor. Following upon the previous modelling work, each of these newly formed phases will be considered as separate components in the model. Unfortunately, at this moment it is not yet possible to say exactly which phases form during the two step process. In-Situ X-ray diffraction measurements should be able to us more about the reaction paths once it is operational. For now, we will work towards a very general model that can deal with an arbitrary amount of components. The details can be dealt with at a later time.

A subtle issue here is whether we should distinguish between different physical phases (i.e. liquid, solid, gas) as well. The first question in this respect is of course whether such different physical phases occur at all. Since initially the different layers of the precursor are all assumed to be in a solid state it would be surprising to see liquid or gas phases occur so this may not even be an issue at room temperature. But, as it turns out, gallium itself has a relatively low melting point and when it mixes with indium, the melting point decreases to a point where melting occurs even at room temperature.[11] It was decided at TNO/Solliance not model this liquidification. The reason being that it is of no particular interested whether gallium or indium are (partially) liquid: as soon as they react with other components to form new crystalline phases they become part of a solid crystal anyways. And it is the formation of the new crystalline phases that are of most interest. The possible liquid phase of We could try model the higher mobility of indium and gallium in their liquid phases by increasing their 'mobility' in the model, i.e. a higher diffusion coefficient.

Furthermore it should be noted that we chose not to include the vacancies within the precursor as a separate component. Like in Chapter 2, we assume the concentration of vacancies to be negligible compared with the concentrations of other components.

As stated before we will assume the partial molar volumes V_i of all the components to be constant. For *atomic* components we will assume V_i to be equal to the molar volume of the pure component i. For intermetallic phases, we will take linear combinations of the partial molar volumes of the constituents. For example, for Cu_9Ga_4 we will say that $V_{Cu_9Ga_4} = 9V_{Cu} + 4V_{Ga}$.

[11] A video demonstrating this behaviour can be found at https://www.youtube.com/watch?v=4-ZDDkamfAc

4.1.4 Diffusion Fluxes

Transport Equations and Interdiffusion Coefficients

Next up is the vector of fluxes $\boldsymbol{F}$. A lot has already been said about fluxes in Chapter 2. We repeat the most important conclusions here. Given a choice of sections to measure fluxes, the diffusion fluxes for the different components within the precursor can be expressed as

$$\boldsymbol{F}_i = \sum_j L_{ij} X_j,$$

with X_j the different driving forces and L_{ij} the corresponding transport coefficients. In general the transport coefficients are rank two tensors. For the precursor model, we assume temperature and pressure to be constant. We also assume the precursor to be electrically neutral. Only the chemical potential gradients $\nabla\mu_i$ of the different components are assumed to be responsible for driving the fluxes. The above transport equations reduce to the following equations for the diffusion fluxes:

$$\boldsymbol{F}_i = \sum_{j=1}^{n} L_{ij} \nabla_x \mu_j.$$

In Chapter 2 we discussed in detail the importance of specifying a reference frame relative to which fluxes are measured. For our model, we will be working with laboratory-fixed sections. Assuming constant partial molar volumes for the components, this choice of sections coincides with volume-fixed sections. Hence we can refer to the fluxes as *interdiffusion fluxes* and we may denote them with tildes. Since we will only be working with one frame of reference in this chapter - a laboratory fixed frame of reference - there is no need to use this specific notation though.

By rewriting the chemical potential gradients in terms of concentration gradients the interdiffusion fluxes could also be expressed as

$$\tilde{\boldsymbol{F}}_i = \sum_{j=1}^{n} -\tilde{D}_{ij} \nabla_x C_j.$$

Here the D_{ij} are the *interdiffusion coefficients*. Like the transport coefficients, they are in general rank two tensors. However, if we assume the precursor to be isotropic then the interdiffusion coefficients will be scalar quantities.

In Chapter 2 we saw that for an n-component system there are in fact only $n-1$ independent interdiffusion fluxes and $(n-1)^2$ interdiffusion coefficients are needed to describe these fluxes. In general these interdiffusion coefficients depend on the concentrations of the different components. We worked on methods to derive (approximations to) the interdiffusion coefficients for the one-dimensional based on experimental measurements with diffusion couples.[12]

Current Working Assumptions for Diffusion Coefficients

At the moment there is not enough data to establish (estimates for) all the relevant interdiffusion coefficients. And even if there is data - it may be difficult to deduce the interdiffusion coefficients

[12]It should be noted here that in Chaper 2, we only considered different *atomic* components. Newly formed intermetallic phases were *not* considered as separate components. This raises the question how to interpret and determine $\tilde{D}_{ij}$ in case i or j refers to an intermetallic phase. A possible solution is to set $\tilde{D}_{ij} = 0$ whenever i refers to an intermetallic phase. In other words, one could render the intermetallic phases immobile. In order for an intermetallic phase to 'diffuse', it would have to decompose (through its reaction term S_i) into its atomic constituents which can then diffuse separately through the precursor. In a similar fashion one could set $\tilde{D}_{ij} = 0$ whenever j refers an intermetallic phase, meaning that atomic components only diffuse under the influence of concentration gradients of atomic components and not under the influence of concentration gradients of intermetallic phases.

because of the difficulties encountered in Chapter 2 with multicomponent diffusion (see also the discussion in Chapter 6. To (partially) overcome these difficulties a few working assumptions have been made at TNO/Solliance. First of all, components are only assumed to diffuse under the influence of their own concentration gradients, i.e.

$$\boldsymbol{F}_i = -\tilde{D}_{ii} \nabla_x C_i.$$

In other words, Fick's law is followed and the so called cross-diffusion terms are ignored by setting $\tilde{D}_{ij} = 0$ whenever $i \neq j$.

Secondly, assumptions have been made on the concentration dependence of the interdiffusion coefficients $\tilde{D}_{ii}$. Say we are following a few particles of component i as they diffuse through an environment consisting entirely of particles of component j. Let $\mathfrak{D}_{ij}$ be diffusion coefficient associated with this process (note that $\tilde{D}_{ij}$ and D_{ij} have a completely different meaning. From now on, just forget about the interdiffusion coefficients $\tilde{D}_{ij}$). If $i = j$, then this coefficient is commonly referred to in literature as the *self or tracer diffusion coefficient*. Otherwise, if $i \neq j$, it is referred to as *impurity diffusion coefficient [37]*. The tracer and impurity diffusion coefficients $\mathfrak{D}_{ij}$ are related to (the solid-state equivalent of) infinite dilutely diffusion processes in which the environment of the diffusing particles is chemically homogenous. But what happens to the diffusion coefficient $\tilde{D}_{ii}$ for a particle in an environment that is not infinitely dilute? As a simple approximation we could take a weighted average of the different tracer and impurity diffusion coefficients. More specifically, let N_j denote the molar fraction of component j. In Chapter 2 we have seen that the molar fractions are related to the concentrations by

$$N_j = C_j V_{\text{mol}},$$

where the molar volume V_{mol} can be expressed as

$$V_{\text{mol}} = \sum_{i=1}^{n} N_i V_i = \frac{1}{\sum_{i=1}^{n} C_i}.$$

We then make the assumptions that the diffusion coefficient $\tilde{D}_{ii}$ for component i can be expressed as

$$\tilde{D}_{ii}(\boldsymbol{C}(x,t)) \quad = \quad \sum_{j=1}^{n} N_j(x,t) \mathfrak{D}_{ij} = \sum_{j=1}^{n} \left(\frac{C_i(x,t)}{\sum_{k=1}^{n} C_k(x,t)} \right) \mathfrak{D}_{ij}.$$

The advantage of this approach is that some of the tracer and impurity diffusion coefficients can be found in literature. On the downside it is not clear how the resulting diffusion coefficients $\tilde{D}_{ii}$ relate to interdiffusion coefficients that can be derived from measurements. And of course the possible effects of cross-diffusion are ignored.

Temperature Dependence of Diffusion Coefficients

In the precursor model we work with a constant temperature θ. However, if at a later time the precursor is selenized, temperature of course starts playing a role. As already mentioned in Chapter 2 diffusion coefficients are assumed to follow the general Arrhenius formula. More specifically, we will say that the D_{ij} have a temperature dependence that can be expressed as

$$\mathfrak{D}_{ij}(\theta) = \mathring{\mathfrak{D}}_{ij} \exp\left(-\frac{Q_{ij}}{R\theta} \right).$$

Here $\mathring{\mathfrak{D}}_{ij}$ is the pre-exponential factor, Q_{ij} the activation energy and R is the universal gas constant. With this assumption we can write

$$\tilde{D}_{ii}(\boldsymbol{C}(x,t),\theta) = \sum_{j=1}^{n} \left(\frac{C_i(x,t)}{\sum_{k=1}^{n} C_k(x,t)} \right) \mathring{\mathfrak{D}}_{ij} \exp\left(-\frac{Q_{ij}}{R\theta} \right).$$

To conclude the above discussion, we introduce a diffusion matrix $\bar{\boldsymbol{D}} := \left[\tilde{D}_{ij} \right]_{i,j=1}^{n}$, with

$$\tilde{D}_{ii}(\boldsymbol{C}(x,t),\theta) = \sum_{j=1}^{n} \left(\frac{C_i(x,t)}{\sum_{k=1}^{n} C_k(x,t)} \right) \mathring{\mathfrak{D}}_{ij} \exp\left(-\frac{Q_{ij}}{R\theta} \right)$$

and

$$\tilde{D}_{ij} = 0 \quad \text{whenever } i \neq j.$$

Then the flux vector $\boldsymbol{F}$ can be written as

$$\boldsymbol{F} = -\bar{\boldsymbol{D}} \nabla_x \boldsymbol{C},$$

where the gradient operator ∇_x is understood to be applied component-wise to the vector of concentrations $\boldsymbol{C}$.

4.1.5 Chemical Reactions

The only possible sources or sinks for the components in the precursor during the first step of the process will be chemical reactions. It is only during the second step that a 'real' selenium source is added but let's forget about that for now. Chemical reactions occur when different atoms meet under the right conditions where they form new bonds and thus new chemical components. X-ray Diffraction measurements performed at TNO/Solliance and other research institutes revealed that, among others, the following chemical reactions may occur at room temperature.

$$\begin{aligned} 9\mathrm{Cu} + 4\mathrm{Ga} &\rightharpoonup \mathrm{Cu_9Ga_4}, \\ \mathrm{Cu} + \mathrm{In} &\rightharpoonup \mathrm{CuIn}, \\ \mathrm{Cu} + 2\mathrm{In} &\rightharpoonup \mathrm{CuIn_2}, \\ 2\mathrm{Cu} + \mathrm{In} &\rightharpoonup \mathrm{Cu_2In}. \end{aligned}$$

Note that all these equations are of the form

$$aA + bB \rightharpoonup cC,$$

where A, B and C are certain components and a, b and c stoichiometric coefficients. The reaction equations tell us nothing about the *rate* at which the reaction occurs though. The only thing can be said is that the rate at which c moles of C are produced should be equal to the rate at which a moles of A are removed. Similarly, the rate at which c moles of C are produced should be equal to the rate at which b moles of B are removed. If we let $[X]$ denote the concentration of component X for a moment, then what we just said can be summarized as:

$$\text{Reaction rate} = \frac{1}{c}\frac{d[C]}{dt} = -\frac{1}{a}\frac{d[A]}{dt} = -\frac{1}{b}\frac{d[B]}{dt}.$$

Now let f be the function which describes the reaction rate. We will assume that f depends on the concentrations $[A], [B]$ and $[C]$ as well as on the temperature θ. Usually f is assumed to take the form

$$f([A],[B],[C],\theta) = \mathring{f} \exp\left(-\frac{E_f}{R\theta} \right) [A]^{\alpha}[B]^{\beta}[C]^{\gamma}.$$

Here $\mathring{f}$ is the temperature-independent pre-exponential factor, E_f the activation energy for the reaction and R is the universal gas constant again. In other words, the reaction rate is assumed to follow the same Arrhenius-type behaviour for temperature as the diffusion coefficients do. The sum $\alpha + \beta + \gamma$ is called the order of the reaction and should be determined experimentally. No matter what the particular form of f is, it does give rise to the following system of time evolution equations:

$$\begin{aligned} \frac{d[A]}{dt} &= -af([A],[B],[C],\theta), \\ \frac{d[B]}{dt} &= -bf([A],[B],[C],\theta), \\ \frac{d[C]}{dt} &= cf([A],[B],[C]). \end{aligned}$$

Experimentally determining the orders of the reactions, the pre-exponential factors and the activation energies for the different reactions may not always be possible. The reason is that the growth of new intermetallic phases in solids is usually *diffusion limited.* This can be explained as follows. Suppose we have a diffusion couple that consists of a beam of pure component A and a beam of pure component B. As the beams are put together, a new phase A_xB_{1-x} grows as a layer in the interdiffusion zone. To keep on growing, atoms of type A need to diffuse through the layer to meet with atoms of type B and react, or, alternatively, atoms of type B must diffuse in the other direction through the layer and meet with atoms of type A. The thicker the layer grows, the longer it takes for the atoms to diffuse through the layer. If the diffusion rates are slow compared to the reaction rates than the reaction rates may not at all be oservable: the rate at which the layer grows is limited by diffusion and hence only reveals information about the diffusion rates. But if we are dealing with diffusion limited growth, it may not be necessary to have an accurate reaction rate in the model. It may simply be enough to say that reactions rates are several orders of magnitudes larger than the diffusion coefficients.

It should also be mentioned that most reactions only occur within a certain temperature range. This could be incorporated into the model by multiplying each reaction rate by a factor that is equal to one for temperatures within the correct temperature range and zero otherwise.

Furthermore it should be noted that it is possible for two or more new phases to grow at the same position. But possibly some phases grow in layers, preventing different phases to be present at the same position in space. This behaviour could perhaps be incorporated in the reaction rates by setting them to zero if some other phase is already present.

In the end, deciding which reactions to include and how to model them is part of the ongoing research at TNO/Solliance and is not the aim of this work. For the numerical method to be developed (and for the subsequent mathematical analysis of the model in the final thesis for the TU/e) the particular choices made in this respect are not really important. We will just say that the reactions for component i are all covered by the general source term $S_i = S_i(\boldsymbol{C}, T)$.

4.1.6 Boundary and Initial Conditions

To complete the model we need boundary conditions and initial conditions. For the initial conditions, it is assumed at TNO/Solliance that assembly of the precursor results in uniform layers of molybdenum, copper, indium and gallium (in that order). This was already illustrated in Figure (4.1). Initially there are no intermetallic compounds present. As for the boundary conditions, it is assumed that no particles can leave the system. In other words, there is *zero*

flux across the boundaries. For each component $1 \leq i \leq n$ this condition can be expressed as

$$\boldsymbol{F}_i(x,t) \cdot \boldsymbol{n} = -\left(\sum_{j=1}^{n} \tilde{D}_{ij}(\boldsymbol{C}(x,t),\theta) \nabla_x C_j(x,t) \right) \cdot \boldsymbol{n} = 0, \quad x \in \partial\Omega.$$

Here $\partial\Omega$ is the boundary of Ω and $\boldsymbol{n}$ is the outward pointing normal to the boundary. For the system as a whole, we will formulate these conditions as

$$\boldsymbol{F}(x,t) \cdot \boldsymbol{n} = -\left(\bar{\boldsymbol{D}}(\boldsymbol{C}(x,t),\theta) \nabla_x \boldsymbol{C}(x,t) \right) \cdot \boldsymbol{n} = 0, \quad x \in \partial\Omega.$$

Here the inner product with the outer normal $\boldsymbol{n}$ is understood to be applied component-wise to the vector of fluxes.

4.1.7 The Precursor Model

Putting all of the above together we arrive at the following model.

Required Data:

1. A choice of n components (Molybdenum, Copper, Indium, Gallium and intermetallic compounds that form during the diffusion process), whose (unknown!) concentrations will be denoted by $C_i = C_i(x,t)$ are stored in the vector of functions $\boldsymbol{C} = [C_1, \ldots, C_n]$;
2. Chemical reactions functions $S_i = S_i(\boldsymbol{C}(x,t),\theta)$ for each $i \in \{1,\ldots,n\}$, collected in the vector of functions $\boldsymbol{S} = [S_1, \ldots, S_n]$;
3. Interdiffusion coefficient functions $\tilde{D}_{ij} = \tilde{D}_{ij}(\boldsymbol{C}(x,t),\theta)$ for each $i \in \{1,\ldots,n\}$, collected in a matrix of functions $\bar{\boldsymbol{D}} = \left[\tilde{D}_{ij}\right]_{i,j=1}^{n}$;
4. Initial concentrations $C_i^{\text{Initial}} = C_i^{\text{Initial}}(x)$ for each $i \in \{1,\ldots,n\}$, collected in the vector of functions $\boldsymbol{C}^{\text{Initial}} = \left[C_1^{\text{Initial}}, \ldots, C_n^{\text{Initial}}\right]$.

System of Equations:

Given the data, time evolution of the n concentrations stored in $\boldsymbol{C}$ is governed by the following system of non-linear, coupled partial differential equations with no-flux boundary conditions:

$$\begin{cases} \partial_t \boldsymbol{C}(x,t) = \operatorname{div}_x \left(\bar{\mathbf{D}}(\boldsymbol{C}(x,t),\theta) \nabla_x \boldsymbol{C}(x,t) \right) + \mathbf{S}(\boldsymbol{C}(x,t),\theta) & (x,t) \in \Omega \times [0,\infty), \\ \left(\bar{\boldsymbol{D}}(\boldsymbol{C}(x,t),\theta) \nabla_x \boldsymbol{C}(x,t) \right) \cdot \boldsymbol{n} = 0 & (x,t) \in \partial\Omega \times [0,\infty), \\ \boldsymbol{C}(x,0) = \boldsymbol{C}_{\text{Initial}}(x) & x \in \Omega. \end{cases} \tag{4.1}$$

Like before, $\nabla_x, \operatorname{div}_x$ and the inner product with $\boldsymbol{n}$ are understood to be applied component-wise.

Now that we have written down a system of equations we would like to know how to solve them for $\boldsymbol{C}$. Analytically solving a system of coupled, non-linear partial differential equations is out of the question though. Numerical methods have been developed at TNO/Solliance in this respect and they will be discussed in the next section.

From a more theoretical point of view we would like to know whether it can be shown upfront that the above system of equations admits a vector of solutions $\boldsymbol{C}(x,t)$ on $\Omega \times [0,\infty)$ in some appropriate setting. If that's the case we would like to know whether the solution is unique, positive and depends smoothly on the data.

4.2 Numerical Implementation

4.2.1 Current Numerical Method

The model (4.1) is formulated in a three-dimensional setting. Because the initial sample is assumed to have a layered structure as depicted in figure (4.1), we only expect to see changes along the vertical direction. Therefore, in doing calculations, we will restrict ourselves to a one-dimensional case. The variable x will be used to denote the depth within the layer.

The current numerical method used at TNO/Solliance [35] to the diffusion-reaction equations is based on the following. Suppose we have a concentration of some species in an inifite, one-dimensional domain. The species is only allowed to diffuse, no reactions occur. The diffusion coefficient D is assumed to be constant. As for the initial condition, we suppose a point source is present at the some point $y \in \mathbb{R}$. This is modelled by the Dirac-delta function δ_y. If C denotes the concentration then the above assumptions lead to the following set of equations for the time evolution of the species:

$$\begin{cases} \partial_t C(x,t) = D\Delta C(x,t), & (x,t) \in \mathbb{R}, \\ C(x,0) = \delta_y(x), & t \in [0,\infty). \end{cases}$$

This equation can be solved exactly [37] with the solution being

$$C(x,t) = \frac{1}{\sqrt{2\pi Dt}} \exp\left(-\frac{(x-y)^2}{4Dt}\right).$$

The solution can be interpreted as a Gaussian curve that spreads out over time. In reality, the domain is not infinite dimensional though. To deal with this numerically, the Gaussian curve is reflected inwards at the physical boundaries. When the reflections meet the physical boundaries at the other side, they are reflected inwards again. And so on. To deal with general initial conditions, note that in principle any initial condition can be written as a superposition of Dirac-delta functions. Because the diffusion coefficient is constant, the diffusion equation is linear and hence the solution C in case of a general initial condition can be written as the superposition of solutions corresponding to different Dirac-delta functions. After computing diffusion using this 'superpositions-and-reflections method', the reactions are computed separately. Then diffusion again, rections, etcetera.

There are a few downsides to this method though. First of all, the method is built around the assumption that the diffusion coefficients are constant. This may not be the case. But then the exact solution is not really an exact solution to begin with and it is unclear how 'superpositions-and-reflections' solutions compare with 'real' solutions. Furthermore, doing all the superpositions and all the reflections at the boundaries is computationally expensive. It is the bottleneck in the current numerical method used for simulating the CIGS formation process.

Below we will work out a numerical method to treat diffusion - and also reactions - using a computationally effecient method that allows us to work with non-constant diffusion coefficients as well. Care has been taken to allow easy incorporaton of new components, intermetallic phases or reactions in the method because it is not yet known which components and reactions to include.

The proposed numerical method can be summarized in two steps:

1. Discretize the partial differential equations in space first (Method of Lines) using a finite volume discretization;
2. Discretize the remaining systems of ordinary differential equations in time.

Both of the steps will be worked out below.

4.2.2 Finite Volume Discretization

Equations for Volume Averages

For the first step, note that the time evolution equation for a single component i in the one-dimensional case reads as

$$\partial_t C_i(x,t) = \partial_x \left(\sum_{j=1}^{n} \tilde{D}_{ij}(\boldsymbol{C}(x,t),\theta)\partial_x C_j(x,t) \right) + S_i(\boldsymbol{C}(x,t),\theta).$$

As one of the working assumptions we said that $\tilde{D}_{ij} = 0$ whenever $i \neq j$. If we use the notation D_i for $\tilde{D}_{ii}$, then the above equation reduces to

$$\partial_t C_i(x,t) = \partial_x \left(D_i(\boldsymbol{C}(x,t),T)\partial_x C_i(x,t) \right) + S_i(\boldsymbol{C}(x,t),T). \tag{4.2}$$

Now we apply the *method of lines* to this partial differential equation. That is, we first discrete the spatial derivatives. This reduces the *partial* differential equation to a *system* of *ordinary* differential equations that will be solved by a particular choice of time integration method. But more on time integration later. For now, note that in the one-dimensional case our physical domain of interest is the interval $[0, H]$. For the spatial discretization of our equations we will employ the so called *finite volume method*. For this method we partition the interval $[0, H]$ into N smaller intervals that we refer to as *finite volumes*. Let Ω^j denote the j-th finite volume, let x_j denote its center and denote the left and right boundaries of the volume by $x_{j-1/2}$ and $x_{j+1/2}$ respectively. With this notation, we see that $\Omega^j = [x_{j-1/2}, x_{j+1/2}]$. The length of Ω^j will be denoted by Δx_j. Because the current numerical method used at TNO/Solliance involves adaptive meshing procedures we do not make the assumption that all Δx_j are equal.

To proceed, we integrate equation (4.2) over a finite volume Ω^j. This yields:

$$\begin{aligned} \int_{\Omega^j} \partial_t C_i(x,t)dx &= \int_{\Omega^j} \partial_x \left(D_i(\boldsymbol{C}(x,t),\theta)\partial_x C_i(x,t) \right) + \int_{\Omega^j} S_i(\boldsymbol{C}(x,t),\theta)dx \\ &= \left[D_i(\boldsymbol{C}(x,t),T)\partial_x C_i(x,t) \right]_{x_{j-1/2}}^{x_{j+1/2}} + \int_{\Omega^j} S_i(\boldsymbol{C}(x,t),\theta)dx. \end{aligned} \tag{4.3}$$

Now define

$$F_i^{j+1/2}(t,\theta) := -D_i(\boldsymbol{C}(x_{j+1/2},t),\theta)\partial_x C_i(x_{j+1/2},t). \tag{4.4}$$

Note that $F_i^{j+1/2}$ represents the flux of component i across the boundary between Ω^j and Ω^{j+1}. With this notation, equation (4.3) can be expressed as

$$\int_{\Omega^j} \partial_t C_i(x,t)dx = -\left[F_i^{j+1/2}(t,\theta) - F_i^{j-1/2}(t,\theta) \right] + \int_{\Omega^j} S_i(\boldsymbol{C}(x,t),\theta)dx. \tag{4.5}$$

Next, define $C_i^{\mathrm{Avg},j}(t)$ to be the average of C_i over volume Ω^j at time t :

$$C_i^{\mathrm{Avg},j}(t) := \frac{1}{\Delta x_j} \int_{\Omega^j} C_i(x,t)dx.$$

In a similar fashion, we define volume average reactions as

$$S_i^{\mathrm{Avg},j}(t,\theta) := \frac{1}{\Delta x_j} \int_{\Omega^j} S_i(\boldsymbol{C}(x,t),\theta)dx.$$

Then the integrated evolution equation (4.5) can be reformulated as

$$\frac{d}{dt} C_i^{\mathrm{Avg},j}(t) = -\frac{F_i^{j+1/2}(t,\theta) - F_i^{j-1/2}(t,\theta)}{\Delta x_j} + S_i^{\mathrm{Avg},j}(t,\theta). \tag{4.6}$$

Note that this equation is exact.

Numerical Average Concentrations

In doing simulations it is the quantities $C_i^{\text{Avg},j}$ that we will be trying to compute. To this end, let $C_i^{\text{Num},j}(t)$ denote the (unknown) *numerical approximation* to $C_i^{\text{Avg},j}(t)$. For convenience, we also define the vectors $\boldsymbol{C}^{\text{Num},j}(t)$ and $\boldsymbol{C}_i^{\text{Num}}(t)$ and a matrix $\bar{\boldsymbol{C}}^{\text{Num}}(t)$ as

$$\begin{aligned}
\boldsymbol{C}^{\text{Num},j}(t) &:= \left[C_1^{\text{Num},j}(t), \ldots, C_n^{\text{Num},j}(t)\right] \in \mathbb{R}^n, \\
\boldsymbol{C}_i^{\text{Num}}(t) &:= \left[C_i^{\text{Num},1}(t), \ldots, C_i^{\text{Num},N}(t)\right]^T \in \mathbb{R}^N, \\
\bar{\boldsymbol{C}}^{\text{Num}}(t) &:= \left[\boldsymbol{C}_1^{\text{Num}}(t), \ldots, \boldsymbol{C}_n^{\text{Num}}(t)\right] = \left[\boldsymbol{C}^{\text{Num},1}(t), \ldots, \boldsymbol{C}^{\text{Num},N}(t)\right]^T \in \mathbb{R}^{n\times N}.
\end{aligned}$$

Numerical Source Terms

Now let's have a look at the source terms $S_i^{\text{Avg},j}$. Because the $S_i^{\text{Avg},j}$ depend on the exact concentrations - which we do not know - we need to approximate the source terms in terms of the numerical solution values contained in $\bar{\boldsymbol{C}}^{\text{Num}}$. To this end, let $S_i^{\text{Num},j}(t,\theta)$ denote the numerical approximation to $S_i^{\text{Avg},j}(t,\theta)$. The most straightforward choice would be to define $S_i^{\text{Num},j}(t,\theta)$ as

$$S_i^{\text{Num},j}(t,\theta) := S_i(\boldsymbol{C}^{\text{Num},j}(t),\theta), \tag{4.7}$$

simply because

$$S_i(\boldsymbol{C}^{\text{Num},j}(t),\theta) = \frac{1}{\Delta x_j}\int_{\Omega^j} S_i(\boldsymbol{C}^{\text{Num},j}(t),\theta)dx \approx \frac{1}{\Delta x_j}\int_{\Omega^j} S_i(\boldsymbol{C}(x,t),\theta)dx = S_i^{\text{Avg},j}(t,\theta).$$

Numerical Fluxes

We see from expression (4.4) that the flux term $F_i^{j+1/2}$ is (minus) the product of the diffusion coefficient and the concentration gradient at the boundary between volumes Ω^j and Ω^{j+1}. Of course we don't know these quantities exactly so we need to introduce approximations. In this respect, we make the assumption that the diffusion coefficients are constant in each respective finite volume. More specifically, we define numerical diffusion coefficients $D_i^{\text{Num},j}$ in the finite volume Ω^j as

$$D_i^{\text{Num},j}(t,\theta) := D_i(\boldsymbol{C}^{\text{Num},j}(t),\theta). \tag{4.8}$$

To obtain an approximation for the diffusion coefficient at the boundary between Ω^j and Ω^{j+1}, one may be tempted to simply take the *arithmetic mean* $\left(D_i^{\text{Num},j+1} + D_i^{\text{Num},j}\right)/2$ of the diffusion coefficients $D_i^{\text{Num},j}$ and $D_i^{\text{Num},j+1}$ in Ω^j and Ω^{j+1} respectively. A better idea would be to take the harmonic mean though. To see this, suppose we introduce an artificial grid point $x_{j+1/2}$ at the boundary between Ω^j and Ω^{j+1}. Doing a simple finite difference approximation to the derivative at $x_{j+1/2}$ - and surpressing the time and temperature dependence from the notations for a moment - we could approximate the flux $F_i^{j+1/2}$ as

$$F_i^{j+1/2} \approx -D_i^{\text{Num},j}\frac{C_i^{\text{Num},j+1/2} - C_i^{\text{Num},j}}{\Delta x_j/2}. \tag{4.9}$$

Note that we only use information from the volume Ω^j in this approximation. In a similar spirit we could approximate the flux $F_i^{j+1/2}$ using information from Ω^{j+1} as

$$F_i^{j+1/2} \approx -D_i^{\text{Num},j+1}\frac{C_i^{\text{Num},j+1} - C_i^{\text{Num},j+1/2}}{\Delta x_{j+1}/2}. \tag{4.10}$$

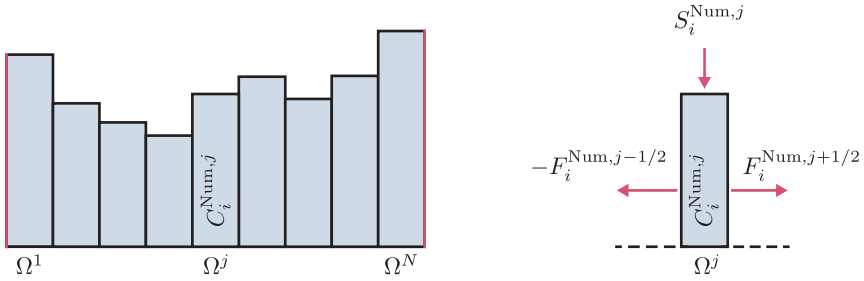

Figure 4.2: Graphical representation of the finite volume method. On the left we see the domain divided into N finite volumes. The heights of the bars represent the average concentrations within these volumes. On the right we see that the average concentration $C^i_{\mathrm{Avg},j}$ over the volume Ω^j can only change due to material flowing through the boundaries of Ω^j and due to sources within Ω^j itself.

Now, of course, we want the two approximations for the flux to be equal. Equating (4.9) and (4.10) and solving for $C_i^{\mathrm{Num},j+1/2}$ yields

$$C_i^{\mathrm{Num},j+1/2} = \frac{\Delta x_j D_i^{\mathrm{Num},j+1} C_i^{\mathrm{Num},j+1} + \Delta x_{j+1} D_i^{\mathrm{Num},j} C_i^{\mathrm{Num},j}}{\left(\Delta x_j D_i^{\mathrm{Num},j+1} + \Delta x_{j+1} D_i^{\mathrm{Num},j}\right)}.$$

The reason for solving for $C_i^{\mathrm{Num},j+1/2}$ is that we do not really want to introduce additional grid points. Substituting the above expression back into either (4.9) or (4.10) and doing some basic algebra shows us that (with time and temperature dependence back in the notation)

$$F_i^{j+1/2} \approx -\frac{2D_i^{\mathrm{Num},j} D_i^{\mathrm{Num},j+1}}{\Delta x_j D_i^{\mathrm{Num},j+1} + \Delta x_{j+1} D_i^{\mathrm{Num},j}} \left(C_i^{\mathrm{Num},j+1} - C_i^{\mathrm{Num},j}\right).$$

This motivates us to define $F_i^{\mathrm{Num},j+1/2}$ and $F_i^{\mathrm{Num},j-1/2}$ as

$$F_i^{\mathrm{Num},j+1/2}(t,\theta) := -\frac{2D_i^{\mathrm{Num},j}(t,\theta) D_i^{\mathrm{Num},j+1}(t,\theta) \left(C_i^{\mathrm{Num},j+1}(t) - C_i^{\mathrm{Num},j}(t)\right)}{\Delta x_j D_i^{\mathrm{Num},j+1}(t,\theta) + \Delta x_{j+1} D_i^{\mathrm{Num},j}(t,\theta)}, \quad (4.11)$$

$$F_i^{\mathrm{Num},j-1/2}(t,\theta) := -\frac{2D_i^{\mathrm{Num},j-1}(t,\theta) D_i^{\mathrm{Num},j}(t,\theta) \left(C_i^{\mathrm{Num},j}(t) - C_i^{\mathrm{Num},j-1}(t)\right)}{\Delta x_{j-1} D_i^{\mathrm{Num},j}(t,\theta) + \Delta x_j D_i^{\mathrm{Num},j-1}(t,\theta)}. \quad (4.12)$$

Note that if $\Delta x_j = \Delta x$ for all $j = 1, \ldots, N$ then

$$F_i^{\mathrm{Num},j+1/2} = -\left(\frac{2D_i^{\mathrm{Num},j} D_i^{\mathrm{Num},j+1}}{D_i^{\mathrm{Num}j+1} + D_i^{\mathrm{Num},j}}\right) \frac{C_i^{\mathrm{Num},j+1} - C_i^{\mathrm{Num},j}}{\Delta x},$$

and we recognize the diffusion coefficient in this expression as being the *harmonic mean* (and hence not the arithmetic mean) of $D_i^{\mathrm{Num},j}$ and $D_i^{\mathrm{Num},j+1}$ and not that arithmetic mean.

At the boundaries ($j = 1, N$), we cannot use the above definitions without invoking artificial volumes. But there is no need to: the no-flux conditions at the boundary of the domain in the model simply tell us to set

$$F_i^{\mathbf{Num},1/2}(t,\theta) = F_i^{\mathbf{Num},N+1/2}(t,\theta) = 0. \quad (4.13)$$

It should be noted that, while we are working with the assumption that the components only diffuse under the influence of their own concentration gradient (Fick's law), the method can easily be extended to include cross-diffusion terms as well.

Semi-Discrete System of Equations

Replacing all the quantities in equation (4.6) by their numerical counterparts (4.7), (4.11) and (4.12) we obtain the semi-discrete system

$$\begin{aligned}
\frac{dC_i^{\mathrm{Num},j}}{dt}(t) &= -\frac{F_i^{\mathrm{Num},j+1/2}(t,\theta) - F_i^{\mathrm{Num},j-1/2}(t,\theta)}{\Delta x_j} + S_i^{\mathrm{Num},j}(t,\theta) \qquad (4.14)\\
&= \frac{1}{\Delta x_j}\left[\frac{2D_i^{\mathrm{Num},j}(t,\theta)D_i^{\mathrm{Num},j+1}(t,\theta)}{\Delta x_j D_i^{\mathrm{Num}j+1}(t,\theta) + \Delta x_{j+1} D_i^{\mathrm{Num},j}(t,\theta)}\right]\left[C_i^{\mathrm{Num},j+1}(t) - C_i^{\mathrm{Num},j}(t)\right]\\
&- \frac{1}{\Delta x_j}\left[\frac{2D_i^{\mathrm{Num},j-1}(t,\theta)D_i^{\mathrm{Num},j}(t,\theta)}{\Delta x_{j-1} D_i^{\mathrm{Num},j}(t,\theta) + \Delta x_j D_i^{\mathrm{Num},j-1}(t,\theta)}\right]\left[C_i^{\mathrm{Num},j}(t) - C_i^{\mathrm{Num},j-1}(t)\right]\\
&+ S_i^{\mathrm{Num},j}(t,\theta),
\end{aligned}$$

for $j = 2, \ldots, N-1$, and at the boundaries

$$\begin{aligned}
\frac{dC_i^{\mathrm{Num},1}}{dt}(t) &= \frac{1}{\Delta x_1}\left[\frac{2D_i^{\mathrm{Num},1}(t,\theta)D_i^{\mathrm{Num},2}(t,\theta)}{\Delta x_j D_i^{\mathrm{Num}2}(t,\theta) + \Delta x_{j+1} D_i^{\mathrm{Num},1}(t,\theta)}\right]\left[C_i^{\mathrm{Num},2}(t) - C_i^{\mathrm{Num},1}(t)\right]\\
&+S_i^{\mathrm{Num},1}(t,\theta),\\
\frac{dC_i^{\mathrm{Num},N}}{dt}(t) &= -\frac{1}{\Delta x_N}\left[\frac{2D_i^{\mathrm{Num},N-1}(t,\theta)D_i^{\mathrm{Num},N}(t,\theta)}{\Delta x_{N-1} D_i^{\mathrm{Num},N}(t,\theta) + \Delta x_j D_i^{\mathrm{Num},N-1}(t,\theta)}\right]\left[C_i^{\mathrm{Num},N}(t) - C_i^{\mathrm{Num},N-1}(t)\right]\\
&+S_i^{\mathrm{Num},N}(t,\theta).
\end{aligned}$$

All diffusion related terms can be put into a triadiagonal $N \times N$ matrix that we will denote by $\boldsymbol{A}_i(t,\theta)$. Careful bookkeeping reveals that

$$\boldsymbol{A}_i(t,\theta) = \begin{pmatrix}
-A_i^1 & A_i^{1+\frac{1}{2}} & & & & & \\
A_i^{2-\frac{1}{2}} & -A_i^2 & A_i^{2+\frac{1}{2}} & & & & \\
& \ddots & \ddots & \ddots & & & \\
& & A_i^{j-\frac{1}{2}} & -A_i^j & A_i^{j+\frac{1}{2}} & & \\
& & & \ddots & \ddots & \ddots & \\
& & & & A_i^{(N-1)-\frac{1}{2}} & -A_i^{N-1} & A_i^{(N-1)+\frac{1}{2}} \\
& & & & & A_i^{N-\frac{1}{2}} & -A_i^N
\end{pmatrix}, \qquad (4.15)$$

where

$$\begin{aligned}
A_i^{j+1/2} = A_i^{j+1/2}(t,\theta) &:= \frac{1}{\Delta x_j}\left[\frac{2D_i^{\mathrm{Num},j}(t,\theta)D_i^{\mathrm{Num},j+1}(t,\theta)}{\Delta x_j D_i^{\mathrm{Num}j+1}(t,\theta) + \Delta x_{j+1} D_i^{\mathrm{Num},j}(t,\theta)}\right],\\
A_i^{j-1/2} = A_i^{j-1/2}(t,\theta) &:= \frac{1}{\Delta x_j}\left[\frac{2D_i^{\mathrm{Num},j-1}(t,\theta)D_i^{\mathrm{Num},j}(t,\theta)}{\Delta x_{j-1} D_i^{\mathrm{Num},j}(t,\theta) + \Delta x_j D_i^{\mathrm{Num},j-1}(t,\theta)}\right],\\
A_i^j = A_i^j(t,\theta) &:= A_i^{j+1/2}(t,\theta) + A_i^{j-1/2}(t,\theta).
\end{aligned}$$

If we further define a source term vector $\boldsymbol{S}_i^{\mathrm{Num}}(t,\theta) \in \mathbb{R}^N$ as

$$\boldsymbol{S}_i^{\mathrm{Num}}(t,\theta) := \left[S_i^{\mathrm{Num},1}(t,\theta), \ldots, S_i^{\mathrm{Num},N}(t,\theta)\right]^T \in \mathbb{R}^N,$$

then the semi-discrete system can be conveniently written as

$$\frac{d\boldsymbol{C}_i^{\mathrm{Num}}}{dt}(t) = \boldsymbol{A}_i(t,\theta)\boldsymbol{C}_i^{\mathrm{Num}}(t) + \boldsymbol{S}_i^{\mathrm{Num}}(t,\theta). \qquad (4.16)$$

4.2.3 Time Integration

Explicit Euler Scheme

To numerically solve the system (4.16), we discretize the equations in time as well. To this end, let Δt be the time step and define $t^k := k\Delta t$. The most straightforward time integration method is the so called *explicit Euler scheme*. For this scheme, we apply a finite difference approximation to the time derivative in equation (4.16) as follows:

$$\frac{\boldsymbol{C}_i^{\text{Num}}(t^{k+1}) - \boldsymbol{C}_i^{\text{Num}}(t^k)}{\Delta t} = \boldsymbol{A}_i(t^k, \theta)\boldsymbol{C}_i^{\text{Num}}(t^k) + \boldsymbol{S}_i^{\text{Num}}(t^k, \theta). \tag{4.17}$$

Upon rewriting, we find that - given the solution at time t^k - we can compute the solution at time t^{k+1} as

$$\boldsymbol{C}_i^{\text{Num}}(t^{k+1}) = \Big(\boldsymbol{I} + \Delta t\boldsymbol{A}_i(t^k, \theta)\Big)\, \boldsymbol{C}_i^{\text{Num}}(t^k) + \boldsymbol{S}_i^{\text{Num}}(t^k, \theta).$$

Here $\boldsymbol{I}$ is the $N \times N$ identity matrix. This scheme is referred to as the explicit Euler scheme is easy to implement but it may put severe constraints on the allowed time step Δt. For Δt larger than the allowed time step the method may become unstable and return useless results. The allowed time steps may be too small for the method to be useful in practice.

Implicit Euler Scheme

To overcome the stability issue issue with the explicit Euler scheme, we might want to use an implicit scheme. In general implicit schemes are able to handle much larger time steps. They may even be unconditionally stable. As an example, we consider the *implicit Euler scheme*. For the implicit Euler scheme, we replace all the t^k's in the right-hand side of equation (4.17) with t^{k+1}'s:

$$\frac{\boldsymbol{C}_i^{\text{Num}}(t^{k+1}) - \boldsymbol{C}_i^{\text{Num}}(t^k)}{\Delta t} = \boldsymbol{A}_i(t^{k+1}, \theta)\boldsymbol{C}_i^{\text{Num}}(t^{k+1}) + \boldsymbol{S}_i^{\text{Num}}(t^{k+1}, \theta).$$

Solving this system for $\boldsymbol{C}_i^{\text{Num}}(t^{k+1})$ is not straight-forward though. Iterative schemes may be needed.

Semi-Implicit Scheme

As an alternative, we could work with a *semi-implicit* method. That is, we discretize (4.16) in time as

$$\frac{\boldsymbol{C}_i^{\text{Num}}(t^{k+1}) - \boldsymbol{C}_i^{\text{Num}}(t^k)}{\Delta t} = \boldsymbol{A}_i(t^k, \theta)\hat{\boldsymbol{C}}_i(t^{k+1}) + \hat{\boldsymbol{S}}_i(t^k, \theta).$$

After rewriting, we see that

$$\Big(\boldsymbol{I} - \Delta t\boldsymbol{A}_i(t^k, \theta)\Big)\, \boldsymbol{C}_i^{\text{Num}}(t^{k+1}) = \boldsymbol{C}_i^{\text{Num}}(t^k) + \boldsymbol{S}_i^{\text{Num}}(t^k, \theta).$$

To solve for $\boldsymbol{C}_i^{\text{Num}}(t^{k+1})$ - the solution at the new timestep - we have to invert the matrix $\boldsymbol{I} - \boldsymbol{A}_i(t^k, T)$:

$$\boldsymbol{C}_i^{\text{Num}}(t^{k+1}) = \Big(\boldsymbol{I} - \Delta t\boldsymbol{A}_i(t^k, \theta)\Big)^{-1} \Big[\boldsymbol{C}_i^{\text{Num}}(t^k) + \boldsymbol{S}_i^{\text{Num}}(t^k, \theta)\Big]. \tag{4.18}$$

The downside of the semi-implicit scheme compared with the explicit Euler scheme is that we have to invert the matrices. For large matrices, this may be computationally expensive. However, for TNO/Solliance, N will be approximately equal to 250 and then `MATLAB` can invert the matrices quickly. More importantly though, semi-implicit schemes can handle much larger time steps than

the explicit Euler scheme. The full implicit Euler scheme is expected to be able to handle even larer time steps, but, as we saw it is more difficult to implement. In practice the semi-implicit method was found to be stable for the time steps used at TNO/Solliance. Hence this is the suggested time integration method for TNO/Solliance. Note that with the current compact matrix-vector notation it is easy to work out - if necessary - other time integration methods as well, like the more general *Runge-Kutta methods*.

4.2.4 Full Algorithm

The complete finite volume scheme with semi-implicit time integration can be summarized in algorithmic form as follows.

1. Given the numerical solution matrix

$$\boldsymbol{C}^{\mathrm{Num}}(t^k) = \left[\boldsymbol{C}_1^{\mathrm{Num}}(t^k), \ldots, \boldsymbol{C}_n^{\mathrm{Num}}(t^k)\right] \in \mathbb{R}^{N\times n}$$

   at time t^k, do the following for each $i \in \{1, \ldots, n\}$:

   (a) Evaluate the numerical diffusion coefficients matrix

$$\boldsymbol{D}^{\mathrm{Num}}(t^k, \theta) := \left[D_i^{\mathrm{Num},j}(t^k, \theta)\right]_{1\le i\le n, 1\le j\le N} \in \mathbb{R}^{N\times n}$$

   using definition (4.8);

   (b) Use the entries from $\boldsymbol{C}^{\mathrm{Num}}(t^k)$ and $\boldsymbol{D}^{\mathrm{Num}}(t^k, \theta)$ to set up the diffusion matrix $\boldsymbol{A}_i(t^k, T) \in \mathbb{R}^{N\times N}$ using definition (4.15);

   (c) Use the entries from $\boldsymbol{C}^{\mathrm{Num}}(t^k)$ to evaluate the reaction vector

$$\boldsymbol{S}_i^{\mathrm{Num}}(t^k, T) = \left[S_1^{\mathrm{Num}}(t, \theta), \ldots, S_1^{\mathrm{Num}}(t, \theta)\right] \in \mathbb{R}^N$$

   using definition (4.7);

   (d) Invert the matrix $\boldsymbol{I} + \Delta t \boldsymbol{A}_i(t^k, \theta)$;

   (e) Compute the concentration vector at the new time $\boldsymbol{C}_i^{\mathrm{Num}}(t^{k+1})$ as

$$\boldsymbol{C}_i^{\mathrm{Num}}(t^{k+1}) = \left(\boldsymbol{I} - \Delta t \boldsymbol{A}_i(t^k, \theta)\right)^{-1} \left[\boldsymbol{C}_i^{\mathrm{Num}}(t^k) + \boldsymbol{S}_i^{\mathrm{Num}}(t^k, \theta)\right].$$

Even though the reactions have been included in the numerical scheme, it is of course still possible to compute diffusion and reactions separately (as may be preferred by TNO/Solliance).

Now that we have a numerical scheme, we would have to show that the numerical solution obtained using this scheme converges to the 'real solution' (assuming it exists) to the equations described in section 4.1.7 as the time step Δt and the spatial steps Δx_j go to zero. Showing the convergence and deriving orders of convergence will be outside the scope of this work though.

The scheme is implemented in a stand-alone `MATLAB` script for testing. The scheme has also been incorporated into the current `MATLAB` script used at TNO/Solliance. It is seen to be approximately 400 times faster than the method based on error functions described in section 4.2.1. The diffusion mechanism is no longer the bottleneck in the full script used at TNO/Solliance and it can handle non-constant diffusion coefficients. Moreover, because everything has been formulated in terms of vectors and matrices it is easy to change the number of components, reactions, etcetera. The scheme also satisfies an important conservation property, as explained in the next paragraph.

4.2.5 Conservation Property

Suppose that no reactions occur within the precursor. Then, if the concentration C_i of component i satisfies the exact equation (4.2), we see that

$$\begin{aligned}
\partial_t \int_{[0,H]} C_i(x,t)dx &= \int_{[0,H]} \partial_t C_i(x,t)dx \\
&= \int_{[0,H]} \partial_x \left(D_i(\boldsymbol{C}(x,t),T)\partial_x C_i(x,t) \right) dx \\
&= \left[D_i(\boldsymbol{C}(x,t),T)\partial_x C_i(x,t) \right]_{x=0}^{x=H} \\
&= 0.
\end{aligned} \tag{4.19}$$

In other words, the total amount of component i is conserved. For TNO/Solliance it is important that a numerical scheme respects this conservation property. The numerical scheme worked out above satisfies this property. Indeed, suppose that at some point in time (most likely at the first timestep) the numerical solutions $C_i^{\mathrm{Num},j}$ coincide with the exact solutions $C_i^{\mathrm{Avg},j}$. Then we see that

$$\begin{aligned}
\sum_{j=1}^{N} C_i^{\mathrm{Num},j}(t^k)\Delta x_j &= \sum_{j=1}^{N} C_i^{\mathrm{Avg},j}(t^k)\Delta x_j \\
&= \sum_{j=1}^{N} \int_{\Omega^j} C_i(x,t^k)dx \\
&= \int_{[0,H]} C_i(x,t^k)dx.
\end{aligned} \tag{4.20}$$

Comparing equations (4.19) and (4.20) we deduce that our numerical method should conserve the sum $\sum_{j=1}^{N} C_i^{\mathrm{Num},j}(t)\Delta x_j$ over time. And it does, since

$$\begin{aligned}
\sum_{j=1}^{N} C_i^{\mathrm{Num},j}(t^{k+1})\Delta x_j &\overset{(4.14)}{=} \left[\sum_{j=1}^{N} C_i^{\mathrm{Num},j}(t^k)\Delta x_j \right] \\
&\quad -\Delta t \left[\sum_{j=1}^{N} F_i^{\mathrm{Num},j+1/2}(t^{k+1},\theta) - F_i^{\mathrm{Num},j-1/2}(t^{k+1},\theta) \right] \\
&\overset{(4.13)}{=} \left[\sum_{j=1}^{N} C_i^{\mathrm{Num},j}(t^k)\Delta x_j \right] \\
&\quad -\Delta t \left[F_i^{\mathrm{Num},N+1/2}(t^{k+1},\theta) - F_i^{\mathrm{Num},1/2}(t^{k+1},\theta) \right] \\
&= \sum_{j=1}^{N} C_i^{\mathrm{Num},j}(t^k)\Delta x_j.
\end{aligned}$$

For the last equality we used the no-flux boundary conditions (4.13). We conclude that the proposed numerical scheme respects the conservation property, as desired. Note that this statement is true for any choice of step sizes $\{\Delta x_j\}_{j=1}^{N}$ and not just in a limiting sense when the step sizes go to zero.

5 Mathematical Analysis of the Precursor Model

In Chapter 4 we presented the so called precursor model. In this chapter we show that the precursor model is well-posed. To this end, we first have to provide a weak formulation for the precursor model. Then, assuming the problem is linear for a moment, we employ the so called Method of Rothe to show that it is indeed well-posed. Then we return to the non-linear problem and treat the case of a single, scalar concentration first. We will not show well-posedness of the full non-linear problem in this thesis.

The main sources used in this chapter are the book *Partial Differential Equations* by Evans [19], *Applied Functional Analysis* by Zeidler [61], *Nonlinear Partial Differential Equations with Applications* by Roubíček [49] and the lecture notes *Parabolic Equations* by Pop [45]. The reader is assumed to have taken courses in functional analysis and (theory of) partial differential equations.

5.1 Towards a Weak Formulation

5.1.1 The Problem and the Objectives

Remember from Chapter 4 the following *precursor model*:

(Problem P) Given $\bar{\mathbf{D}}, \mathbf{S}, \boldsymbol{C}_{\text{Initial}}$ and θ, find $\boldsymbol{C} = \boldsymbol{C}(x,t)$ such that

$$\begin{cases} \partial_t \boldsymbol{C}(x,t) = \operatorname{div}_x \left(\bar{\mathbf{D}}(\boldsymbol{C}(x,t),\theta) \nabla_x \boldsymbol{C}(x,t) \right) + \mathbf{S}(\boldsymbol{C}(x,t),\theta) & \text{for } (x,t) \in \Omega \times (0,\infty), \\ \bar{\mathbf{D}}(\boldsymbol{C}(x,t),\theta) \nabla_x \boldsymbol{C}(x,t) \cdot \boldsymbol{n} = 0 & \text{for } (x,t) \in \partial\Omega \times (0,\infty), \\ \boldsymbol{C}(x,0) = \boldsymbol{C}_{\text{Initial}}(x) & \text{for } x \in \Omega. \end{cases} \tag{5.1}$$

Remember the general convention that bold symbols without bars refer to vector-like quantities while bold symbols with bars refer to matrices. The divergence operator, the gradient operator and the inner product with the normal derivative were understood to be applied component-wise. From now on, we will refer to the above set of equations as problem P. Our goal in this chapter will be to show that problem P is *well-posed,* meaning that:

1. Problem P admits a solution $\boldsymbol{C}$ in an appropriate setting to be specified later (the weak formulation as defined in problem WP);
2. The solution $\boldsymbol{C}$ is unique in this setting;
3. The solution $\boldsymbol{C}$ depends smoothly on the data $\boldsymbol{D}, \boldsymbol{S}, \theta$ and $\boldsymbol{C}_{\text{Initial}}$;
4. Because we are dealing with concentrations, we want each component of the solution vector $\boldsymbol{C}$ to take on positive values only.

In this chapter we will only work on proving existence and uniqueness. As a first step, we cast problem P in dimensionless form because we do not want to worry about dimensions when doing the mathematics.

5.1.2 Dimensionless Model

Let L and T represent dimensions of length and time respectively. Then C_i, D_i and S_i are seen to have dimensions

$$\begin{aligned}[C_i] &= \frac{\text{mol}}{L^3},\\ [D_i] &= \frac{\text{mol}L^2}{T},\\ [S_i] &= \frac{\text{mol}}{L^3 T},\end{aligned}$$

respectively.

Now introduce new independent, dimensionless variables $\hat{x}, \hat{t}$ and $\hat{\theta}$ as:

$$\begin{aligned}\hat{x} &:= x/x_{\text{ref}},\\ \hat{t} &:= t/t_{\text{ref}},\\ \hat{\theta} &:= \theta/\theta_{\text{ref}}.\end{aligned}$$

Here $x_{\text{ref}}, t_{\text{ref}}$ and θ_{ref} are reference quantities having the same dimensions as x, t and θ respectively. It will not be necessary to further specify them here. With these independent, dimensionless variables we define new dependent, dimensionless variables:

$$\begin{aligned}\hat{C}_i(\hat{x},\hat{t}) &:= C_i(x,t)/C_{\text{ref}},\\ \hat{\boldsymbol{C}}(\hat{x},\hat{t}) &:= \left[\hat{C}_1(\hat{x},\hat{t}),\ldots,C_n(\hat{x},\hat{t})\right],\\ \hat{S}_i(\hat{\boldsymbol{C}}(\hat{x},\hat{t}),\hat{\theta}) &:= S_i(C_i(x,t))/S_{\text{ref}}\\ \hat{\boldsymbol{S}}(\hat{\boldsymbol{C}}(\hat{x},\hat{t}),\hat{\theta}) &:= \left[\hat{S}_1(\hat{\boldsymbol{C}}(\hat{x},\hat{t}),\theta),\ldots,\hat{S}_n(\hat{\boldsymbol{C}}(\hat{x},\hat{t}),\theta)\right],\\ \hat{D}_{ij}(\hat{\boldsymbol{C}}(\hat{x},\hat{t}),\hat{\theta}) &:= D_{ij}(\boldsymbol{C}(x,t)\theta)/D_{\text{ref}},\\ \hat{\bar{\boldsymbol{D}}}(\hat{\boldsymbol{C}}(\hat{x},\hat{t}),\hat{\theta}) &:= \left[\hat{D}_1(\hat{\boldsymbol{C}}(\hat{x},\hat{t}),\theta),\ldots,\hat{D}_n(\hat{\boldsymbol{C}}(\hat{x},\hat{t}),\theta)\right].\end{aligned}$$

Again, $C_{\text{ref}}, D_{\text{ref}}$ and S_{ref} are reference quantities having the same dimensions as C_i, D_{ij} and S_i respectively. Now, using the chain rule it follows that

$$\begin{aligned}\partial_t \boldsymbol{C}(x,t) &= C_{\text{ref}}\partial_t \hat{\boldsymbol{C}}(\hat{x},\hat{t})\\ &= \frac{C_{\text{ref}}}{t_{\text{ref}}}\partial_{\hat{t}}\hat{\boldsymbol{C}}(\hat{x},\hat{t}),\end{aligned} \tag{5.2}$$

$$\begin{aligned}\bar{\mathbf{D}}(\boldsymbol{C}(x,t),T)\nabla_x \boldsymbol{C}(x,t) &= C_{\text{ref}}D_{\text{ref}}\hat{\bar{\boldsymbol{D}}}(\hat{\boldsymbol{C}}(\hat{x},\hat{t}),\hat{\theta})\nabla_x\hat{\boldsymbol{C}}(\hat{x},\hat{t})\\ &= \frac{C_{\text{ref}}D_{\text{ref}}}{x_{\text{ref}}}\hat{\bar{\boldsymbol{D}}}(\hat{\boldsymbol{C}}(\hat{x},\hat{t}),\hat{\theta})\nabla_{\hat{x}}\hat{\boldsymbol{C}}(\hat{x},\hat{t}),\end{aligned} \tag{5.3}$$

and

$$\begin{aligned}\text{div}_x\left(\bar{\mathbf{D}}(\boldsymbol{C}(x,t),T)\nabla_x\boldsymbol{C}(x,t)\right) &= \frac{C_{\text{ref}}D_{\text{ref}}}{x_{\text{ref}}}\text{div}_x\left(\hat{\bar{\boldsymbol{D}}}(\hat{\boldsymbol{C}}(\hat{x},\hat{t}),\hat{\theta})\nabla_{\hat{x}}\hat{\boldsymbol{C}}(\hat{x},\hat{t})\right)\\ &= \frac{C_{\text{ref}}D_{\text{ref}}}{x_{\text{ref}}^2}\text{div}_{\hat{x}}\left(\hat{\bar{\boldsymbol{D}}}(\hat{\boldsymbol{C}}(\hat{x},\hat{t}),\hat{\theta})\nabla_{\hat{x}}\hat{\boldsymbol{C}}(\hat{x},\hat{t})\right).\end{aligned} \tag{5.4}$$

Substituting (5.2) and (5.4) into the partial differential equation of *Problem P* yields

$$\frac{C_{\text{ref}}}{t_{\text{ref}}}\partial_{\hat{t}}\hat{\boldsymbol{C}}(\hat{x},\hat{t}) = \frac{C_{\text{ref}}D_{\text{ref}}}{x_{\text{ref}}^2}\text{div}_{\hat{x}}\left(\hat{\bar{\boldsymbol{D}}}(\hat{\boldsymbol{C}}(\hat{x},\hat{t}),\hat{\theta})\nabla_{\hat{x}}\hat{\boldsymbol{C}}(\hat{x},\hat{t})\right) + S_{\text{ref}}\hat{\boldsymbol{S}}(\hat{\boldsymbol{C}}(\hat{x},\hat{t}),\hat{\theta}).$$

After dividing both sides by $C_{\text{ref}}/t_{\text{ref}}$ we find that

$$\partial_{\hat{t}}\hat{\boldsymbol{C}}(\hat{x},\hat{t}) = \frac{t_{\text{ref}}D_{\text{ref}}}{x_{\text{ref}}^2}\text{div}_{\hat{x}}\left(\hat{\bar{\boldsymbol{D}}}(\hat{\boldsymbol{C}}(\hat{x},\hat{t}),\hat{\theta})\nabla_{\hat{x}}\hat{\boldsymbol{C}}(\hat{x},\hat{t})\right) + \frac{t_{\text{ref}}S_{\text{ref}}}{C_{\text{ref}}}\hat{\boldsymbol{S}}(\hat{\boldsymbol{C}}(\hat{x},\hat{t}),\hat{\theta}).$$

If we choose $D_{\text{ref}} = x_{\text{ref}}/t_{\text{ref}}^2$ and $S_{\text{ref}} = C_{\text{ref}}/t_{\text{ref}}$ then we see that

$$\partial_{\hat{t}}\hat{\boldsymbol{C}}(\hat{x},\hat{t}) = \text{div}_{\hat{x}}\left(\hat{\bar{\boldsymbol{D}}}(\hat{\boldsymbol{C}}(\hat{x},\hat{t}),\hat{\theta})\nabla_{\hat{x}}\hat{\boldsymbol{C}}(\hat{x},\hat{t})\right) + \hat{\boldsymbol{S}}(\hat{\boldsymbol{C}}(\hat{x},\hat{t}),\hat{\theta}).$$

Finally, using the initial conditions for the dimensional concentrations we arrive at initial conditions for the dimensionless concentrations:

$$\hat{\boldsymbol{C}}(x,0) = \boldsymbol{C}_{\text{Initial}}(x)/C_{\text{ref}} \quad \text{for } x \in \Omega.$$

Similarly, for the boundary conditions use equation (5.3) together with the boundary conditions for the dimensional concentrations to arrive at

$$\hat{\bar{\boldsymbol{D}}}(\hat{\boldsymbol{C}}(\hat{x},\hat{t}),\hat{\theta})\nabla_{\hat{x}}\hat{\boldsymbol{C}}(\hat{x},\hat{t})\cdot\boldsymbol{n} = \frac{x_{\text{ref}}}{C_{\text{ref}}D_{\text{ref}}}\bar{\mathbf{D}}(\boldsymbol{C}(x,t),T)\nabla_x\boldsymbol{C}(x,t)\cdot\boldsymbol{n} = 0 \quad \text{for } (x,t)\in\partial\Omega\times(0,\infty).$$

From now on we drop the hats from the notation. Note that all the equations are the same as before then, except for the initial conditions. Since dividing the initial conditions by C_{ref} is not going to have any influence on the analysis below, we will simply think of problem P being dimensionless already.

5.1.3 Notations and Preliminaries

Before we start the mathematical analysis of the (dimensionless) model we introduce notations and state results that will be used throughout this chapter. Most of the definitions and results should be familiar, or at least understandable, to the reader with a background in functional analysis. References to proofs of the results will be given when appropriate.

Hilbert and Banach Spaces: Let H be a real *Hilbert space* with inner product $\langle\cdot,\cdot\rangle_H : H\times H\to\mathbb{R}$. Recall that the inner product gives rise to a norm $\|\cdot\|_H : H\to\mathbb{R}$ defined as

$$\|f\|_H = \sqrt{\langle f,f\rangle_H}.$$

A fundamental inequality that relates the norm and the inner product on a Hilbert space is the *Cauchy-Schwartz inequality*, which can be expressed as

$$|\langle f,g\rangle_H| \le \|f\|_H\|g\|_H$$

for all $f,g\in H$. Now suppose that $H_1,\dots,H_n$ are Hilbert spaces with inner products $\langle\cdot,\cdot\rangle_{H_1},\dots,\langle\cdot,\cdot\rangle_{H_n}$ and consider the product space

$$H := H_1\times\cdots\times H_n.$$

Then $\langle\cdot,\cdot\rangle_H : H\times H\to\mathbb{R}$ defined as

$$\langle\boldsymbol{f},\boldsymbol{g}\rangle_H := \sum_{i=1}^n \langle f_i,g_i\rangle_{H_i},$$

is an inner product on H that turns it into a Hilbert space. In a similar fashion we find that for Banach spaces $X_1,\dots,X_n$ the product space

$$X := X_1\times\ldots\times X_n$$

is a Banach space again when endowed with the norm

$$\|\boldsymbol{f}\|_X := \left(\sum_{i=1}^n \|f_i\|_{X_i}\right)^{1/2}.$$

L^2 Spaces: Let $\Omega \subset \mathbb{R}^d$. The space $L^2(\Omega)$ consists of equivalence classes of measurable functions $f : \Omega \to \mathbb{R}$ which are square-integrable over Ω. For ease of notation, when $f \in L^2(\Omega)$, we simply think of f being a square-integrable function and gently ignore the fact that f is an equivalence class of functions. The space $L^2(\Omega)$ becomes a Hilbert space if we endow it with the inner product

$$\langle f, g \rangle_{L^2(\Omega)} := \int_\Omega f(x)g(x)dx.$$

Weak Derivatives: Given a function $f \in L^2(\Omega)$, we say that f has a *weak $i-th$ partial derivative* in $L^2(\Omega)$ if there exists a function $g_i \in L^2(\Omega)$ such that

$$\int_\Omega f(x)\partial_{x_i}\varphi(x)dx = -\int_\Omega g_i(x)\varphi(x)dx$$

holds for all $\varphi \in C_c^\infty(\Omega)$, where $C_c^\infty(\Omega)$ is the space compactly supported smooth functions on Ω. We denote the weak $i-$th partial derivative g_i of f by $\partial_{x_i} f$. The vector of (weak) partial derivatives will be denoted by ∇f. With these notations the *Sobolev space* $W^{1,2}(\Omega)$ is defined as

$$W^{1,2}(\Omega) := \left\{ f \in L^2(\Omega) \;\middle|\; \partial_{x_i} f \in L^2(\Omega) \text{ for all } 1 \le i \le d \right\}.$$

A natural inner product that turns $W^{1,2}(\Omega)$ into a Hilbert space is given by

$$\begin{aligned} \langle f, g \rangle_{W^{1,2}(\Omega)} &:= \int_\Omega f(x)g(x)dx + \sum_{i=1}^d \int_\Omega \partial_{x_i} f(x)\partial_{x_i} g(x)dx \\ &= \int_\Omega f(x)g(x)dx + \int_\Omega \nabla_x f(x) \cdot \nabla_x g(x)dx. \end{aligned}$$

Dual Spaces: Given a Banach space X, its *dual space* X' is defined as

$$X' := \{\varphi : X \to \mathbb{R} \;\mid\; \varphi \text{ linear and bounded}\}.$$

The duality pairing between X and X' will be denoted by $\langle \cdot, \cdot \rangle_{X',X}$. That is,

$$\langle \varphi, f \rangle_{X',X} := \varphi(f).$$

A natural norm on X' which turns it into a Banach space is given by

$$\|\varphi\|_{X'} := \sup_{f \in X, \|f\|_X = 1} \left| \langle \varphi, f \rangle_{X',X} \right|.$$

In the particular case of $X = W^{1,2}(\Omega)$ we denote the dual space by $W^{-1,2}(\Omega)$. If $X = X_1 \times \ldots \times X_n$ then - in the same way as for inner product on products of Hilbert spaces - we define the duality pairing $\langle \cdot, \cdot \rangle_{X',X} : X' \times X \to \mathbb{R}$ as

$$\langle \boldsymbol{\varphi}, \boldsymbol{f} \rangle_{X',X} := \sum_{i=1}^n \langle \varphi_i, f_i \rangle_{X_i',X_i}.$$

A Banach space X is said to be *reflexive* if it is isomorphic to $(X')'$, the dual of its dual.

Strong and Weak Convergence: Let X be a Banach space. A sequence $\{f_n\}_{n \ge 0}$ in X is said to converge *strongly* to $f \in X$ if

$$\|f_n - f\|_X \to 0$$

as $n \to \infty$. A sequence $\{f_n\}_{n \ge 0}$ in X is said to converge *weakly* to $f \in X$ if for each $\varphi \in X'$ we have that

$$\langle \varphi, f - f_n \rangle_{X',X} \to 0$$

as $n \to \infty$.

Eberlein–ȩmulian Theorem: An important result in real analysis is that bounded sequences in $X = \mathbb{R}^n$ have *strongly* convergent subsequences. If X is an infinite dimensional space this statement is no longer true. However, the Eberlein–ȩmulian theorem tells us that if X is a reflexive Banach space, then any bounded sequence $\{x_n\}$ in X has a *weakly* convergent subsequence. Moreover, if all subsequences of $\{x_n\}$ converge weakly to the same limit x in X, then $\{x_n\}$ itself converges weakly to x. See Appendix D of Evans [19].

Riesz' Representation Theorem: Riesz' Representation theorem establishes a connection between a Hilbert space H and its dual H'. More specifically, the theorem states that if H is a Hilbert space, then for each $\varphi \in H'$ there exists a unique $f \in H$ such that

$$\langle \varphi, g \rangle_{H',H} = \langle f, g \rangle_H$$

and $\|\varphi\|_{H'} = \|g\|_H$. See Appendix D of Evans [19].

Lax-Milgram Theorem: Riesz' Representation theorem is in fact a special case of a more general theorem: the Lax-Milgram theorem. This latter theorem is an important tool in showing well-posedness of time-independent (elliptic boundary value) problems. Given $B : H \times H \to \mathbb{R}$ and $F : H \to \mathbb{R}$, the linear Lax-Milgram theorem states that if B is bounded and bilinear, and there exists a positive constant α such that $B(f, f) \geq \alpha \|f\|_H^2$ for all $f \in H$ (i.e. B is *coercive*), then for any $\varphi \in H'$ there exists a unique $g \in H$ such that for every $f \in H$ the equality

$$B(f, g) = \varphi(f)$$

holds. A proof of the linear Lax-Milgram theorem can be found in Chapter 6.2.1 of Evans [19].

There is also a non-linear version of the Lax-Milgram theorem. This version states that if for each fixed $f \in H$, the functional $B(f, \cdot) : H \to \mathbb{R}$ is bounded and linear, and if there exist positive constants α and β such that for every $f, g, h \in H$

$$B(f, f - g) - B(g, f - g) \geq \alpha \|f - g\|_H^2$$

and

$$|B(f, h) - B(g, h)| \leq \beta \|f - g\|_H \|h\|_H;$$

then for any $\varphi \in H'$ there exists a unique $g \in H$ such that for every $f \in H$ the equality

$$B(f, g) = \varphi(f)$$

holds. A proof of the non-linear Lax-Milgram theorem can be found in Chapter 2.15 of Zeidler [61].

Continuous and Compact Embeddings: Let X and Y be Banach spaces such that $X \subset Y$. We will say that X is *continously embedded* in Y if there exists a constant $C > 0$ such that for every $f \in X$ we have that

$$\|f\|_Y \leq C \|f\|_X.$$

If, on top of that, every bounded sequence has a strongly convergent subsequence in Y, then we say that X is *compactly embedded* in Y.

The space $W^{1,2}(\Omega)$ is compactly embedded in $L^2(\Omega)$, see Chapter 5.7 of Evans [19]. The space $L^2(\Omega)$ is easily seen to be continuously embedded in $W^{-1,2}(\Omega)$.

$L^2(0,T;X)$ Spaces: Note that a solution $\boldsymbol{C} = [C_1, \dots, C_n]$ to *Problem P* is expected to have both a space (x) and a time (t) dependence. However, instead of thinking of the functions C_i as being defined on the whole space-time domain $\Omega \times [0,\infty)$, we will think of each C_i as a function mapping a the time interval $[0,T]$ into some appropriate set of functions on Ω. For example, let $t \in [0,T]$ and suppose that $C_i(t) \in L^2(\Omega)$. Then $[C_i(t)](x) \in \mathbb{R}$ can be identified with $C_i(x,t)$. As for conditions on C_i itself, it will be convenient to work in a Hilbert space setting. To make this more precise, let X be a Banach space. Then the space $L^2(0,T;X)$ is defined as the set of all (equivalence classes of) measurable functions $f : [0,T] \to X$ for which

$$\left(\int_0^T \|f(t)\|_X^2 dt \right)^{1/2} < \infty.$$

It should be noted here that measurability - and also (Lebesgue) integrability - of functions taking on values in a Banach space is defined in exactly the same way as for real-valued functions. The norm

$$\|f\|_{L^2(0,T;X)} := \left(\int_0^T \|f(t)\|_X^2 dt \right)^{1/2}$$

turns $L^2(0,T;X)$ into a Banach space. The dual of $L^2(0,T;X)$ is given by $L^2(0,T;X')$ (see Proposition 1.38 of Roubíček [49]). Moreover, if X is a Hilbert space then $L^2(0,T;X)$ endowed with the inner product

$$\langle f, g \rangle_{L^2(0,T;X)} := \int_0^T \langle f(t), g(t) \rangle_X \, dt$$

is a Hilbert space as well. The notions of continuity and of weak derivatives can easily be extended to Banach-valued functions as well. This allows us to work with spaces like $C(0,T;X)$ and $W^{1,2}(0,T;X)$. See Chapter 5.9.2 of Evans [19] for more details on these types of spaces. In this chapter the space X will either be $L^2(\Omega), W^{1,2}(\Omega)$ or $W^{-1,2}(\Omega)$.

The space $\mathcal{W}$: Another important space will be

$$\mathcal{W} := \left\{ f \in L^2(0,T;W^{1,2}(\Omega)) \;\middle|\; f \text{ has a weak derivative } \partial_t f \in L^2(0,T;W^{-1,2}(\Omega)) \right\}.$$

This space resembles $W^{1,2}(0,T;W^{1,2}(\Omega))$, but the difference is that we impose less strict conditions on the weak derivative. Indeed, since $W^{1,2}(\Omega) \subset L^2(\Omega)$ it follows directly that the reverse inclusion holds for their dual spaces. Identifying $L^2(\Omega)$ with its dual $L^{-2}(\Omega)$ using Riesz' Representation theorem we find that

$$W^{1,2}(\Omega) \subset L^2(\Omega) = L^{-2}(\Omega) \subset W^{-1,2}(\Omega).$$

The space $\mathcal{W}$ is a Banach space when endowed with the norm

$$\begin{aligned} \|f\|_{\mathcal{W}} &:= \left(\|f\|^2_{L^2(0,T;W^{1,2}(\Omega))} + \|\partial_t f\|^2_{L^2(0,T;W^{-1,2}(\Omega))} \right)^{1/2} \\ &= \left(\int_0^T \|f\|^2_{W^{1,2}(\Omega)} + \|\partial_t f\|^2_{W^{-1,2}(\Omega)} dt \right)^{1/2} \end{aligned}$$

Moreover, it is continuously embedded in $C(0,T;W^{1,2}(\Omega))$ and compactly embedded in $L^2(0,T;L^2(\Omega))$, see Theorem 3 in Chapter 5.9.2 of Evans [19] and Lemma 7.7 of Roubíček [49].

In the space $\mathcal{W}$ we also have an analogue of integration by parts: Let $f, g \in \mathcal{W}$, then for a.e. $t \in [0,T]$ the equality

$$\frac{d}{dt} \int_\Omega f(t) g(t) dx = \langle \partial_t f(t), g(t) \rangle_{W^{-1,2}, W^{1,2}} + \langle \partial_t g(t), f(t) \rangle_{W^{-1,2}, W^{1,2}} \tag{5.5}$$

holds, see Lemma 7.3 of Roubíček [49].

The need for the space $\mathcal{W}$ will be motivated later.

Limits and Derivatives: Suppose X and Y are Banach spaces such that X is continuously embedded in Y. Let $\{f_n\}_{n\geq 0}$ be a sequence in $L^2(0,T;X)$ converging weakly to $f \in L^2(0,T;X)$ and let $\{g_n\}_{n\geq 0}$ be a sequence in $L^2(0,T;Y)$ converging weakly to $g \in L^2(0,T;Y)$. If $\partial_t f_n = g_n$ for all $n \geq 0$ then $\partial_t f = g$ (in a weak sense).

We will be using this theorem in the setting of the space $\mathcal{W}$ where $X = W^{1,2}(\Omega)$ and $Y = W^{-1,2}(\Omega)$.

5.1.4 The Weak Formulation of Problem

With the notations and preliminaries from the previous subsection in mind, we define the weak formulation of problem P as follows:

(Problem WP) Find $\boldsymbol{C} \in \mathcal{W}^n$ such that $\boldsymbol{C}(0) = \boldsymbol{C}_{\text{Initial}}$ and for all $\boldsymbol{\varphi} = [\varphi_1, \ldots, \varphi_n] \in \left[W^{1,2}(\Omega)\right]^n$ the equality

$$\langle \partial_t \boldsymbol{C}(t), \boldsymbol{\varphi} \rangle_{[W^{-1,2}]^n,[W^{1,2}]^n} \quad = \quad -\left\langle \bar{\boldsymbol{D}}(\boldsymbol{C}(t),\theta)\nabla_x \boldsymbol{C}(t), \nabla_x \boldsymbol{\varphi} \right\rangle_{[L^2]^n} + \langle \mathbf{S}(\boldsymbol{C}(t),\theta), \boldsymbol{\varphi} \rangle_{[L^2]^n} \quad (5.6)$$

holds for a.e. $t \in [0,T]$.

To motivate the particular form of the weak formulation, consider the following model diffusion equation defined on the space-time domain $\Omega \times [0,T]$:

$$P_{\text{Model}} := \begin{cases} \partial_t C = \Delta_x C & \text{in } \Omega \times (0,T], \\ \nabla_x C \cdot \mathbf{n} = 0 & \text{on } \partial\Omega \times (0,T], \\ C = C_{\text{Initial}} & \text{on } \Omega \times \{0\}. \end{cases}$$

To show existence of solutions to such a problem we have to decide what kind of solutions we are looking for in the first place. More specifically, in what kind of function space do we want/expect our possible solution to be found?

- Suppose we have found a solution C to P_{Model}. As mentioned before, instead of thinking of C as being defined on $\Omega \times [0,T]$, we think of C as mapping $[0,T]$ into some appropriate function space on Ω. But what should this appropriate function space be? The *strong formulation* of the partial differential equation in P_{Model}, i.e. $\partial_t C = \Delta C$, requires $C(t)$ to be twice differentiable. However, if we multiply the equation by some smooth test function φ, integrate over Ω and apply partial integration - as one also does for time independent problems - we find that

  $$\int_\Omega \partial_t C(t)\varphi dx = -\int_\Omega \nabla_x C(t) \cdot \nabla_x \varphi dx. \quad (5.7)$$

  In terms of this *weak formulation*, we see that it is sufficient for $C(t)$ to have weak first (spatial) derivatives only. This motivates us to require $C(t) \in W^{1,2}(\Omega)$. That is, $C : [0,T] \to W^{1,2}(\Omega)$.

- What about C itself? From the weak formulation (5.7) we deduce that C should have a weak derivative with respect to time. And because it is convenient to work in the setting of Hilbert spaces, we should require C to be square-integrable. Hence, we may be tempted to require $C \in W^{1,2}(0,T;W^{1,2})$.

- If C were in this space $W^{1,2}(0,T;W^{1,2})$, its weak derivative $\partial_t C$ would have to be square integrable as well. We can relax this requirement though. This can be seen by looking at the weak formulation (5.7) again. For $C(t) \in W^{1,2}(\Omega)$, the mapping

  $$f \mapsto -\int_\Omega \nabla_x C(t) \cdot \nabla_x f dx.$$

defines a bounded linear functional on $W^{1,2}(\Omega)$. In other words, the mapping is an element of $W^{-1,2}(\Omega)$, the dual space of $W^{1,2}(\Omega)$. On the other hand,

$$\int_\Omega \partial_t C(t)\varphi dx = \langle \partial_t C(t), \varphi \rangle_{L^2(\Omega)}.$$

Using Riesz' Representation theorem we identify a unique element in $W^{-1,2}(\Omega)$, which we will simply denote by $\partial_t C(t)$ as well, such that

$$\langle \partial_t C(t), \varphi \rangle_{W^{-1,2}(\Omega), W^{-1,2}(\Omega)} = \langle \partial_t C(t), \varphi \rangle_{L^2(\Omega)}.$$

Putting the above together, we see that the weak formulation (5.7) can be rewritten as

$$\langle \partial_t C(t), \varphi \rangle_{W^{-1,2}(\Omega), W^{-1,2}(\Omega)} = -\int_\Omega \nabla_x C(t) \cdot \nabla_x \varphi dx.$$

This equality shows us that it is sufficient for $\partial_t C(t)$ to be in $W^{-1,2}(\Omega)$.

- Putting everything together we see that C should be in the space

$$\left\{ f \in L^2(0,T; W^{1,2}(\Omega)) \;\middle|\; f \text{ has a weak derivative } \partial_t f \in L^2(0,T; W^{-1,2}(\Omega)) \right\},$$

and that is precisely the space $\mathcal{W}$ defined earlier.

- One problem that we ignored so far is that in an L^2 setting specifying a function at a particular point is meaningless. Hence the initial condition $C(0) = C_{\text{Initial}}$ is no condition at all in such a setting! Luckily though, as mentioned in the preliminaries, the space $\mathcal{W}$ is continuously embedded in $C(0,T; L^2(\Omega))$. This means that any function in $\mathcal{W}$ can be interpreted as a continuous function of time (after a possible redefinition on a set of measure zero). And then it makes sense again to impose an initial condition.

- For the real problem, we proceed as above for each component i (where this time a source term but that does not change anything) and in the end we sum over i. This yields Problem WP.

It should be noted that problem WP is a non-linear problem. In general non-linear problems are difficult to solve. Therefore we will first reduce the problem to a linear problem WP_{Linear}. After having showed well-posedness of the linear problem WP_{Linear} we will return to the non-linear problem WP.

5.2 The Linear Case

5.2.1 Formulation of the Linear Problem

To make the problem linear, we assume that the diffusion coefficients do not depend on the concentrations. That is, the diffusion coefficient matrix $\bar{\boldsymbol{D}}$ is assumed to be independent of $\boldsymbol{C}$ (but it is still allowed to depend on the temperature θ). Furthermore, we assume the source terms only to depend on the concentrations in a linear fashion. That is, we assume the function S_i to be of the form

$$S_i(\boldsymbol{C}(t)) = \sum_{j=1}^{n} S_{ij}(\theta) C_j(t),$$

where the S_{ij} are possibly temperature-dependent reaction coefficients. We use the notation $\bar{\boldsymbol{S}}(\theta)$ for the matrix with entries $S_{ij}(\theta)$. Then the problem P_{Linear} becomes:

(Problem P_{Linear}) Find $\boldsymbol{C} = \boldsymbol{C}(x,t)$ such that

$$\begin{cases} \partial_t \boldsymbol{C}(x,t) = \operatorname{div}_x \left(\bar{\mathbf{D}}(\theta) \nabla_x \boldsymbol{C}(x,t) \right) + \bar{\boldsymbol{S}}(\theta) \boldsymbol{C}(x,t) & \text{for } (x,t) \in \Omega \times (0,\infty), \\ \bar{\mathbf{D}}(\theta) \nabla_x \boldsymbol{C}(x,t) \cdot \boldsymbol{n} = 0 & \text{for } (x,t) \in \partial\Omega \times (0,\infty), \\ \boldsymbol{C}(x,0) = \boldsymbol{C}_{\text{Initial}}(x) & \text{for } x \in \Omega. \end{cases} \tag{5.8}$$

Note that the equations are linear in $\boldsymbol{C}$. The corresponding weak formulation problem WP_{Linear} becomes:

(Problem WP_{Linear}) Find $\boldsymbol{C} \in \mathcal{W}^n$ such that $\boldsymbol{C}(0) = \boldsymbol{C}_{\text{Initial}}$ and for all $\boldsymbol{\varphi} \in \left[W^{1,2}(\Omega)\right]^n$ the equality

$$\langle \partial_t \boldsymbol{C}(t), \boldsymbol{\varphi} \rangle_{[W^{-1,2}]^n, [W^{1,2}]^n} = - \left\langle \bar{\boldsymbol{D}}(\theta) \nabla_x \boldsymbol{C}(t), \nabla_x \boldsymbol{\varphi} \right\rangle_{[L^2]^n} + \left\langle \bar{\mathbf{S}}(\theta) \boldsymbol{C}(t), \boldsymbol{\varphi} \right\rangle_{[L^2]^n}$$

holds for a.e. $t \in [0,T]$. For similar problems of this type, see the paper *Multiscale Modeling of Colloidal Dynamics in Porous Media: Capturing Aggregation and Deposition Effects* by Krehel et al [33].

In what follows we are going to make the following assumptions:

($H_{\textbf{Linear},1}$) The entries D_{ij} of the diffusion coefficient matrix $\bar{\boldsymbol{D}}$ are continuous functions of temperature θ. For each temperature θ the matrix $\bar{\boldsymbol{D}}(\theta)$ is symmetric positive-definite.

($H_{\textbf{Linear},2}$) The entries S_{ij} of the reaction coefficient matrix $\bar{\boldsymbol{S}}$ are continuous functions of temperature θ. For each temperature θ the matrix $\bar{\boldsymbol{S}}(\theta)$ is symmetric negative-definite.

($H_{\textbf{Linear},3}$) The initial conditions are weakly differentiable in space, i.e. $\boldsymbol{C}_{\text{Initial}} \in \left[W^{1,2}(\Omega)\right]^n$.

We refer to these assumptions collectively as *the assumptions* H_{Linear}. Note that under the assumptions H_{Linear} the eigenvalues of $\bar{\boldsymbol{D}}(\theta)$ are real and positive whereas the eigenvalues of $\bar{\boldsymbol{S}}(\theta)$ are real and negative.

Now how do we proceed in showing existence and uniqueness of problem WP_{Linear}? To this end, remember that in deriving a numerical scheme for finding numerical approximations to the solution of problem P in Chapter 4 we used the so called Method of Lines. That is, we first discretized the equations of problem P in space to arrive at systems of ordinary differential equations. Then we discretized the ordinary differential equations in time using either an explicit, implicit or semi-implicit scheme. This time, we are going to start discretizing the weak formulation in time using the implicit Euler scheme and a time step Δt. This results in a bunch of time-independent (elliptic) problems that can be solved rather easily. Then the idea is to patch together these solutions to obtain a 'solution' defined on the whole time interval $[0,T]$. If we then pass to the limit $\Delta t \to 0$, this 'solution' turns out to be *the solution* to problem WP_{Linear}. This method of showing existence and uniqueness to partial differential equations is referred to as the *Method of Rothe* [49, 48].

Let us start now by discretizing the weak formulation in time.

5.2.2 Discretizing in Time

Given the time interval $[0,T]$ and a natural number N, we define a time step Δt and introduce discrete times $t^k := k\Delta t$. For each $k \in \{1, \dots, N\}$ we introduce the problem $WP^k_{\text{Linear},\Delta t}$ as:

(Problem $WP^k_{\text{Linear},\Delta t}$) Given $\boldsymbol{C}^{k-1} \in \left[W^{1,2}(\Omega)\right]^n$, find $\boldsymbol{C}^k \in \left[W^{1,2}(\Omega)\right]^n$ such that for all $\boldsymbol{\varphi} \in \left[W^{1,2}(\Omega)\right]^n$ the equality

$$\left\langle \frac{\boldsymbol{C}^k - \boldsymbol{C}^{k-1}}{\Delta t}, \boldsymbol{\varphi} \right\rangle_{[L^2]^n} = - \left\langle \bar{\boldsymbol{D}}(\theta) \nabla_x \boldsymbol{C}^k, \nabla_x \boldsymbol{\varphi} \right\rangle_{[L^2]^n} + \left\langle \bar{\mathbf{S}}(\theta) \boldsymbol{C}^k, \boldsymbol{\varphi} \right\rangle_{[L^2]^n} \tag{5.9}$$

holds. The sequence of problem is initialized by $\boldsymbol{C}^0 := \boldsymbol{C}^{\text{Initial}}$. Our claim is that each $WP^k_{\text{Linear},\Delta t}$ has a unique solution.

Lemma 1. *Under the assumptions H_{Linear}, the problem $WP^k_{Linear,\Delta t}$ admits a unique solution $\boldsymbol{C}^k \in \left[W^{1,2}(\Omega)\right]^n$ for each $k \in \{1,\dots,N\}$.*

Proof. Note that each problem $WP^k_{\text{Linear},\Delta t}$ is independent of time as a result of the time-discretization. Hence we can apply the Lax Milgram theorem to show the existence and uniqueness of solutions to $WP^k_{\text{Linear},\Delta t}$. To this end define $B^k_{\Delta t} : \left[W^{1,2}(\Omega)\right]^n \times \left[W^{1,2}(\Omega)\right]^n \to \mathbb{R}$ as

$$B^k_{\Delta t}(\boldsymbol{u},\boldsymbol{v}) \quad := \quad \langle \boldsymbol{u},\boldsymbol{v}\rangle_{[L^2]^n} + \Delta t \left\langle \bar{\boldsymbol{D}}(\theta)\nabla_x \boldsymbol{u}, \nabla_x \boldsymbol{v}\right\rangle_{[L^2]^n} - \Delta t \left\langle \bar{\mathbf{S}}(\theta)\boldsymbol{u},\boldsymbol{v}\right\rangle_{[L^2]^n}$$

and $F^k_{\Delta t} : \left[W^{1,2}(\Omega)\right]^n \to \mathbb{R}$ as

$$F^k_{\Delta t}(\boldsymbol{v}) \quad := \quad \left\langle \boldsymbol{C}^{k-1}, \boldsymbol{v}\right\rangle_{[L^2]^n}.$$

Then the time-discretized weak formulation (5.9) can be expressed as

$$B^k_{\Delta t}(\boldsymbol{C}^k, \boldsymbol{\varphi}) = F^k_{\Delta t}(\boldsymbol{\varphi}).$$

To apply the Lax-Milgram theorem, we have to show that $B^k_{\Delta t}$ and $F^k_{\Delta t}$ are *linear* and *bounded* and that $B^k_{\Delta t}$ is *coercive*. Showing linearity is easy and will not be presented here. For boundedness of $F^k_{\Delta t}$, let $\boldsymbol{v} \in \left[W^{1,2}(\Omega)\right]^n$. Then

$$\begin{aligned} |F^k_{\Delta t}(\boldsymbol{v})| &= \left|\left\langle \boldsymbol{C}^{k-1}, \boldsymbol{v}\right\rangle_{[L^2]^n}\right| \\ &\overset{\text{Cauchy-Schwartz}}{\leq} \|\boldsymbol{C}^k\|_{[L^2]}\|\boldsymbol{v}\|_{[L^2]} \\ &= \|\boldsymbol{C}^k_i\|_{[W^{1,2}]^n}\|\boldsymbol{v}\|_{[W^{1,2}]^n}, \end{aligned}$$

which proves that $F^k_{\Delta t}$ is bounded. For boundedness of $B^k_{\Delta t}$, remember first that the matrix $\bar{\boldsymbol{D}}(\theta)$ is assumed to be symmetric. Hence we can diagonalize $\bar{\boldsymbol{D}}(\theta)$ as

$$\bar{\boldsymbol{D}}(\theta) = \bar{\boldsymbol{U}}^T(\theta)\bar{\boldsymbol{\Lambda}}(\theta)\bar{\boldsymbol{U}}(\theta),$$

with $\bar{\boldsymbol{U}}(\theta)$ an orthogonal matrix and $\bar{\boldsymbol{\Lambda}}(\theta)$ a diagonal matrix containing the eigenvalues of $\bar{\boldsymbol{D}}(\theta)$ as entries. Let $\lambda_1(\theta),\dots,\lambda_n(\theta)$ be the eigenvalues of $\bar{\boldsymbol{D}(\theta)}$ and define

$$\lambda_\infty(\theta) := \max_{1\leq i\leq n} |\lambda_i(\theta)|$$

Then, for any $\boldsymbol{u},\boldsymbol{v} \in \left[L^2(\Omega)\right]^n$ we find that

$$\begin{aligned} \left\langle \bar{\boldsymbol{D}}(\theta)\boldsymbol{u},\boldsymbol{v}\right\rangle_{[L^2]^n} &= \left\langle \bar{\boldsymbol{U}}^T(\theta)\bar{\boldsymbol{\Lambda}}(\theta)\bar{\boldsymbol{U}}(\theta)\boldsymbol{u},\boldsymbol{v}\right\rangle_{[L^2]^n} \\ &= \left\langle \bar{\boldsymbol{\Lambda}}(\theta)\bar{\boldsymbol{U}}(\theta)\boldsymbol{u}, \bar{\boldsymbol{U}}(\theta)\boldsymbol{v}\right\rangle_{[L^2]^n} \\ &= \sum_{i=1}^n \lambda_i(\theta)\left\langle \left(\bar{\boldsymbol{U}}(\theta)\boldsymbol{u}\right)_i, \left(\bar{\boldsymbol{U}}(\theta)\boldsymbol{v}\right)_i\right\rangle_{L^2} \\ &\overset{\bar{U}(\theta)\text{ orthogonal}}{=} \sum_{i=1}^n \lambda_i(\theta)\left\langle u_i, v_i\right\rangle_{L^2}. \\ &\leq \lambda_\infty \sum_{i=1}^n \left\langle u_i, v_i\right\rangle_{L^2} \\ &= \lambda_\infty \left\langle \boldsymbol{u},\boldsymbol{v}\right\rangle_{[L^2]^n}. \end{aligned} \tag{5.10}$$

Similarly, we can decompose $\bar{\boldsymbol{S}}(\theta)$ as

$$\bar{\boldsymbol{S}}(\theta) = \bar{\boldsymbol{V}}^T(\theta)\bar{\boldsymbol{M}}(\theta)\bar{\boldsymbol{V}}(\theta),$$

with $\bar{\boldsymbol{V}}(\theta)$ an orthogonal matrix and $\bar{\boldsymbol{M}}(\theta)$ a diagonal matrix containing the eigenvalues $\mu_1(\theta), \dots, \mu_n(\theta)$ of $\bar{\boldsymbol{S}}(\theta)$. Like before, define

$$\mu_\infty(\theta) := \max_{1\le i\le n} |\mu_i(\theta)|$$

By following the same steps as in equation (5.10) we find that

$$\left\langle \bar{\boldsymbol{S}}(\theta)\boldsymbol{u}, \boldsymbol{v}\right\rangle_{[L^2]^n} \le \mu_\infty(\theta)\left\langle \boldsymbol{u}, \boldsymbol{v}\right\rangle_{[L^2]^n} \tag{5.11}$$

for all $\boldsymbol{u}, \boldsymbol{v} \in \left[L^2(\Omega)\right]$. Using equations (5.10) and (5.11) we can show that $B^k_{\Delta t}$ is bounded. Indeed, let $\boldsymbol{u}, \boldsymbol{v} \in W^{1,2}(\Omega)$. Then

$$\begin{aligned}
|B^k_{\Delta t}(\mathbf{u},\mathbf{v})| &\le \left|\langle \boldsymbol{u}, \boldsymbol{v}\rangle_{[L^2]^n}\right| + \Delta t\left|\left\langle \bar{\boldsymbol{D}}(\theta)\nabla_x\boldsymbol{u}, \nabla_x\boldsymbol{v}\right\rangle_{[L^2]^n}\right| + \Delta t\left|\left\langle \bar{\mathbf{S}}(\theta)\boldsymbol{u}, \boldsymbol{v}\right\rangle_{[L^2]^n}\right| \\
&\overset{(5.10),(5.11)}{\le} \left|\langle \boldsymbol{u}, \boldsymbol{v}\rangle_{[L^2]^n}\right| + \Delta t\lambda_\infty\left|\langle \nabla_x\boldsymbol{u}, \nabla_x\boldsymbol{v}\rangle_{[L^2]^n}\right| + \Delta t\mu_\infty\left|\langle \boldsymbol{u}, \boldsymbol{v}\rangle_{[L^2]^n}\right| \\
&\overset{\text{Cauchy-Schwartz}}{\le} \|\mathbf{u}\|_{[L^2]^n}\|\mathbf{v}\|_{[L^2]^n} + \Delta t\lambda_\infty\|\nabla\mathbf{u}\|_{[L^2]^n}\|\nabla\mathbf{v}\|_{[L^2]^n} + \Delta t\mu_\infty\|\mathbf{u}\|_{[L^2]^n}\|\mathbf{v}\|_{[L^2]^n} \\
&\le (1 + \Delta t\lambda_\infty(\theta) + \Delta t\mu_\infty(\theta))\,\|\mathbf{u}\|_{[W^{1,2}]^n}\|\mathbf{v}\|_{[W^{1,2}]^n},
\end{aligned}$$

and we see that $B^k_{\Delta t}$ is bounded.

Finally, for coerciveness of $B^k_{\Delta t}$, let $\boldsymbol{u} \in \left[W^{1,2}(\Omega)\right]^n$ .Remember that the eigenvalues of $\bar{\boldsymbol{D}}(\theta)$ are real and positive whereas the eigenvalues of $\bar{\boldsymbol{S}}(\theta)$ are real and negative. Let $\lambda_-(\theta)$ be the smallest eigenvalue of $\bar{\boldsymbol{D}}(\theta)$ and $\mu_+(\theta)$ the largest (but still negative!) eigenvalue of $\bar{\boldsymbol{S}}(\theta)$. Then we see that

$$\begin{aligned}
B^k_{\Delta t}(\boldsymbol{u},\boldsymbol{u}) &= \langle \boldsymbol{u}, \boldsymbol{u}\rangle_{[L^2]^n} + \Delta t\left\langle \bar{\boldsymbol{D}}(\theta)\nabla_x\boldsymbol{u}, \nabla_x\boldsymbol{u}\right\rangle_{[L^2]^n} - \Delta t\left\langle \bar{\mathbf{S}}(\theta)\boldsymbol{u}, \boldsymbol{u}\right\rangle_{[L^2]^n} \\
&\ge \Delta t\lambda_-(\theta)\left\langle \nabla_x\boldsymbol{u}, \nabla_x\boldsymbol{u}\right\rangle_{[L^2]^n} - \Delta t\mu_+(\theta)\left\langle \boldsymbol{u}, \boldsymbol{u}\right\rangle_{[L^2]^n} \\
&\ge \Delta t\min\{\lambda_-(\theta), -\mu_+(\theta)\}\left(\|\nabla\boldsymbol{u}\|^2_{[L^2(\Omega)]^n} + \|\boldsymbol{u}\|^2_{[W^{1,2}(\Omega)]^n}\right) \\
&= \Delta t\underbrace{\min\{\lambda_-(\theta), -\mu_+(\theta)\}}_{:=\beta(\theta)}\|\boldsymbol{u}\|^2_{[W^{1,2}(\Omega)]^n}.
\end{aligned} \tag{5.12}$$

This proves the coercivity of $B^k_{\Delta t}$. Now all the conditions for the linear Lax-Milgram theorem are fulfilled and we are allowed to conclude for each $k \in \{1, \dots, N\}$ the problem $WP^k_{\text{Linear},\Delta t}$ admits a unique solution $\boldsymbol{C}^k$ in $\left[W^{1,2}(\Omega)\right]^n$. □

The general idea is to patch together these solutions to cover the whole time interval $[0, T]$, take the limit $\Delta t \to 0$ and show that the result is a solution to the original weak formulation. Of course there are technicalities that need to be dealt with. As a first step, we will derive so called *a priori* estimates. These estimates - together with the Eberlein–ęmulian theorem (see Subsection 5.1.3) - ensure that the solutions converge to a limit as $\Delta t \to 0$.

5.2.3 *A Priori* Estimates

We obtain *a priori* estimates by inserting the solutions $\boldsymbol{C}^k$ to the problems $WP^k_{\text{Linear},\Delta t}$ as test functions in the weak formulation (5.9). We do the same for the gradients $\nabla\boldsymbol{C}^k$. The following technical prosposition will be needed.

Proposition 2. *Let $a, b \in \mathbb{R}$ and $\rho > 0$. Then*

$$\begin{aligned} 2a(a-b) &= a^2 - b^2 + (a-b)^2, && (5.13)\\ (a+b)^2 &\leq 2a^2 + 2b^2, && (5.14)\\ |ab| &\leq \rho a^2 + \frac{b^2}{4\rho}. && (5.15) \end{aligned}$$

Proof. See Appendix B of Evans [19]. □

Lemma 3. *Let $\{\boldsymbol{C}^k\}_{k=1}^N$ be the sequence of solutions to the problems $WP^k_{Linear,\Delta t}$ obtained under the assumptions H_{Linear}, then for any $p \in \{0, \dots, N\}$ the following inequality holds:*

$$\|\boldsymbol{C}^p\|^2_{[L^2]^n} + \sum_{k=1}^{p} \|\boldsymbol{C}^k - \boldsymbol{C}^{k-1}\|^2_{[L^2]^n} + \Delta t \beta(\theta) \sum_{k=1}^{p} \left(\|\nabla_x \boldsymbol{C}^k\|^2_{[L^2]^n} + \|\boldsymbol{C}^k\|^2_{[L^2]^n} \right) \leq \|\boldsymbol{C}_{\boldsymbol{Initial}}\|^2_{[L^2]^n}.$$

Proof. We insert the solution $\boldsymbol{C}^k \in \left[W^{1,2}(\Omega)\right]^n$ to problem $WP^k_{\text{Linear},\Delta t}$ in the weak formulation (5.9). This yields

$$\left\langle \boldsymbol{C}^k - \boldsymbol{C}^{k-1}, \boldsymbol{C}^k \right\rangle_{[L^2]^n} + \Delta t \left\langle \bar{\boldsymbol{D}}(\theta) \nabla_x \boldsymbol{C}^k, \nabla_x \boldsymbol{C}^k \right\rangle_{[L^2]^n} - \Delta t \left\langle \bar{\boldsymbol{S}}(\theta) \boldsymbol{C}^k, \boldsymbol{C}^k \right\rangle_{[L^2]^n} = 0. \quad (5.16)$$

For the first term in (5.16), we use equation (5.13) from Proposition 2 to find that

$$\left\langle \boldsymbol{C}^k - \boldsymbol{C}^{k-1}, \boldsymbol{C}^k \right\rangle_{[L^2]^n} = \frac{1}{2} \left(\|\boldsymbol{C}^k\|^2_{[L^2]^n} - \|\boldsymbol{C}^{k-1}\|^2_{[L^2]^n} + \|\boldsymbol{C}^k - \boldsymbol{C}^{k-1}\|^2_{[L^2]^n} \right).$$

For the second and third term in (5.16), when proving the coerciveness of $B^k_{\Delta t}$ in (5.12), we saw that

$$\begin{aligned} \Delta t \left\langle \bar{\boldsymbol{D}}(\theta) \nabla \boldsymbol{C}^k, \nabla \boldsymbol{C}^k \right\rangle_{[L^2]^n} - \Delta t \left\langle \bar{\boldsymbol{S}}(\theta) \boldsymbol{C}^k, \boldsymbol{C}^k \right\rangle_{[L^2]^n} &\geq \Delta t \beta(\theta) \left(\|\nabla_x \boldsymbol{C}^k\|^2_{[L^2]^n} + \|\boldsymbol{C}^k\|^2_{[L^2]^n} \right)\\ &\geq \frac{1}{2} \Delta t \beta(\theta) \left(\|\nabla_x \boldsymbol{C}^k\|^2_{[L^2]^n} + \|\boldsymbol{C}^k\|^2_{[L^2]^n} \right), \end{aligned}$$

where β is the coercivity constant. Then it follows that

$$\begin{aligned} &\frac{1}{2} \left(\|\boldsymbol{C}^k\|^2_{[L^2]^n} - \|\boldsymbol{C}^{k-1}\|^2_{[L^2]^n} + \|\boldsymbol{C}^k - \boldsymbol{C}^{k-1}\|^2_{[L^2]^n} \right)\\ &+ \frac{1}{2} \Delta t \beta(\theta) \left(\|\nabla_x \boldsymbol{C}^k\|^2_{[L^2]^n} + \|\boldsymbol{C}^k\|^2_{[L^2]^n} \right)\\ \leq\ & 0. \end{aligned} \quad (5.17)$$

Now let $p \in \{1, \dots, N)$, sum up equation (**??**) from $k = 1$ to p and multiply by 2 to find that

$$\|\boldsymbol{C}^p\|^2_{[L^2]^n} + \sum_{k=1}^{p} \|\boldsymbol{C}^k - \boldsymbol{C}^{k-1}\|^2_{[L^2]^n} + \Delta t \beta(\theta) \sum_{k=1}^{p} \left(\|\nabla_X \boldsymbol{C}^k\|^2_{[L^2]^n} + \|\boldsymbol{C}^k\|^2_{[L^2]^n} \right) \leq \|\boldsymbol{C}^0\|^2_{[L^2]^n}.$$

Because $\boldsymbol{C}^0 = \boldsymbol{C}_{\text{Initial}}$, we see that the lemma has been proven. □

Lemma 4. *Let $\{\boldsymbol{C}^k\}_{k=1}^N$ be the sequence of solutions to the problems $WP^k_{Linear,\Delta t}$ obtained under the assumptions H_{Linear}. There exists a positive constant $\gamma(\theta)$ such that for any $p \in \{0, \dots, N\}$ the following inequality holds:*

$$\frac{1}{\Delta t} \sum_{k=1}^{p} \left\| \boldsymbol{C}^k - \boldsymbol{C}^{k-1} \right\|^2_{[L^2]^n} + \|\nabla_x \boldsymbol{C}^p\|^2_{[L^2]^n} + \sum_{k=1}^{p} \|\nabla_x \boldsymbol{C}^k - \nabla_x \boldsymbol{C}^{k-1}\|^2_{[L^2]^n} \leq \gamma(\theta) \|\boldsymbol{C}_{Initial}\|^2_{[W^{1,2}]^n}.$$

Proof. This time, we are going to test with $\boldsymbol{C}^k - \boldsymbol{C}^{k-1} \in \left[W^{1,2}(\Omega)\right]^n$ in the weak formulation (5.9). We find that

$$\left\|\boldsymbol{C}^k - \boldsymbol{C}^{k-1}\right\|^2_{[L^2]^n} + \Delta t \left\langle \bar{\boldsymbol{D}}(\theta)\nabla_x \boldsymbol{C}^k, \nabla_x \left(\boldsymbol{C}^k - \boldsymbol{C}^{k-1}\right)\right\rangle_{[L^2]^n} - \Delta t \left\langle \bar{\boldsymbol{S}}(\theta)\boldsymbol{C}^k, \boldsymbol{C}^k - \boldsymbol{C}^{k-1}\right\rangle_{[L^2]^n} = 0 \tag{5.18}$$

Using the fact that $\bar{\boldsymbol{D}}(\theta)$ is diagonalizable under the assumptions H, and recalling some of the steps taken in (5.10) to show boundedness of $B^k_{\Delta t}$, we find that

$$\begin{aligned}
& \Delta t \left\langle \bar{\boldsymbol{D}}(\theta)\nabla_x \boldsymbol{C}^k, \nabla_x \left(\boldsymbol{C}^k - \boldsymbol{C}^{k-1}\right)\right\rangle_{[L^2]^n} \\
\overset{(5.10)}{=} \; & \Delta t \sum_{i=1}^n \lambda_i(\theta) \left(\left(\nabla_x \boldsymbol{C}^k\right)_i, \left(\nabla_x \boldsymbol{C}^k - \nabla_x \boldsymbol{C}^{k-1}\right)_i \right)_{L^2} \\
\overset{(5.13)}{=} \; & \frac{\Delta t}{2} \sum_{i=1}^n \lambda_i(\theta) \left\{ \|\nabla \boldsymbol{C}^k_i\|^2_{L^2} - \|\nabla \boldsymbol{C}^{k-1}_i\|^2_{L^2} + \|\nabla_x \boldsymbol{C}^k_i - \nabla_x \boldsymbol{C}^{k-1}_i\|^2_{L^2} \right\} \\
\geq \; & \frac{\Delta t}{2} \lambda_-(\theta) \left\{ \|\nabla_x \boldsymbol{C}^k\|^2_{[L^2]^n} - \|\nabla_x \boldsymbol{C}^{k-1}\|^2_{[L^2]^n} + \|\nabla_x \boldsymbol{C}^k - \nabla_x \boldsymbol{C}^{k-1}\|^2_{[L^2]^n} \right\}
\end{aligned} \tag{5.19}$$

Next, using the fact that $\bar{\boldsymbol{S}}(\theta)$ can be diagonalized and using Proposition 2 with $\rho = \alpha \Delta t \mu_\infty$, where α is some positive constant to be determined later, we find that

$$\begin{aligned}
& \Delta t \left\langle \bar{\boldsymbol{S}}(\theta)\boldsymbol{C}^k, \boldsymbol{C}^k - \boldsymbol{C}^{k-1}\right\rangle_{[L^2]^n} \\
= \; & \Delta t \sum_{i=1}^n \mu_i(\theta) \left(\boldsymbol{C}^k_i, \boldsymbol{C}^k_i - \boldsymbol{C}^{k-1}_i\right)_{L^2} \\
\overset{\text{Cauchy-Schwarz}}{\leq} \; & \Delta t \mu_\infty(\theta) \sum_{i=1}^n \|\boldsymbol{C}^k_i\|_{L^2} \|\boldsymbol{C}^k_i - \boldsymbol{C}^{k-1}_i\|_{L^2} \\
\overset{(5.15) \text{ with } \rho = \alpha \Delta t \mu_\infty}{\leq} \; & \Delta t \mu_\infty(\theta) \sum_{i=1}^n \left\{ \alpha \Delta t \mu_\infty \|\boldsymbol{C}^k_i\|^2_{L^2} + \frac{1}{4\alpha \Delta t \mu_\infty} \|\boldsymbol{C}^k_i - \boldsymbol{C}^{k-1}_i\|^2_{L^2} \right\} \\
= \; & \sum_{i=1}^n \left\{ \alpha \left(\Delta t \mu_\infty(\theta)\right)^2 \|\boldsymbol{C}^k_i\|^2_{L^2} + \frac{1}{4\alpha} \|\boldsymbol{C}^k_i - \boldsymbol{C}^{k-1}_i\|^2_{L^2} \right\} \\
= \; & \alpha \left(\Delta t \mu_\infty(\theta)\right)^2 \|\boldsymbol{C}^k\|^2_{[L^2]^n} + \frac{1}{4\alpha} \|\boldsymbol{C}^k - \boldsymbol{C}^{k-1}\|^2_{[L^2]^n}. \\
\overset{(5.22)}{\leq} \; & \alpha \left(\Delta t \mu_\infty(\theta)\right)^2 \|\boldsymbol{C}_{\text{Initial}}\|^2_{[L^2]^n} + \frac{1}{4\alpha} \|\boldsymbol{C}^k - \boldsymbol{C}^{k-1}\|^2_{[L^2]^n}.
\end{aligned} \tag{5.20}$$

Combining equation (5.18) with equations(5.19) and (5.20) allows us to conclude that

$$\begin{aligned}
& \left\|\boldsymbol{C}^k - \boldsymbol{C}^{k-1}\right\|^2_{[L^2]^n} + \frac{\Delta t}{2} \lambda_-(\theta) \left\{ \|\nabla \boldsymbol{C}^k\|^2_{[L^2]^n} - \|\nabla \boldsymbol{C}^{k-1}\|^2_{[L^2]^n} + \|\nabla \boldsymbol{C}^k - \nabla \boldsymbol{C}^{k-1}\|^2_{[L^2]^n} \right\} \\
\leq \; & \Delta t \left\langle \bar{\boldsymbol{S}}(\theta)\boldsymbol{C}^k, \boldsymbol{C}^k - \boldsymbol{C}^{k-1}\right\rangle_{[L^2]^n} \\
\leq \; & \alpha \left(\Delta t \mu_\infty(\theta)\right)^2 \|\boldsymbol{C}_{\text{Initial}}\|^2_{[L^2]^n} + \frac{1}{4\alpha} \|\boldsymbol{C}^k - \boldsymbol{C}^{k-1}\|^2_{[L^2]^n}.
\end{aligned}$$

Now let $p \in \{1, \ldots, N\}$ and sum up the above equation from 1 to p to find that

$$\begin{aligned}
& \left(1 - \frac{1}{4\alpha}\right) \sum_{k=1}^p \left\|\boldsymbol{C}^k - \boldsymbol{C}^{k-1}\right\|^2_{[L^2]^n} + \frac{\Delta t}{2} \lambda_-(\theta) \left\{ \|\nabla \boldsymbol{C}^p\|^2_{[L^2]^n} + \sum_{k=1}^p \|\nabla \boldsymbol{C}^k - \nabla \boldsymbol{C}^{k-1}\|^2_{[L^2]^n} \right\} \\
\leq \; & \frac{\Delta t}{2} \lambda_-(\theta) \|\nabla \boldsymbol{C}_{\text{Initial}}\|^2_{[L^2]^n} + \alpha \left(\Delta t \mu_\infty(\theta)\right)^2 p \|\boldsymbol{C}_{\text{Initial}}\|^2_{[L^2]^n}.
\end{aligned} \tag{5.21}$$

Next, let α be such that

$$\left(1 - \frac{1}{4\alpha}\right) \;=\; \frac{1}{2}\lambda_-(\theta)$$

and assume α to be positive.[13] Then, finally, we divide both sides of equation (5.21) by $(\Delta t/2)\lambda_-(\theta)$ to find that

$$\begin{aligned}
&\frac{1}{\Delta t}\sum_{k=1}^{p}\left\|\boldsymbol{C}^k - \boldsymbol{C}^{k-1}\right\|^2_{[L^2]^n} + \|\nabla_x \boldsymbol{C}^p\|^2_{[L^2]^n} + \sum_{k=1}^{p}\|\nabla_x \boldsymbol{C}^k - \nabla_x \boldsymbol{C}^{k-1}\|^2_{[L^2]^n} \\
\leq\;& \|\nabla_x \boldsymbol{C}_{\text{Initial}}\|^2_{[L^2]^n} + \frac{2\alpha}{\lambda_-(\theta)}\left(\mu_\infty(\theta)\right)^2 p\Delta t\|\boldsymbol{C}_{\text{Initial}}\|^2_{[L^2]^n} \\
\leq\;& \|\nabla_x \boldsymbol{C}_{\text{Initial}}\|^2_{[L^2]^n} + \frac{2\alpha}{\lambda_-(\theta)}\left(\mu_\infty(\theta)\right)^2 T\|\boldsymbol{C}_{\text{Initial}}\|^2_{[L^2]^n} \\
\leq\;& \underbrace{\max\left\{1, \frac{2\alpha}{\lambda_-(\theta)}\left(\mu_\infty(\theta)\right)^2 T\right\}}_{:=\gamma(\theta)}\|\boldsymbol{C}_{\text{Initial}}\|^2_{[W^{1,2}]^n}.
\end{aligned}$$

□

Corollary 5. *Let $\{\boldsymbol{C}^k\}_{k=1}^N$ be the sequence of solutions to the problems $WP^k_{Linear,\Delta t}$ obtained under the assumptions H_{Linear}. There exists a positive constant $\gamma(\theta)$ such*

$$\begin{aligned}
\|\boldsymbol{C}^k\|^2_{[L^2]^n} &\leq \|\boldsymbol{C}_{Initial}\|^2_{[W^{1,2}]^n}, \\
\|\nabla_x \boldsymbol{C}^k\|^2_{[L^2]^n} &\leq \gamma(\theta)\|\boldsymbol{C}_{Initial}\|^2_{[W^{1,2}]^n},
\end{aligned} \tag{5.22}$$

and hence

$$\|\boldsymbol{C}^k\|^2_{[W^{1,2}]^n} \leq (1+\gamma(\theta))\|\boldsymbol{C}_{Initial}\|^2_{[W^{1,2}]^n}$$

for each $k \in \{0,\ldots,N\}$. Observe that each of the above three upper bounds is independent of the time step Δt.

5.2.4 Time Interpolation of the Discrete-Time Solutions

Now the idea is to patch together the solutions $\{\boldsymbol{C}^k\}_{k=1}^N$ obtained under Lemma 1 to obtain a solution defined on the whole time domain $[0,T]$. We do this in two different ways: we either extend the solutions in a (piecewise) constant manner or in a (piecewise) linear over the time domain. More specifically, we define $\boldsymbol{C}^{\text{Constant}}_{\Delta t}, \boldsymbol{C}^{\text{Linear}}_{\Delta t} : [0,T] \to \left[W^{1,2}(\Omega)\right]^n$ as

$$\boldsymbol{C}^{\text{Constant}}_{\Delta t}(t) := \boldsymbol{C}^k, \quad t \in (t^{k-1}, t^k], \tag{5.23}$$

and

$$\boldsymbol{C}^{\text{Linear}}_{\Delta t}(t) := \boldsymbol{C}^{k-1} + \frac{t - t^{k-1}}{\Delta t}\left(\boldsymbol{C}^k - \boldsymbol{C}^{k-1}\right), \quad t \in (t^{k-1}, t^k]. \tag{5.24}$$

The reasons for choosing these two interpolations will present themselves later. For now, we claim the following.

Lemma 6. *Let $\Delta t > 0$. Then $\boldsymbol{C}^{Constant}_{\Delta t} \in \left[L^2(0,T;W^{1,2}(\Omega)\right]^n$ and*

$$\boldsymbol{C}^{Linear}_{\Delta t} \in \mathcal{W}^n = \left\{\boldsymbol{C} \in \left[L^2\left(0,T;W^{1,2}(\Omega)\right)\right]^n \;\middle|\; \partial_t \boldsymbol{C} \in \left[L^2(0;T;W^{-1,2}(\Omega))\right]^n\right\}.$$

Moreover, the norms $\|\boldsymbol{C}^{Constant}_{\Delta t}\|_{[L^2(0,T;W^{1,2}(\Omega))]^n}$ and $\|\boldsymbol{C}^{Linear}_{\Delta t}\|_{\mathcal{W}^n}$ are bounded uniformly with respect to Δt.

[13] Solving for α gives $\alpha = (4 - 2\lambda_-(\theta))^{-1}$. We have to make sure that the scaling parameter D_{ref} from Subsection 5.1.2 is chosen in such a way that $\lambda_-(\theta) < 1/2$, otherwise α becomes negative.

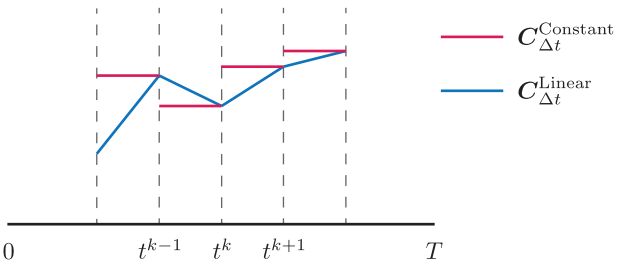

Figure 5.1: Graphical representations of $\boldsymbol{C}_{\Delta t}^{\text{Constant}}$ and $\boldsymbol{C}_{\Delta t}^{\text{Linear}}$.

Proof. Since $\boldsymbol{C}_{\Delta t}^{\text{Constant}}$ is a piecewise constant function (with respect to time), it is measurable. Furthermore, we see that

$$\begin{aligned}
\int_0^T \|\boldsymbol{C}_{\Delta t}^{\text{Constant}}(t)\|_{[W^{1,2}]^n}^2 dt &= \sum_{k=0}^{N} \Delta t \|\boldsymbol{C}^k\|_{[W^{1,2}]^n}^2 dt \\
&\leq T \max_{k=0,\dots,N} \|\boldsymbol{C}^k\|_{[W^{1,2}]^n}^2 \\
&\overset{\text{Corrollary 5}}{\leq} T(1+\gamma(\theta))\|\boldsymbol{C}^{\text{Initial}}\|_{[W^{1,2}]^n}^2 \\
&< \infty.
\end{aligned}$$

We conclude that $\boldsymbol{C}_{\Delta t}^{\text{Constant}} \in \left[L^2(0,T;W^{1,2}(\Omega)\right]^n$. Moreover, the second to last inequality shows us that the norm $\|\boldsymbol{C}_{\Delta t}^{\text{Constant}}\|_{[L^2(0,T;W^{1,2}(\Omega)]^n}^2$ is bounded uniformly with respect to Δt.

To show that $\boldsymbol{C}_{\Delta t}^{\text{Linear}}$ is measurable, let $p \in \mathbb{N}$ and define

$$t_j^k := t^{k-1} + \frac{j}{p}\Delta t.$$

With this notation, we can define a piecewise constant (with respect to t) function $\boldsymbol{C}_{\Delta t,p} : [0,T] \to \left[W^{1,2}(\Omega)\right]^n$ as

$$\boldsymbol{C}_{\Delta t,p}(t) := \boldsymbol{C}^{k-1} + \frac{t_j^k - t^{k-1}}{\Delta t}\left(\boldsymbol{C}^k - \boldsymbol{C}^{k-1}\right), \quad t \in (t_{j-1}^k, t_j^k].$$

Like $\boldsymbol{C}_{\Delta t}^{\text{Linear}}$, each $\boldsymbol{C}_{\Delta t,p}$ is measurable. Now, for $t \in (t_{j-1}^k, t_j^k]$, we see that

$$\begin{aligned}
\|\boldsymbol{C}_{\Delta t}^{\text{Linear}}(t) - \boldsymbol{C}_{\Delta t,p}(t)\|_{[W^{1,2}]^n} &= \frac{t - t_j^k}{\Delta t}\|\boldsymbol{C}^k - \boldsymbol{C}^{k-1}\|_{[W^{1,2}]^n} \\
&\leq \frac{1}{p}\|\boldsymbol{C}^k - \boldsymbol{C}^{k-1}\|_{[W^{1,2}]^n}.
\end{aligned}$$

The upper bound clearly goes to zero as $p \to \infty$. That means $\boldsymbol{C}_{\Delta t}^{\text{Linear}}$ is the pointwise limit of sequence of measurable functions and from measure theory we know that $\boldsymbol{C}_{\Delta t}^{\text{Linear}}$ is itself

measurable in that case. Because

$$\begin{aligned}
\int_0^T \|\boldsymbol{C}_{\Delta t}^{\text{Linear}}\|_{[W^{1,2}]^n}^2 dt &= \sum_{k=1}^{N} \int_{t^{k-1}}^{t^k} \left\| \boldsymbol{C}^{k-1} + \frac{t-t^{k-1}}{\Delta t}\left(\boldsymbol{C}^k - \boldsymbol{C}^{k-1}\right)\right\|_{[W^{1,2}]^n}^2 dt \\
&= \sum_{k=1}^{N} \int_{t^{k-1}}^{t^k} \left\| \frac{t-t^{k-1}}{\Delta t}\boldsymbol{C}^k + \frac{t^k-t}{\Delta t}\boldsymbol{C}^{k-1}\right\|_{[W^{1,2}]^n}^2 dt \\
&\overset{(5.14)}{\leq} 2\sum_{k=1}^{N}\left[\int_{t^{k-1}}^{t^k}\left(\frac{t-t^{k-1}}{\Delta t}\right)^2 \left\|\boldsymbol{C}^k\right\|_{[W^{1,2}]^n}^2 + \left(\frac{t^k-t}{\Delta t}\right)^2 \left\|\boldsymbol{C}^{k-1}\right\|_{[W^{1,2}]^n}^2 dt\right] \\
&= \frac{2\Delta t}{3}\sum_{k=1}^{N}\left[\left\|\boldsymbol{C}^k\right\|_{[W^{1,2}]^n}^2 + \left\|\boldsymbol{C}^{k-1}\right\|_{[W^{1,2}]^n}^2\right] \\
&\leq \frac{4\Delta t}{3}\sum_{k=0}^{N}\left\|\boldsymbol{C}^k\right\|_{[W^{1,2}]^n}^2 \\
&< \infty,
\end{aligned}$$

we can conclude $\boldsymbol{C}_{\Delta t}^{\text{Linear}} \in \left[L^2\left(0,T;W^{1,2}(\Omega)\right)\right]^n$. Moreover, using Corollary 5 we state that $\sum_{k=0}^{N}\left\|\boldsymbol{C}^k\right\|_{[W^{1,2}]^n}^2$ and hence $\|\boldsymbol{C}_{\Delta t}^{\text{Linear}}\|_{[L^2(0,T;W^{1,2})]^n}^2$ is bounded uniformly with respect to Δt.

The next step is to show that $\partial_t \boldsymbol{C}_{\Delta t}^{\text{Linear}} \in \left[L^2(0;T;W^{-1,2}(\Omega))\right]^n$. To this end, note that the piecewise linear form of $\boldsymbol{C}_{\Delta t}^{\text{Linear}}$ suggests that it has a strong derivative (defined for a.e. $t \in (0,T]$). Indeed, let $t \in (t^{k-1},t^k)$ and suppose h is small enough for $t+h$ to be in (t^{k-1},t^k) as well. Then

$$\frac{\boldsymbol{C}_{\Delta t}^{\text{Linear}}(t+h) - \boldsymbol{C}_{\Delta t}^{\text{Linear}}(t)}{h} = \frac{\boldsymbol{C}^k - \boldsymbol{C}^{k-1}}{\Delta t}$$

and hence

$$\lim_{h\to 0}\left\|\frac{\boldsymbol{C}_{\Delta t}^{\text{Linear}}(t+h) - \boldsymbol{C}_{\Delta t}^{\text{Linear}}(t)}{h} - \frac{\boldsymbol{C}^k - \boldsymbol{C}^{k-1}}{\Delta t}\right\|_{[L^2]^n} = 0.$$

We conclude that $\partial_t \boldsymbol{C}_{\Delta t}^{\text{Linear}} = \left(\boldsymbol{C}^k - \boldsymbol{C}^{k-1}\right)/\Delta t$ for $t \in (t^{k-1},t^k)$. Moreover,

$$\begin{aligned}
\int_0^T \|\partial_t \boldsymbol{C}_{\Delta t}^{\text{Linear}}(t)\|_{[L^2]^n}^2 dt &= \sum_{k=1}^{N}\int_{t^{k-1}}^{t^k}\left\|\frac{\boldsymbol{C}^k - \boldsymbol{C}^{k-1}}{\Delta t}\right\|_{[L^2]^n}^2 dt \\
&= \sum_{k=1}^{N}\left\{\left\|\boldsymbol{C}^k - \boldsymbol{C}^{k-1}\right\|_{[L^2]^n}^2 \int_{t^{k-1}}^{t^k}\frac{1}{(\Delta t)^2} dt\right\} \\
&= \frac{1}{\Delta t}\sum_{k=1}^{N}\left\|\boldsymbol{C}^k - \boldsymbol{C}^{k-1}\right\|_{[L^2]^n}^2 \\
&< \infty,
\end{aligned}$$

and we can say that $\boldsymbol{C}_{\Delta t}^{\text{Linear}} \in \mathcal{W}^n$. From Lemma 4 we see that $\frac{1}{\Delta t}\sum_{k=1}^{N}\left\|\boldsymbol{C}^k - \boldsymbol{C}^{k-1}\right\|_{[L^2]^n}^2$ and hence $\|\partial_t \boldsymbol{C}_{\Delta t}^{\text{Linear}}(t)\|_{[L^2(0,T;L^2)]^n}^2$ is bounded uniformly with respect to Δt as well. Note that we have in fact shown that $\partial_t \boldsymbol{C}_{\Delta t}^{\text{Linear}} \in L^2(\left[0,T;L^2(\Omega)\right)]^n$. Since $L^2(\Omega)$ is continuously embedded in $W^{-1,2}(\Omega)$ (see Subsection 5.1.3), it follows that $\partial_t \boldsymbol{C}_{\Delta t}^{\text{Linear}} \in \left[L^2(0,T;W^{-1,2}(\Omega))\right]^n$ and the norm in the latter space is then bounded uniformly with respect to Δt as well. This concludes the proof that $\boldsymbol{C}_{\Delta t}^{\text{Linear}} \in \mathcal{W}^n$ and that

$$\|\boldsymbol{C}_{\Delta t}^{\text{Linear}}\|_{\mathcal{W}^n} = \left(\|\boldsymbol{C}_{\Delta t}^{\text{Linear}}\|_{[L^2(0,T;W^{1,2})]^n}^2 + \|\partial_t \boldsymbol{C}_{\Delta t}^{\text{Linear}}\|_{[L^2(0,T;W^{-1,2})]^n}^2\right)^{1/2}$$

is bounded uniformly with respect to Δt. □

Having all these norm-estimates ready, it is finally time to pass to the limit $\Delta t \to 0$.

5.2.5 Passing to the Limit

Let $\{(\Delta t)_j\}$ be a sequence of time steps such that $(\Delta t)_j \to 0$ as $j \to \infty$. In the previous section we have seen that $\boldsymbol{C}_{\Delta t}^{\text{Linear}}$ and $\partial_t \boldsymbol{C}_{\Delta t}^{\text{Linear}}$ are bounded uniformly with respect to Δt in $\left[L^2(0,T;W^{1,2}(\Omega)\right]^n$ and $\left[L^2(0,T;W^{-1,2}(\Omega)\right]^n$ respectively. Hence the sequences $\{\boldsymbol{C}_{(\Delta t)_j}^{\text{Linear}}\}$ and $\{\partial_t \boldsymbol{C}_{(\Delta t)_j}^{\text{Linear}}\}$ remain bounded in their respective spaces. Using the Eberlein–ɕmulian theorem we can then extract weakly converging subsequences. That is, we can deduce the existence of a $\boldsymbol{C}^{\text{Linear}} \in \left[L^2(0,T;W^{1,2}(\Omega)\right]^n$, a $\partial_t \boldsymbol{C}^{\text{Linear}} \in \left[L^2(0,T;W^{-1,2}(\Omega)\right]^n$ and a subsequence of time steps $\{(\Delta t)_{j_k}\}$ such that

$$\left\{\boldsymbol{C}_{(\Delta t)_{j_k}}^{\text{Linear}}\right\} \rightharpoonup \boldsymbol{C}^{\text{Linear}} \text{ weakly in } \left[L^2(0,T;W^{1,2}(\Omega)\right]^n, \tag{5.25}$$

$$\left\{\partial_t \boldsymbol{C}_{(\Delta t)_{j_k}}^{\text{Linear}}\right\} \rightharpoonup \partial_t \boldsymbol{C}^{\text{Linear}} \text{ weakly in } \left[L^2(0,T;W^{-1,2}(\Omega)\right]^n. \tag{5.26}$$

From now on we will no longer specify the particular subsequence along which the weak convergence takes place. We will simply say that $\boldsymbol{C}_{\Delta t}^{\text{Linear}}$ converges weakly to $\boldsymbol{C}^{\text{Linear}}$ as $\Delta t \to 0$ and similarly for $\partial_t \boldsymbol{C}_{\Delta t}^{\text{Linear}}$. Note that the theorem stated under *Limits and Derivatives* in the preliminaries (Section 5.1.3) allows us to conclude that $\partial_t \boldsymbol{C}^{\text{Linear}}$ really is the weak derivative of $\boldsymbol{C}^{\text{Linear}}$ and hence $\boldsymbol{C}^{\text{Linear}} \in \mathcal{W}^n$.

For the sequence $\{\boldsymbol{C}_{(\Delta t)_j}^{\text{Constant}}\}$ we can use similar arguments to deduce the existence of a $\boldsymbol{C}^{\text{Constant}} \in \left[L^2(0,T;W^{1,2}(\Omega))\right]^n$ such that

$$\boldsymbol{C}_{\Delta t}^{\text{Constant}} \rightharpoonup \boldsymbol{C}^{\text{Constant}} \text{ weakly in } \left[L^2(0,T;W^{1,2}(\Omega))\right]^n. \tag{5.27}$$

Lemma 7. *The weak limits of $\boldsymbol{C}_{\Delta t}^{Constant}$ and $\boldsymbol{C}_{\Delta t}^{Linear}$ coincide. That is, $\boldsymbol{C}^{Constant} = \boldsymbol{C}^{Linear}$.*

Proof. The idea of the proof is to show that

$$\left\langle \boldsymbol{C}^{\text{Linear}} - \boldsymbol{C}^{\text{Constant}}, \boldsymbol{\psi} \right\rangle_{[L^2(0,T;L^2)]^n} = 0$$

for all $\boldsymbol{\psi} \in \left[L^2(0,T;L^2(\Omega))\right]^n$. Then it would follow that $\boldsymbol{C}^{\text{Linear}} = \boldsymbol{C}^{\text{Constant}}$ in $\left[L^2(0,T;L^2(\Omega))\right]^n$. To show that this is the case, we use the triangle inequality and find that

$$\begin{aligned} \left|\left\langle \boldsymbol{C}^{\text{Linear}} - \boldsymbol{C}^{\text{Constant}}, \boldsymbol{\psi} \right\rangle_{[L^2(0,T;L^2)]^n}\right| &\leq \left|\left\langle \boldsymbol{C}^{\text{Linear}} - \boldsymbol{C}_{\Delta t}^{\text{Constant}}, \boldsymbol{\psi} \right\rangle_{[L^2(0,T;L^2)]^n}\right| \\ &+ \left|\left\langle \boldsymbol{C}_{\Delta t}^{\text{Constant}} - \boldsymbol{C}^{\text{Constant}}, \boldsymbol{\psi} \right\rangle_{[L^2(0,T;L^2)]^n}\right|. \end{aligned} \tag{5.28}$$

For the second term on the right-hand side of equation (5.28), we observe that for any given $\boldsymbol{\psi} \in \left[L^2(0,T;L^2(\Omega))\right]^n$, the mapping on $\left[L^2(0,T;W^{1,2}(\Omega))\right]^n \to \mathbb{R}$ defined by

$$\boldsymbol{\phi} \mapsto \langle \boldsymbol{\phi}, \boldsymbol{\psi} \rangle_{[L^2(0,T;L^2)]^n}$$

is linear and bounded. Hence $\boldsymbol{\phi} \mapsto \langle \boldsymbol{\phi}, \boldsymbol{\psi} \rangle_{[L^2(0,T;L^2)]^n}$ is in the dual space of $\left[L^2(0,T;W^{1,2}(\Omega))\right]^n$. Since $\boldsymbol{C}_{\Delta t}^{\text{Constant}}$ converges weakly to $\boldsymbol{C}^{\text{Constant}}$ in $\left[L^2(0,T;W^{1,2}(\Omega))\right]^n$ (see (5.27)) it follows directly that

$$\left\langle \boldsymbol{C}_{\Delta t}^{\text{Constant}} - \boldsymbol{C}^{\text{Constant}}, \boldsymbol{\psi} \right\rangle_{[L^2(0,T;L^2)]^n} \to 0, \tag{5.29}$$

as $\Delta t \to 0$. For the first term on the right-hand side of equation (5.28), we want to show that $\boldsymbol{C}_{\Delta t}^{\text{Constant}}$ converges weakly to $\boldsymbol{C}^{\text{Linear}}$ as $\Delta t \to 0$. To this end, we use the triangle inequality again to find that

$$\begin{aligned} \left\|\boldsymbol{C}_{\Delta t}^{\text{Constant}} - \boldsymbol{C}^{\text{Linear}}\right\|_{[L^2(0,T;L^2)]^n} &\leq \left\|\boldsymbol{C}_{\Delta t}^{\text{Constant}} - \boldsymbol{C}_{\Delta t}^{\text{Linear}}\right\|_{[L^2(0,T;L^2)]^n} \\ &+ \left\|\boldsymbol{C}_{\Delta t}^{\text{Linear}} - \boldsymbol{C}^{\text{Linear}}\right\|_{[L^2(0,T;L^2)]^n}. \end{aligned} \tag{5.30}$$

Using the fact that $\mathcal{W}$ is compactly embedded in $\left[L^2(0,T;L^2(\Omega))\right]^n$ (see Section 5.1.3) we can say that $\boldsymbol{C}_{\Delta t}^{\text{Linear}}$ converges strongly to $\boldsymbol{C}^{\text{Linear}}$ in $\left[L^2(0,T;L^2(\Omega))\right]^n$. Hence the second term on the right-hand side of (5.30) goes to zero as $\Delta t \to 0$. For the first term on the right-hand side of (5.30), let $t \in (t^{k-1}, t^k]$. Then

$$\begin{aligned}\boldsymbol{C}_{\Delta t}^{\text{Constant}}(t) - \boldsymbol{C}_{\Delta t}^{\text{Linear}}(t) &= \boldsymbol{C}^k - \left(\boldsymbol{C}^{k-1} + \frac{t-t^{k-1}}{\Delta t}\left(\boldsymbol{C}^k - \boldsymbol{C}^{k-1}\right)\right)\\ &= \frac{t^k - t}{\Delta t}\left(\boldsymbol{C}^k - \boldsymbol{C}^{k-1}\right).\end{aligned}$$

It follows that

$$\begin{aligned}\left\|\boldsymbol{C}_{\Delta t}^{\text{Constant}} - \boldsymbol{C}_{\Delta t}^{\text{Linear}}\right\|^2_{[L^2(0,T;L^2)]^n} &= \int_0^T \left\|\boldsymbol{C}_{\Delta t}^{\text{Constant}}(t) - \boldsymbol{C}_{\Delta t}^{\text{Linear}}(t)\right\|^2_{[L^2]^n} dt\\ &= \sum_{k=1}^{N}\int_{t^{k-1}}^{t^k}\left(\frac{t^k - t}{\Delta t}\right)^2 \left\|\boldsymbol{C}^k - \boldsymbol{C}^{k-1}\right\|^2_{[L^2]^n} dt\\ &= \frac{\Delta t}{3}\sum_{k=1}^{N}\left\|\boldsymbol{C}^k - \boldsymbol{C}^{k-1}\right\|^2_{[L^2]^n}\\ &\overset{(3)}{<} \Delta t \|\boldsymbol{C}_{\text{Initial}}\|^2_{[L^2]^n}.\end{aligned}$$

We conclude that first term on the right-hand side of (5.30) also goes to zero as $\Delta t \to 0$ and hence $\boldsymbol{C}_{\Delta t}^{\text{Constant}}$ converges strongly to $\boldsymbol{C}^{\text{Linear}}$ in $\left[L^2(0,T;L^2(\Omega))\right]^n$. Since strong convergence implies weak convergence, it follows that

$$\left\langle \boldsymbol{C}^{\text{Linear}} - \boldsymbol{C}_{\Delta t}^{\text{Constant}}, \boldsymbol{\psi}\right\rangle_{[L^2(0,T;L^2)]^n} \to 0$$

as $\Delta t \to 0$. Now equation (5.28) allows us to conclude that $\boldsymbol{C}^{\text{Linear}} = \boldsymbol{C}^{\text{Constant}}$ in $\left[L^2(0,T;L^2(\Omega))\right]^n$. □

Corollary 8. *There exists a $\boldsymbol{C} \in \mathcal{W}^n$ such that*

$$\begin{aligned}\boldsymbol{C}_{\Delta t}^{Constant} &\rightharpoonup \boldsymbol{C}, \text{ weakly in } \left[L^2(0,T;W^{1,2}(\Omega))\right]^n,\\ \boldsymbol{C}_{\Delta t}^{Constant} &\to \boldsymbol{C}, \text{ strongly in } \left[L^2(0,T;L^2(\Omega))\right]^n\\ \boldsymbol{C}_{\Delta t}^{Linear} &\rightharpoonup \boldsymbol{C}, \text{ weakly in } \left[L^2(0,T;W^{1,2}(\Omega))\right]^n,\\ \boldsymbol{C}_{\Delta t}^{Linear} &\to \boldsymbol{C}, \text{ strongly in } \left[L^2(0,T;L^2(\Omega))\right]^n,\\ \partial_t \boldsymbol{C}_{\Delta t}^{Linear} &\rightharpoonup \partial_t \boldsymbol{C}, \text{ weakly in } \left[L^2(0,T;W^{-1,2}(\Omega))\right]^n,\end{aligned}$$

as $\Delta t \to 0$. Like before, the convergence is to be understood as taking place along a particular sequence of time steps converging to zero.

What remains to be shown is that the limit $\boldsymbol{C}$ is in fact a (weak) solution to the problem $WP_{\textbf{Linear}}$.

Lemma 9. *The limit $\boldsymbol{C} \in \mathcal{W}^n$ obtained under the assumptions H_{Linear} is a (weak) solution to the problem $WP_{\textbf{Linear}}$.*

Proof. Let $\boldsymbol{\varphi} \in \left[L^2(0,T;W^{1,2}(\Omega)\right]^n$. From the problem $WP^k_{\text{Linear},\Delta t}$ we know that

$$\left\langle \boldsymbol{C}^k - \boldsymbol{C}^{k-1}, \boldsymbol{\varphi}(t)\right\rangle_{[L^2]^n} = -\Delta t \left\langle \bar{\boldsymbol{D}}\nabla_x \boldsymbol{C}^k, \nabla\boldsymbol{\varphi}(t)\right\rangle_{[L^2]^n} + \Delta t \left\langle \bar{\boldsymbol{S}}\boldsymbol{C}^k, \boldsymbol{\varphi}(t)\right\rangle_{[L^2]^n}$$

for a.e. $t \in (t^{k-1}, t^k]$ (wherever $\boldsymbol{\varphi}$ is defined). Also remember from (5.23) and (5.24) that $\boldsymbol{C}_{\Delta t}^{\text{Constant}}(t) = \boldsymbol{C}^k$ and $\partial_t \boldsymbol{C}_{\Delta t}^{\text{Linear}}(t) = \left(\boldsymbol{C}^k - \boldsymbol{C}^{k-1}\right)/\Delta t$ on $t \in (t^{k-1}, t^k]$. Hence the above equality can be rewritten as

$$\left\langle \partial_t \boldsymbol{C}_{\Delta t}^{\text{Linear}}(t), \boldsymbol{\varphi}(t) \right\rangle_{[L^2]^n} = -\left\langle \bar{\boldsymbol{D}}(\theta) \nabla_x \boldsymbol{C}_{\Delta t}^{\text{Constant}}(t), \nabla \boldsymbol{\varphi}(t) \right\rangle_{[L^2]^n} + \left\langle \bar{\boldsymbol{S}}(\theta) \boldsymbol{C}_{\Delta t}^{\text{Constant}}(t), \boldsymbol{\varphi}(t) \right\rangle_{[L^2]^n}.$$

Integrating the whole equation in time yields

$$\begin{aligned} \int_0^T \left\langle \partial_t \boldsymbol{C}_{\Delta t}^{\text{Linear}}(t), \boldsymbol{\varphi}(t) \right\rangle_{[L^2]^n} dt &= -\int_0^T \left\langle \bar{\boldsymbol{D}}(\theta) \nabla_x \boldsymbol{C}_{\Delta t}^{\text{Constant}}(t), \nabla \boldsymbol{\varphi}(t) \right\rangle_{[L^2]^n} dt \\ &\quad + \int_0^T \left\langle \bar{\boldsymbol{S}}(\theta) \boldsymbol{C}_{\Delta t}^{\text{Constant}}(t), \boldsymbol{\varphi}(t) \right\rangle_{[L^2]^n} dt. \end{aligned} \tag{5.31}$$

Using the weak convergence results summarized in Corollary (8) we know that

$$\begin{aligned} \int_0^T \left\langle \partial_t \boldsymbol{C}_{\Delta t}^{\text{Linear}}(t), \boldsymbol{\varphi}(t) \right\rangle_{[L^2]^n} dt &= \int_0^T \left\langle \partial_t \boldsymbol{C}_{\Delta t}^{\text{Linear}}(t), \boldsymbol{\varphi}(t) \right\rangle_{[W^{-1,2}]^n, [W^{1,2}]^n} dt \\ &\to \int_0^T \left\langle \partial_t \boldsymbol{C}(t), \boldsymbol{\varphi}(t) \right\rangle_{[W^{-1,2}]^n, [W^{1,2}]^n} dt. \end{aligned} \tag{5.32}$$

Furthermore, because

$$\int_0^T \left\langle \nabla_x \boldsymbol{C}_{\Delta t}^{\text{Constant}}(t), \nabla \boldsymbol{\varphi}(t) \right\rangle_{[L^2]^n} dt \to \int_0^T \left\langle \bar{\boldsymbol{D}} \nabla_x \boldsymbol{C}(t), \nabla \boldsymbol{\varphi}(t) \right\rangle_{[L^2]^n} dt,$$

it follows that

$$\begin{aligned} \int_0^T \left\langle \bar{\boldsymbol{D}}(\theta) \nabla_x \boldsymbol{C}_{\Delta t}^{\text{Constant}}(t), \nabla \boldsymbol{\varphi}(t) \right\rangle_{[L^2]^n} dt &= \sum_{i=1}^n \int_0^T \lambda_i(\theta) \left\langle \left(\nabla_x \boldsymbol{C}_{\Delta t}^{\text{Constant}}\right)_i(t), \nabla_x \boldsymbol{\varphi}_i(t) \right\rangle_{L^2} dt \\ &\to \sum_{i=1}^n \int_0^T \lambda_i(\theta) \left\langle \left(\nabla_x \boldsymbol{C}\right)_i(t), \nabla_x \boldsymbol{\varphi}_i(t) \right\rangle_{L^2} dt \\ &= \int_0^T \left\langle \bar{\boldsymbol{D}}(\theta) \nabla_x \boldsymbol{C}(t), \nabla_x \boldsymbol{\varphi}(t) \right\rangle_{[L^2]^n} dt. \end{aligned} \tag{5.33}$$

In a similar fashion,

$$\begin{aligned} \int_0^T \left(\bar{\boldsymbol{S}}(\theta) \boldsymbol{C}_{\Delta t}^{\text{Constant}}(t), \boldsymbol{\varphi}(t)\right)_{[L^2]^n} dt &= \sum_{i=1}^n \int_0^T \mu_i(\theta) \left\langle \left(\boldsymbol{C}_{\Delta t}^{\text{Constant}}\right)_i(t), \boldsymbol{\varphi}_i(t) \right\rangle_{L^2} dt \\ &\to \sum_{i=1}^n \int_0^T \mu_i(\theta) \left\langle \boldsymbol{C}_i(t), \boldsymbol{\varphi}_i(t) \right\rangle_{L^2} dt \\ &= \int_0^T \left\langle \bar{\boldsymbol{S}}(\theta) \boldsymbol{C}(t), \boldsymbol{\varphi}(t) \right\rangle_{[L^2]^n} dt. \end{aligned} \tag{5.34}$$

Putting (5.31), (5.32), (5.33) and (5.34) together we find that $\boldsymbol{C}$ satisfies

$$\int_0^T \left\langle \partial_t \boldsymbol{C}(t), \boldsymbol{\varphi}(t) \right\rangle_{[W^{-1,2}]^n, [W^{1,2}]^n} + \left\langle \bar{\boldsymbol{D}}(\theta) \nabla \boldsymbol{C}(t), \nabla_x \boldsymbol{\varphi}(t) \right\rangle_{[L^2]^n} - \left\langle \bar{\boldsymbol{S}} \boldsymbol{C}(t), \boldsymbol{\varphi}(t) \right\rangle_{[L^2]^n} dt = 0. \tag{5.35}$$

Because this is true for all $\boldsymbol{\varphi} \in \left[L^2(0,T; W^{1,2}(\Omega)\right]^n$, we can conclude that for all $\boldsymbol{\varphi} \in \left[W^{1,2}(\Omega)\right]^n$ the equation

$$\left\langle \partial_t \boldsymbol{C}(t), \boldsymbol{\varphi} \right\rangle_{[W^{-1,2}]^n, [W^{1,2}]^n} + \left\langle \bar{\boldsymbol{D}}(\theta) \nabla_x \boldsymbol{C}(t), \nabla_x \boldsymbol{\varphi} \right\rangle_{[L^2]^n} - \left\langle \bar{\boldsymbol{S}} \boldsymbol{C}(t), \boldsymbol{\varphi} \right\rangle_{[L^2]^n} \tag{5.36}$$

holds for a.e. $t \in [0,T]$. Indeed, suppose there exists a $\boldsymbol{\varphi} \in \left[W^{1,2}(\Omega)\right]^n$ and subset $S \subset [0,T]$ with non-zero measure on which (5.36) is non-zero. Let χ_S be the characteristic function of S and use $\boldsymbol{\psi}(t) := \chi_S \boldsymbol{\varphi} \in \left[L^2(0,T; W^{1,2}(\Omega)\right]^n$ as a test function in (5.36) to arrive at a contradiction. □

Lemma 10. *The limit $\boldsymbol{C}$ obtained under the assumptions H_{Linear} is the unique (weak) solution to problem WP_{Linear}.*

Proof. To show uniqueness, we start by substituting the solution $\boldsymbol{C}$ as a test function in the weak formulation WP_{Linear} and integrating from 0 to T to find that

$$\int_0^T \langle \partial_t \boldsymbol{C}(t), \boldsymbol{C}(t) \rangle_{[W^{-1,2}]^n,[W^{1,2}]^n} \, dt + \int_0^T \left\langle \bar{\boldsymbol{D}}(\theta) \nabla_x \boldsymbol{C}(t), \nabla_x \boldsymbol{C}(t) \right\rangle_{[L^2]^n} = \left\langle \bar{\boldsymbol{S}}(\theta) \boldsymbol{C}(t), \boldsymbol{C}(t) \right\rangle_{[L^2]^n} dt. \tag{5.37}$$

Using partial integration, see equation (5.5), it follows that

$$\begin{aligned} \int_0^T \langle \partial_t \boldsymbol{C}(t), \boldsymbol{C}(t) \rangle_{[W^{-1,2}]^n,[W^{1,2}]^n} \, dt &= \frac{1}{2} \int_0^T \left[\frac{d}{dt} \langle \boldsymbol{C}(t), \boldsymbol{C}(t) \rangle_{[L^2]^n} \right] dt \\ &= \frac{1}{2} \|\boldsymbol{C}(T)\|^2_{[L^2]^n} - \frac{1}{2} \|\boldsymbol{C}(0)\|^2_{[L^2]^n} \\ &= \frac{1}{2} \|\boldsymbol{C}(T)\|^2_{[L^2]^n} - \frac{1}{2} \|\boldsymbol{C}_{\text{Initial}}\|^2_{[L^2]^n} . \end{aligned}$$

By repeating arguments used to show that the functionals $B^k_{\Delta t}$ were coercive (see Lemma 1), we can show that

$$\int_0^T \left\langle \bar{\boldsymbol{D}}(\theta) \nabla_x \boldsymbol{C}(t), \nabla_x \boldsymbol{C}(t) \right\rangle_{[L^2]^n} - \left\langle \bar{\boldsymbol{S}}(\theta) \boldsymbol{C}(t), \boldsymbol{C}(t) \right\rangle_{[L^2]^n} dt \geq \int_0^T \beta(\theta) \|\mathbf{C}(t)\|^2_{[W^{1,2}]^n} dt.$$

Substituting these last two expressions into equation (5.37) reveal that

$$\frac{1}{2} \|\boldsymbol{C}(T)\|^2_{[L^2(\Omega)]^n} + \beta(\theta) \int_0^T \|\mathbf{C}(t)\|^2_{[W^{1,2}(\Omega)]^n} dt \leq \frac{1}{2} \|\boldsymbol{C}_{\text{Initial}}\|^2_{[L^2(\Omega)]^n}. \tag{5.38}$$

Now suppose suppose $\tilde{\boldsymbol{C}}$ is another weak solution. Using linearity of the problem, it follows that $\boldsymbol{C} - \tilde{\boldsymbol{C}}$ is a solution as well. Substituting this 'new solution' into equation (5.38) yields

$$\begin{aligned} \frac{1}{2} \left\| \boldsymbol{C}(T) - \tilde{\boldsymbol{C}}(T) \right\|^2_{[L^2]^n} + \beta(\theta) \int_0^T \|\boldsymbol{C}(t) - \tilde{\boldsymbol{C}}(t)\|^2_{[W^{1,2}]^n} dt &\leq \frac{1}{2} \|\boldsymbol{C}(0) - \tilde{\boldsymbol{C}}(0)\|^2_{[L^2]^n} \\ &= \frac{1}{2} \|\boldsymbol{C}_{\text{Initial}} - \boldsymbol{C}_{\text{Initial}}\|^2_{[L^2]^n} \\ &= 0. \end{aligned}$$

Because both terms on the left-hand side are positive, it follows that

$$\int_0^T \|\boldsymbol{C}(t) - \tilde{\boldsymbol{C}}(t)\|^2_{[W^{1,2}]^n} dt = \|\boldsymbol{C} - \tilde{\boldsymbol{C}}\|^2_{[L^2(0,T;W^{1,2})]^n} dt = 0.$$

We conclude that $\tilde{\boldsymbol{C}} = \boldsymbol{C}$. □

The results obtained in this section are summarized in the following theorem.

Theorem 11. *Under the assumptions H_{Linear} the problem WP_{Linear} has a unique solution.*

5.3 Non-Linear Case: Single Component

5.3.1 Problem Formulation and Assumptions

With the results for the linear case in mind, it is time to go back to the non-linear problem. But not the full one yet. First we show well-posedness of a non-linear diffusion reaction equation in case there is just one component. The corresponding weak problem can be formulated as follows:

(Problem WP_1) Find $C \in \mathcal{W}$ such that $C(0) = C_{\text{Initial}}$ and for all $\varphi \in W^{1,2}(\Omega)$ the equality

$$\langle \partial_t C(t), \varphi \rangle_{W^{-1,2}, W^{1,2}} = -\langle D(C(t), \theta) \nabla_x C(t), \nabla_x \varphi \rangle_{L^2} + \langle S(C(t), \theta), \varphi \rangle_{L^2} \tag{5.39}$$

holds for a.e. $t \in [0, T]$.

(H_1) The diffusion coefficient D depends continuously on C and on θ. Moreover, there exists constants $m(\theta)$ and $M(\theta)$ such that

$$m(\theta) \leq D(x, \theta) \leq M(\theta)$$

for all $x \in \mathbb{R}$.

(H_2) The chemical reaction function S is Lipschitz continuous with positive Lipschitz constant $\mathcal{L}$ depending on temperature θ only. That is,

$$|S(x) - S(y)| \leq \mathcal{L} |x - y|$$

for all $x, y \in \mathbb{R}$. Moreover, we assume that $S(x) = 0$ for $x \leq 0$ (no concentrations, no chemical reactions).

(H_3) The initial condition is weakly differentiable in space, i.e. $C_{\text{Initial}} \in W^{1,2}(\Omega)$.

From now on we fix the temperature θ and surpress it from the notations. It is nothing but a parameter.

Furthermore, note that assumption H_2 leads to the growth condition

$$|S(x)| = |S(x) - S(0)| \leq \mathcal{L} |x|$$

for all $x \in \mathbb{R}$.

5.3.2 Kirchoff Transform

The non-linearity due to the diffusion coefficient being concentration dependent may be difficult to deal with. Luckily, by making a a clever transformation - the so called *Kirchoff transformation* - we can get rid of this non-linearity [1]. Indeed, define $\kappa : [0, \infty) \to [0, \infty)$ by

$$\kappa(C) = \int_0^C D(y) dy.$$

Because D is continuous, it follows that κ is differentiable with $\kappa'(x) = D(x)$. Because D is assumed to be strictly positive, the inverse function theorem then tells us that κ is invertible and that its inverse κ^{-1} is continuously differentiable with

$$\left(\kappa^{-1}\right)'(\kappa(C)) = \frac{1}{\kappa'(C)} = \frac{1}{D(C)}. \tag{5.40}$$

Now observe that

$$\nabla_x \kappa(C) = \frac{d\kappa}{dC} \nabla_x C = D(C) \nabla_x.$$

Substituting this into equation (5.39) yields

$$\langle \partial_t C(t), \varphi \rangle_{W^{-1,2}, W^{1,2}} = -\langle \nabla_x \kappa(C(t)), \nabla_x \varphi \rangle_{L^2} + \langle S(C(t)), \varphi \rangle_{L^2}.$$

If we further define $U(t) := \kappa(C(t))$ (so U is the Kirchoff transform of C) problem WP_1 can be reformulated as follows:

(Problem WPK_1) Find $U \in \mathcal{W}$ such that $U(0) = U_{\text{Initial}} := \kappa(C_{\text{Initial}})$ and for all $\varphi \in W^{1,2}(\Omega)$ the equality

$$\left\langle \partial_t \kappa^{-1}(U(t)), \varphi \right\rangle_{W^{-1,2},W^{1,2}} = -\left\langle \nabla_x U(t), \nabla_x \varphi \right\rangle_{L^2} + \left\langle S(\kappa^{-1}(U(t)), \varphi \right\rangle_{L^2} \tag{5.41}$$

holds for a.e. $t \in [0,T]$.

We see that the non-linearity due to the concentration dependent diffusion coefficient is gone. But this comes at a price: we have introduced a non-linearity in the time derivative. Fortunately, we can deal with this. Because κ^{-1} is differentiable the mean value theorem familiar from real analysis tells us that for every $0 \leq a < b$ there exists a $\xi \in (a,b)$ such that

$$\frac{\kappa^{-1}(b) - \kappa^{-1}(a)}{b-a} = \left[\kappa^{-1}\right]'(\xi) = \frac{1}{\kappa'(\kappa^{-1}(\xi))} \overset{(5.40)}{=} \frac{1}{D(\kappa^{-1}(\xi))}.$$

Because of the bounds on the diffusion coefficient D, see assumption H_1, it follows that

$$\frac{b-a}{M} \leq \kappa^{-1}(b) - \kappa^{-1}(a) \leq \frac{b-a}{m}. \tag{5.42}$$

Now note that $\kappa(0) = 0$ and hence $\kappa^{-1}(0) = 0$ as well. If we take $a = 0$ then the above equation shows us that

$$\frac{b}{M} \leq \kappa^{-1}(b) \leq \frac{b}{m}$$

for every $b \geq 0$. In particular, if $U \in \mathcal{W}$ then

$$\frac{U(t)}{M} \leq \kappa^{-1}(U(t)) \leq \frac{U(t)}{m} \tag{5.43}$$

for all $t \in [0,T]$. This shows that whenever we have an upper or a lower bound for $\kappa^{-1}(U)$, we have an upper or lower bound for U and it works the other way around as well.

With all of this in mind, we are going to proceed like in the linear case. That is, we are going to discretize in problem WPK_1 in time using the implicit Euler scheme. We show existence and uniqueness to the time-discrete problems. Then we derive *a priori* estimates, interpolate in time and pass to the limit. For some of the steps we will not work out the proofs because they are basically copies of the proofs for the linear case.

5.3.3 Discretizing in Time

As in the linear case, we employ the Method of Rothe.[14] That is, we discretize the weak formulation WPK in time. If Δt denotes the time step again, this yields the problems $WPK^k_{\Delta t}$ defined as follows:

(Problem $WPK^k_{1,\Delta t}$) Given $U^{k-1} \in W^{1,2}(\Omega)$, find $U^k \in W^{1,2}(\Omega)$ such that for all $\varphi \in W^{1,2}(\Omega)$ the equality

$$\left\langle \frac{\kappa^{-1}(U^k) - \kappa^{-1}(U^{k-1})}{\Delta t}, \varphi \right\rangle_{L^2} = -\left\langle \nabla_x U^k, \nabla_x \varphi \right\rangle_{L^2} + \left\langle S(\kappa^{-1}(U^k)), \varphi \right\rangle_{L^2} \tag{5.44}$$

holds. The sequence of problem is initialized by $U^0 := \kappa^1(C^{\text{Initial}})$. Our claim is that each $WPK^k_{\text{Linear},\Delta t}$ has a unique solution.

[14] More specifically, for the *a priori* estimates and the existence proof that will follow we refer to the articles *Regularization schemes for degenerate Richards equations and outflow conditions* by Pop and Schweizer [46] and *Error estimates for the finite volume discretization for the porous medium equation* by Pop et al [47].

Lemma 12. *Under the assumptions H, the problem $WPK^k_{1,\Delta t}$ admits a unique solution $U^k \in W^{1,2}(\Omega)$ for each $k \in \{1,\dots,N\}$.*

Proof. Given U^{k-1}, define the functional $B^k_{\Delta t} : W^{1,2}(\Omega) \times W^{1,2}(\Omega) \to \mathbb{R}$ as

$$B^k_{\Delta t}(u,v) = \left\langle \kappa^{-1}(u), v \right\rangle_{L^2} + \Delta t \left\langle \nabla_x u, \nabla_x v \right\rangle_{L^2} - \Delta t \left\langle S(\kappa^{-1}(u)), v \right\rangle_{L^2}$$

and the functional $F^k_{\Delta t} : W^{1,2}(\Omega)^n \to \mathbb{R}$ as

$$F^k_{\Delta t}(v) = \left\langle \kappa^{-1}(U^{k-1}), v \right\rangle_{L^2}.$$

Then problem $WPK^k_{1,\Delta t}$ can be formulated as: find $U^k \in W^{1,2}(\Omega)$ such that for all $\varphi \in W^{1,2}(\Omega)$ the equality

$$B^k_{\Delta t}(U^k, \varphi) = F^k_{\Delta t}(\varphi)$$

holds. The idea is, of course, to apply the non-linear Lax-Milgram theorem. To this end, note that $F^k_{\Delta t}$ is easily seen to be linear bounded and for each fixed u, we see that $B^k_{\Delta t}(u,\cdot)$ is linear and bounded as well. Now let $u, v, w \in \mathcal{W}$. Then

$$\begin{aligned}
\left| B^k_{\Delta t}(u,w) - B^k_{\Delta t}(v,w) \right| \quad &\leq \quad \left| \left\langle \kappa^{-1}(u) - \kappa^{-1}(v), w \right\rangle_{L^2} \right| \\
&\quad + \Delta t \left| \left\langle \nabla_x u - \nabla_x v, \nabla_x w \right\rangle_{L^2} \right| \\
&\quad + \Delta t \left| \left\langle S(\kappa^{-1}(u)) - S(\kappa^{-1}(v)), w \right\rangle_{L^2} \right| \\
&\leq \quad \left\| \kappa^{-1}(u) - \kappa^{-1}(v) \right\|_{L^2} \|w\|_{L^2} \\
&\quad + \Delta t \left\| \nabla_x u - \nabla_x v \right\|_{L^2} \|\nabla_x w\|_{L^2} \\
&\quad + \Delta t \left\| S(\kappa^{-1}(u)) - S(\kappa^{-1}(v)) \right\|_{L^2} \|w\|_{L^2} \\
&\overset{(5.42)}{\leq} \quad \frac{1}{m} \|u - v\|_{L^2} \|w\|_{L^2} + \Delta t \left\| \nabla_x u - \nabla_x v \right\|_{L^2} \|\nabla_x w\|_{L^2} \\
&\quad + \frac{\mathcal{L}}{m} \Delta t \|u - v\|_{L^2} \|w\|_{L^2} \\
&\leq \quad \max\left\{ \frac{1}{m}, 2\Delta t, \frac{\mathcal{L}}{m} \Delta t \right\} \|u - v\|_{W^{1,2}} \|w\|_{W^{1,2}}.
\end{aligned}$$

Moreover, we see that

$$\begin{aligned}
B^k_{\Delta t}(u, u-v) - B^k_{\Delta t}(v, u-v) \quad &= \quad \left\langle \kappa^{-1}(u) - \kappa^{-1}(v), u - v \right\rangle_{L^2} \\
&\quad + \Delta t \left\langle \nabla_x (u-v), \nabla_x (u-v) \right\rangle_{L^2} \\
&\quad - \Delta t \left\langle S(\kappa^{-1}(u)) - S(\kappa^{-1}(v)), u - v \right\rangle_{L^2} \\
&\geq \quad \frac{1}{M} \|u - v\|^2_{L^2} + \Delta t \left\| \nabla_x u - \nabla_x v \right\|^2_{L^2} \\
&\quad - \Delta t \left\langle S(\kappa^{-1}(u)) - S(\kappa^{-1}(v)), u - v \right\rangle_{L^2} \\
&\overset{H_2,(5.42)}{\geq} \quad \frac{1}{M} \|u - v\|^2_{L^2} + \Delta t \left\| \nabla_x u - \nabla_x v \right\|^2_{L^2} \\
&\quad + \Delta t \frac{\mathcal{L}}{M} \|u - v\|^2_{L^2} \\
&\geq \quad \min\left\{ \frac{1}{M} + \Delta t \frac{\mathcal{L}}{M}, \Delta t \right\} \|u - v\|^2_{W^{1,2}}.
\end{aligned}$$

We see that all the conditions for the non-linear Lax-Milgram theorem are fulfilled and we are allowed to conclude for each $k \in \{1,\dots,N\}$ the problem $WPK^k_{1,\Delta t}$ admits a unique solution U^k in $W^{1,2}(\Omega)$. □

5.3.4 *A Priori* Estimates

Lemma 13. *Let $\{U^k\}_{k=1}^N$ be the sequence of solutions to the problems $WPK^k_{1,\Delta t}$ obtained under the assumptions H, then there exists a constant $\mathcal{C}_1$ such that for any $p \in \{0,\dots,N\}$ the following inequality holds:*

$$\|U^p\|_{L^2}^2 + \sum_{k=1}^{p} \|U^k - U^{k-1}\|_{L^2}^2 + \Delta t \sum_{k=1}^{p} \left(\|\nabla_x U^k\|_{L^2}^2 + \frac{2M\mathcal{L}}{m} \|U^k\|_{L^2}^2 \right) \leq \mathcal{C}_1 \|U_{\textit{Initial}}\|_{L^2}^2.$$

Proof. Take $\varphi = \kappa^{-1}(U^k)$ as a test function in problem $WPK^k_{1,\Delta t}$ to find that

$$\left\langle \kappa^{-1}(U^k) - \kappa^{-1}(U^{k-1}), \kappa^{-1}(U^k) \right\rangle_{L^2} + \Delta t \left\langle \nabla_x U^k, \nabla_x \kappa^{-1}(U^k) \right\rangle_{L^2} = \Delta t \left\langle S(\kappa^{-1}(U^k), \kappa^{-1}(U^k) \right\rangle_{L^2}.$$

For the first term, we use equation (5.13) to find that

$$\begin{aligned} \left\langle \kappa^{-1}(U^k) - \kappa^{-1}(U^{k-1}), \kappa^{-1}(U^k) \right\rangle_{L^2} &= \frac{1}{2}\|\kappa^{-1}(U^k)\|_{L^2}^2 - \frac{1}{2}\|\kappa^{-1}(U^{k-1})\|_{L^2}^2 \\ &\quad + \frac{1}{2}\|\kappa^{-1}(U^k) - \kappa^{-1}(U^{k-1})\|_{L^2}^2 \end{aligned}$$

For the second term, we use equation (5.43):

$$\begin{aligned} \Delta t \left\langle \nabla_x U^k, \nabla_x \kappa^{-1}(U^k) \right\rangle_{L^2} &\geq \Delta t \int_\Omega \nabla_x U^k \cdot \nabla_x \kappa^{-1}(U^k) dx \\ &= \Delta t \int_\Omega \nabla_x U^k \cdot \left(\kappa^{-1}\right)'(U^k)(\nabla_x U^k) dx \\ &\geq \frac{\Delta t}{M} \int_\Omega \nabla_x U^k \cdot \nabla_x U^k dx \\ &\geq \frac{\Delta t}{2M} \|\nabla_x U^k\|_{L^2}^2. \end{aligned}$$

Finally, for the third term, we use the Lipschitz-continuity of the chemical reaction term S:

$$\begin{aligned} \Delta t \left\langle S(\kappa^{-1}(U^k)), \kappa^{-1}(U^k) \right\rangle_{L^2} &= \Delta t \int_\Omega S(\kappa^{-1}(U^k(x)))\kappa^{-1}(U^k(x)) dx \\ &\overset{H_2}{\geq} -\mathcal{L}\Delta t \int_\Omega \left[\kappa^{-1}(U^k(x))\right]^2 dx \\ &= -\mathcal{L}\Delta t \left\| \kappa^{-1}(U^k) \right\|^2 \\ &\overset{(5.43)}{\geq} -\frac{\mathcal{L}\Delta t}{m} \left\| U^k \right\|^2. \end{aligned}$$

Combining the above results, it follows that

$$\frac{1}{2}\|\kappa^{-1}(U^k)\|_{L^2}^2 - \frac{1}{2}\|\kappa^{-1}(U^{k-1})\|_{L^2}^2 + \frac{1}{2}\|\kappa^{-1}(U^k) - \kappa^{-1}(U^{k-1})\|_{L^2}^2 + \frac{\Delta t}{M}\|\nabla_x U^k\|_{L^2}^2 + \frac{\mathcal{L}\Delta t}{m} \left\| U^k \right\|^2 \leq 0$$

Now take $p \in \{1, \dots, N\}$, sum from $k = 1$ to p and use equation (5.43) to find that

$$\begin{aligned}
& \frac{1}{2M}\|U^p\|_{L^2}^2 + \sum_{k=1}^{p}\left\{\frac{1}{2M}\|U^k - U^{k-1}\|_{L^2}^2 + \frac{\Delta t}{2M}\|\nabla_x U^k\|_{L^2}^2 + \frac{\mathcal{L}\Delta t}{m}\left\|U^k\right\|^2\right\} \\
\overset{(5.43)}{\leq}\ & \frac{1}{2}\|\kappa^{-1}(U^p)\|_{L^2}^2 + \sum_{k=1}^{p}\left\{\frac{1}{2}\|\kappa^{-1}(U^k) - \kappa^{-1}(U^{k-1})\|_{L^2}^2 + \frac{\Delta t}{2M}\|\nabla_x U^k\|_{L^2}^2 + \frac{\mathcal{L}\Delta t}{m}\left\|U^k\right\|^2\right\} \\
\leq\ & \frac{1}{2}\|\kappa^{-1}(U^0)\|_{L^2}^2. \\
\overset{(5.43)}{\leq}\ & \frac{1}{2m}\|U^0\|_{L^2}^2 \\
=\ & \frac{1}{2m}\|U_{\text{Initial}}\|_{L^2}^2.
\end{aligned}$$

We multiply both sides by $2M$ to conclude the proof. □

Lemma 14. *Let $\{U^k\}_{k=1}^N$ be the sequence of solutions to the problems $WPK_{1,\Delta t}^k$ obtained under the assumptions H. There exists a positive constant $\mathcal{C}_2$ such that for any $p \in \{0, \dots, N\}$ the following inequality holds:*

$$\frac{1}{\Delta t}\sum_{k=1}^{p}\left\|U^k - U^{k-1}\right\|_{L^2}^2 + \|\nabla_x U^p\|_{L^2}^2 + \sum_{k=1}^{p}\|\nabla_x U^k - \nabla_x U^{k-1}\|_{L^2}^2 \leq \mathcal{C}_2\|U_{Initial}\|_{W^{1,2}}^2.$$

This proof is basically the same as the linear case, except that we have to use the mean value theorem every now and then to obtains bounds on U from bounds on $\kappa^{-1}(U)$.

Corollary 15. *Let $\{U^k\}_{k=1}^N$ be the sequence of solutions to the problems $WPK_{1,\Delta t}^k$ obtained under the assumptions H. There exist a positive constants $\mathcal{C}_1$ and $\mathcal{C}_2$ such that*

$$\begin{aligned}
\|U^k\|_{L^2}^2 &\leq \mathcal{C}_1\|U_{Initial}\|_{W^{1,2}}^2, \\
\|\nabla U^k\|_{L^2}^2 &\leq \mathcal{C}_2\|U_{Initial}\|_{W^{1,2}}^2,
\end{aligned} \tag{5.45}$$

and hence

$$\|U^k\|_{W^{1,2}}^2 \leq (\mathcal{C}_1 + \mathcal{C}_2)\|U_{Initial}\|_{W^{1,2}}^2$$

for each $k \in \{0, \dots, N\}$. Observe that each of the above three upper bounds is independent of the time step Δt.

5.3.5 Time Interpolation of Discrete-Time Solutions

As in the linear case, we define piecewise constant and piecewise linear interpolations $U_{\Delta t}^{\text{Constant}}, U_{\Delta t}^{\text{Linear}} : [0, T] \to W^{1,2}(\Omega)$ as follows:

$$U_{\Delta t}^{\text{Constant}}(t) := U^k, \quad t \in (t^{k-1}, t^k], \tag{5.46}$$

$$U_{\Delta t}^{\text{Linear}}(t) := U^{k-1} + \frac{t - t^{k-1}}{\Delta t}\left(U^k - U^{k-1}\right), \quad t \in (t^{k-1}, t^k]. \tag{5.47}$$

Lemma 16. *Let $\Delta t > 0$. Then $U_{\Delta t}^{Constant} \in L^2(0, T; W^{1,2}(\Omega))$ and*

$$U_{\Delta t}^{Linear} \in \mathcal{W} = \left\{U \in L^2\left(0, T; W^{1,2}(\Omega)\right) \mid \partial_t U \in L^2(0; T; W^{-1,2}(\Omega))\right\}.$$

Moreover, the norms $\|U_{\Delta t}^{Constant}\|_{L^2(0,T;W^{1,2}(\Omega))}$ and $\|U_{\Delta t}^{Linear}\|_{\mathcal{W}}$ are bounded uniformly with respect to Δt.

This lemma can be proved in exactly the same way as its linear counterpart Lemma 6.

5.3.6 Passing to the Limit

Using similar arguments as in the linear case we can deduce the existence of (sub)sequences $\{U_{\Delta t}^{\text{Linear}}\}$ and $\{U_{\Delta t}^{\text{Constant}}\}$ such that

$$\begin{aligned} \{U_{\Delta t}^{\text{Linear}}\} &\rightharpoonup U^{\text{Linear}} \text{ weakly in } L^2(0,T;W^{1,2}(\Omega)), && (5.48)\\ \{\partial_t U_{\Delta t}^{\text{Linear}}\} &\rightharpoonup \partial_t U^{\text{Linear}} \text{ weakly in } L^2(0,T;W^{-1,2}(\Omega)), && (5.49)\\ \{U_{\Delta t}^{\text{Constant}}\} &\rightharpoonup U^{\text{Constant}} \text{ weakly in } L^2(0,T;W^{1,2}(\Omega)). && (5.50) \end{aligned}$$

Like before, the theorem stated under *Limits and Derivatives* in the preliminaries (Section 5.1.3) allows us to conclude that $\partial_t U^{\text{Linear}}$ really is the weak derivative of U^{Linear} and hence $U^{\text{Linear}} \in \mathcal{W}$.

Lemma 17. *The weak limits of $U_{\Delta t}^{Constant}$ and $U_{\Delta t}^{Linear}$ coincide. That is, $U^{Constant} = U^{Linear}$.*

The proof mimics the proof of the linear counterpart Lemma 7. During the proof, one actually discovers that $U_{\Delta t}^{\text{Constant}} \to U$ and $U_{\Delta t}^{\text{Linear}} \to U$ strongly in $L^2(0,T;L^2(\Omega))$. Combined with the continuity of κ^{-1} we arrive at the following corollary.

Corollary 18. *There exists a $U \in \mathcal{W}^n$ such that*

$$\begin{aligned} U_{\Delta t}^{Constant} &\rightharpoonup U, \text{ weakly in } L^2(0,T;W^{1,2}(\Omega)),\\ U_{\Delta t}^{Constant} &\to U, \text{ strongly } L^2(0,T;L^2(\Omega)),\\ U_{\Delta t}^{Linear} &\rightharpoonup U, \text{ weakly in } L^2(0,T;W^{1,2}(\Omega)),\\ U_{\Delta t}^{Linear} &\to U, \text{ strongly in } L^2(0,T;L^2(\Omega)),\\ \partial_t U_{\Delta t}^{Linear} &\rightharpoonup \partial_t U, \text{ weakly in } L^2(0,T;W^{-1,2}(\Omega)),\\ \kappa^{-1}(U_{\Delta t}^{Constant}) &\to \kappa^{-1}(U), \text{ strongly in } L^2(0,T;L^2(\Omega)),\\ \kappa^{-1}(U_{\Delta t}^{Linear}) &\to \kappa^{-1}(U), \text{ strongly in } L^2(0,T;L^2(\Omega)), \end{aligned}$$

as $\Delta t \to 0$. Like before, the convergence is to be understood as taking place along a particular sequence of time steps converging to zero.

What remains to be shown is that the limit U is in fact a (weak) solution to the problem WPK.

Lemma 19. *The limit $U \in \mathcal{W}$ obtained under the assumptions H is a (weak) solution to the problem WPK.*

Proof. Let $\varphi \in L^2(0,T;W^{1,2}(\Omega))$. From the problem $WP^k_{\text{Linear},\Delta t}$ we know that

$$\left\langle \kappa^{-1}(U^k) - \kappa^{-1}(U^{k-1}), \varphi(t) \right\rangle_{L^2} = -\Delta t \left\langle \nabla_x U^k, \nabla_x \varphi(t) \right\rangle_{L^2} + \Delta t \left\langle S(\kappa^{-1}(U^k), \varphi(t) \right\rangle_{L^2} \quad (5.51)$$

for a.e. $t \in (t^{k-1}, t^k]$ (wherever φ is defined). Also remember from (5.46) and (5.24) that $U_{\Delta t}^{\text{Constant}}(t) = U^k$ and $\partial_t U_{\Delta t}^{\text{Linear}}(t) = \left(U^k - U^{k-1}\right)/\Delta t$ on $t \in (t^{k-1}, t^k]$.

Unlike before, we also define

$$K_{\Delta t}(t) = \kappa^{-1}(U^{k-1}) + \frac{t - t^{k-1}}{\Delta t}\left(\kappa^{-1}(U^k) - \kappa^{-1}(U^{k-1})\right), \quad t \in [t^{k-1}, t^k).$$

Note that one can use the same arguments as for $U_{\Delta t}^{\text{Linear}}$ to show that $K_{\Delta t} \in \mathcal{W}$ with $\partial_t K_{\Delta t}(t) = \left(\kappa^{-1}(U^k) - \kappa^{-1}(U^{k-1})\right)/\Delta t$ for $t \in (t^{k-1}, t^k)$. With this notation, equality (5.51) can be rewritten as

$$\langle \partial_t K_{\Delta t}, \varphi \rangle_{L^2} = -\Delta t \left\langle \nabla_x U_{\Delta t}^{\text{Constant}}(t), \nabla_x \varphi(t) \right\rangle_{L^2} + \Delta t \left\langle S(\kappa^{-1}(U_{\Delta t}^{\text{Constant}}), \varphi \right\rangle_{L^2}$$

Integrating the whole equation in time yields

$$\begin{aligned}\int_0^T \langle \partial_t K_{\Delta t}(t), \varphi(t)\rangle_{L^2}\, dt &= -\int_0^T \left\langle \nabla_x U_{\Delta t}^{\text{Constant}}(t), \nabla_x \varphi(t)\right\rangle_{L^2} dt \\ &\quad + \int_0^T \left\langle S(\kappa^{-1}(U_{\Delta t}^{\text{Constant}}(t)), \varphi(t)\right\rangle_{L^2} dt.\end{aligned}$$

Our first goal will be to show that

$$\int_0^T \langle \partial_t K_{\Delta t}(t), \varphi(t)\rangle_{L^2}\, dt \to \int_0^T \left\langle \partial_t \kappa^{-1}(U(t)), \varphi(t)\right\rangle_{L^2} dt$$

as $\Delta t \to 0$. To this end, note that in a similar way as for $U_{\Delta t}^{\text{Linear}}$ we can prove that there exists a $K \in \mathcal{W}$ such that

$$\begin{aligned} K_{\Delta t} &\rightharpoonup K, \text{ weakly in } L^2(0,T;W^{1,2}(\Omega)), \\ K_{\Delta t} &\to K, \text{ strongly } L^2(0,T;L^2(\Omega)).\end{aligned}$$

Similar to Lemma 17 it can be shown that $K_{\Delta t}$ and $\kappa^{-1}(U_{\Delta t}^{\text{Constant}})$ have the same limits as $\Delta t \to 0$. Now, because $U_{\Delta t}^{\text{Constant}}$ converges strongly to U in $L^2(0,T;L^2(\Omega))$ it follows from the continuity of κ^{-1} that $\kappa^{-1}(U_{\Delta t}^{\text{Constant}})$ converges strongly to $\kappa^{-1}(U)$ in $L^2(0,T;L^2(\Omega))$ and hence $K = \kappa^{-1}(U)$. Now, if we assume $\phi \in C_0^\infty(0,T), \psi \in W^{1,2}(\Omega)$ and consider $\varphi = \phi\psi$ as a test function, then we can use partial integration to find that

$$\begin{aligned}\langle \partial_t K_{\Delta t}(t), \varphi(t)\phi\rangle_{L^2} &\overset{\text{Partial Integration}}{=} -\langle K_{\Delta t}(t), \partial_t\phi(t)\psi\rangle_{L^2} \\ &\to -\langle K(t), \partial_t\phi(t)\psi\rangle_{L^2} \\ &= -\left\langle \kappa^{-1}(U(t)), \partial_t\phi(t)\psi\right\rangle_{L^2} \\ &= -\left\langle \partial_t\phi(t)\psi, \kappa^{-1}(U(t))\right\rangle_{W^{-1,2},W^{1,2}} \\ &\overset{\text{Partial Integration}}{=} \left\langle \partial_t\kappa^{-1}(U(t)), \phi(t)\psi\right\rangle_{W^{-1,2},W^{1,2}}\end{aligned}$$

for a.e. $t \in [0,T]$. Because test functions of the type $\phi\psi$ are dense in $L^2(0,T;W^{1,2}(\Omega))$, we conclude that

$$\int_0^T \langle \partial_t K_{\Delta t}(t), \varphi(t)\rangle_{L^2}\, dt \to \int_0^T \left\langle \partial_t\kappa^{-1}(U(t)), \varphi(t)\right\rangle_{W^{-1,2},W^{1,2}} dt$$

for all $\varphi \in L^2(0,T;W^{1,2}(\Omega))$ as $\Delta t \to 0$.

Next, because $U_{\Delta t}^{\text{Constant}} \rightharpoonup U$ weakly in $L^2(0,T;W^{1,2}(\Omega))$, we see that

$$\int_0^T \left\langle \nabla U_{\Delta t}^{\text{Constant}}(t), \nabla\varphi(t)\right\rangle_{[L^2]^n} dt \to \int_0^T \langle \nabla U(t), \nabla\varphi(t)\rangle_{L^2}\, dt$$

as $\Delta t \to 0$. Because S and κ^{-1} are continuous we know that $S(\kappa^{-1}(U_{\Delta t}^{\text{Constant}})) \to S(\kappa^{-1}(U))$ strongly (and hence also weakly) in $L^2(0,T;L^2(\Omega))$. That is,

$$\int_0^T \left\langle S(\kappa^{-1}(U_{\Delta t}^{\text{Constant}}(t)), \varphi\right\rangle_{L^2} dt \to \int_0^T \left\langle S(\kappa^{-1}(U(t))), \varphi(t)\right\rangle_{L^2} dt$$

as $\Delta t \to 0$. Putting the three limits together we find that U satisfies

$$\int_0^T \left\langle \partial_t\kappa^{-1}(U(t)), \varphi(t)\right\rangle_{W^{-1,2},W^{1,2}} dt + \int_0^T \langle \nabla U(t), \nabla\varphi(t)\rangle_{L^2}\, dt - \int_0^T \left\langle S(\kappa^{-1}(U(t))), \varphi(t)\right\rangle_{L^2} dt = 0. \tag{5.52}$$

As in the linear case, it follows that for all $\varphi \in W^{1,2}(\Omega)$ the equality

$$\left\langle \partial_t \kappa^{-1}(U(t)), \varphi \right\rangle_{W^{-1,2},W^{1,2}} + \left\langle \nabla_x U(t), \nabla \varphi \right\rangle_{L^2} - \left\langle S(\kappa^{-1}(U(t))), \varphi \right\rangle_{L^2}$$

holds for a.e. $t \in [0,T]$. We conclude that U is a (weak) solution to problem WPK_1. □

The next step would be to prove uniqueness of the solution to problem WPK_1. While we were able to prove this in case of Dirichlet boundary conditions (on at least a part of the boundary), showing uniqueness turned out to be problematic for the homogeneous Neumann boundary conditions and is left as an open problem. Showing existence and uniqueness for the full non-linear (system) case is also left as an open problem.

6 Applying the Tools to Real Data

In this chapter we are going to apply the tools developed in Chapters 2, 3 and 4 to real data. The data was obtained at TNO as follows. Early in the morning a sample was prepared by depositing layers of copper (Cu), indium (In) and gallium (Ga) (in that order) on top of molybdenum (Mo) coated soda-lime glass (Si and O). Then the sample was loaded into an EDX device where it was measured by E. Balder Msc. The raw data output of the EDX was processed into atomic fraction profiles by dr.ing. J. Emmelkamp. The idea in this chapter will be as follows:

1. The experimentally obtained atomic fraction profiles suffer from blur and noise. Using the methods discussed in Chapter 3 we will try to deblur (and denoise) the profiles. The deblurred profiles may reveal more information about the diffusion-reaction process occuring within the precursor stack.

2. We will then utilize the methods developed in Chapter 2 to extract diffusion coefficients from the deblurred atomic fraction profiles.

3. To check whether the computed diffusion coefficients are 'correct' we will try to recover the deblurred profiles using the model and numerical scheme described in Chapter 4.

Before performing the above steps, it will be useful to say something about what we expected to see from the measurements. To this end, we note that the production of the sample and the subsequent measurements took place at room temperature. The glass is not expected to participate in the diffusion process or react with any of the other components (that is why glass is chosen as a substrate!). Furthermore, as a general rule, components with a high melting diffuse slowly while components with low melting points diffuse much faster. Hence, because the melting point of molybdenum is high ($2623°C$), we do not expect molybdenum to participate in the diffusion process. X-ray diffraction measurements at room temperatures also show that no intermetallic phases with molybdenum form (during the selenization process the molybdenum is seen to react with selenium though). The melting points of copper ($1085°C$), indium ($156,6°C$) and gallium ($29,77°C$) are much lower and hence it is expected that these three components will interdiffuse and possibly form new intermetallic, solid phases. The EDX and X-ray diffraction measurements have confirmed this is indeed the case. More specifically, new intermetallic phases of copper and gallium atoms (Cu_xGa_y) and intermetallic phases of copper and indium atoms (Cu_xIn_y) have been observed to form. Gallium and indium do not form new intermetallic, solid phases. In fact, it may even be possible that gallium and indium form a liquid mixture at room temperature. The reason is that alloys of indium and gallium can have melting points below room temperature, as was mentioned already in Chapter 4. Hence we expect that in really short times the indium and gallium form a uniform mixture on top of the layer of copper. If we assume this to be the starting point of the diffusion process, then we are dealing with a *diffusion couple* as defined in Chapter 2. And that is good because the Den Broeder method for extracting interdiffusion coefficients requires such a setup (i.e. uniform concentrations to the left and the right of the initial contact plane).

Either way, for gallium and indium to react with copper an interdiffusion process must take place. We were hoping capture this process in the EDX measurements. Let us have a look at what the EDX measurements reveal now.

6.1 Deblurring EDX Measurements

In Figure 6.1 a plot of the atomic fraction profiles after 75 minutes obtained with the EDX device is presented. The horizontal axis represents the depth into the precursor as measured from the top, i.e. where initially the gallium was present.

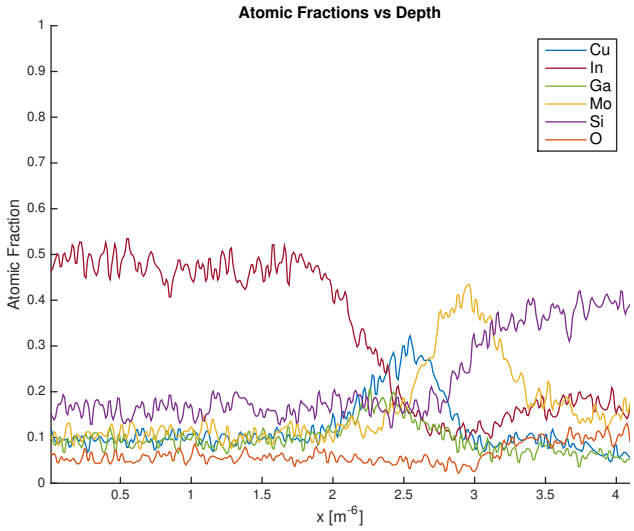

Figure 6.1: Raw data from EDX Measurement 75 minutes after prepartion of the sample.

We observe that the results are polluted with random noise. Moreover, remember that the sample consists of molybdenum (Mo) coated soda-lime glass (Si and O) on top of which layers of copper (Cu), indium (In) and gallium (Ga) have been deposited. Now Figure 6.1 seems to suggest that the glass (Si and O) has mixed with the other components. In reality this is not expected, certainly not after such a short time at room temperature! In the left half of the plot (i.e. the top of the sample) we see that glass seems to be present while there should be none. It appears as if there is a certain base level of noise present. The base level of noise appears to be different for different components. In the right half of the plot it appears as if the molybdenum and glass have diffused into each other. Again, this is not what we expect to happen at room temperature. The mixing is most likely the result of the limited resolution of the EDX measurement device (i.e. blur). We conclude that at least three types of errors are present in the plots: *random noise*, *blur* and *base level noise*. Before we try to mitigate these errors, note that in the end we are not interested in the glass and the molybdenum because they are not expected to participate in the diffusion-reaction process at room temperature. Therefore we eliminate these components and normalize the remaining components. This yields the plot presented in Figure 6.2.

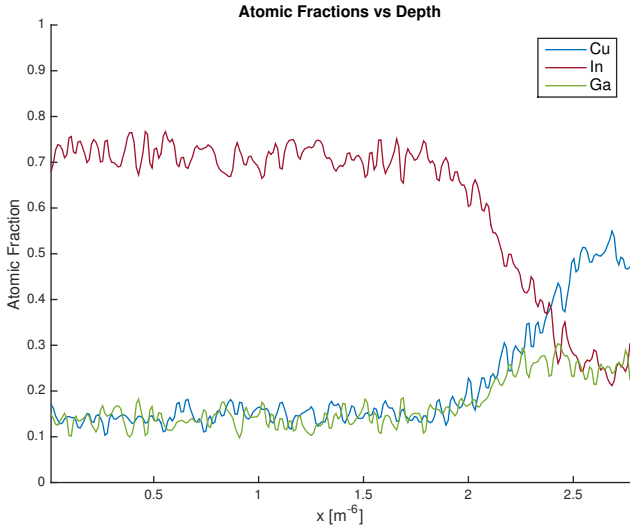

Figure 6.2: Raw data from EDX Measurement 75 minutes after prepartion of the sample with Si, O and Mo eliminated.

Note that we have cut off the plot at a depth of approximately $2,75\mu m$: at this depth the layer of molybdenum is expected to be found and we do not expect there to be any copper, indium or gallium. To proceed, we could mitigate the random noise in the results by smoothing with a simple moving average scheme. This would introduces more blur though. As we saw in Chapter 3, our deblurring schemes are capable of dealing with noise so there is no need to smoothen the profile before deblurring. To mitigate the base level errors, we could subtract for each component the minimum value attained by this component and then renormalize. Because of the random noise currently present it is more effective to do this *after* the deblurring. So, now is the time to try the deblurring algorithms presented in Chapter 3. Employing these algorithms is not straightforward though. This can be explained as follows. In Chapter 3 we saw that the deblurring algorithms require a lot of input parameters: a parameter $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2)$ for the Tikhonov filters and parameters β and γ for the Total Variation filter. Moreover, we do not have details on the blur caused by the EDX device, i.e. we do not know the blurring operator. As in Chapter 3.3 we will make the assumption that blurring operator is linear and can be represented by convolution with a response function. For the response function we make the assumption that it is either Gaussian- or square shaped with a certain width σ. In the test cases considered in Chapter 3.5 we tested the deblurring methods by taking some (random) profile, blurring it with a chosen blurring operator, adding some random noise and then we tried to deblur the resulting profile. The methods looked promising but in a certain sense we were cheating because we knew the blurring operator that was used. Moreover, because we knew what the real, original profiles looked like, we could compare the deblurred images with the true images to get a feeling for what would be 'good' parameters.[15] But this time, we do not know what the true image would look like. We only have physical intuition to guide us in choosing the parameters. Let us it give it a try though. In the next three figures some results are presented. The profiles are deblurred using three different sets of parameters. Note each time after deblurring, the fractions do not sum up to 1 everywhere (the yellow dashed lines). This is because the deblurring algorithm is applied to each component separately. We tried to implement an algorithm that deblurs all the profiles simultaneously under the constraint that at all times the profiles should sum to 1 (using Lagrange multipliers), but without succes: the constraint was too strong in the sense that the method did not feel any freedom to do actual deblurring. That is why we decided to simply normalize the profiles *after* deblurring. We also remove the base level noise for each component before normalizing. The resulting 'Corrected' profiles are shown in the right-most plot in the figures.

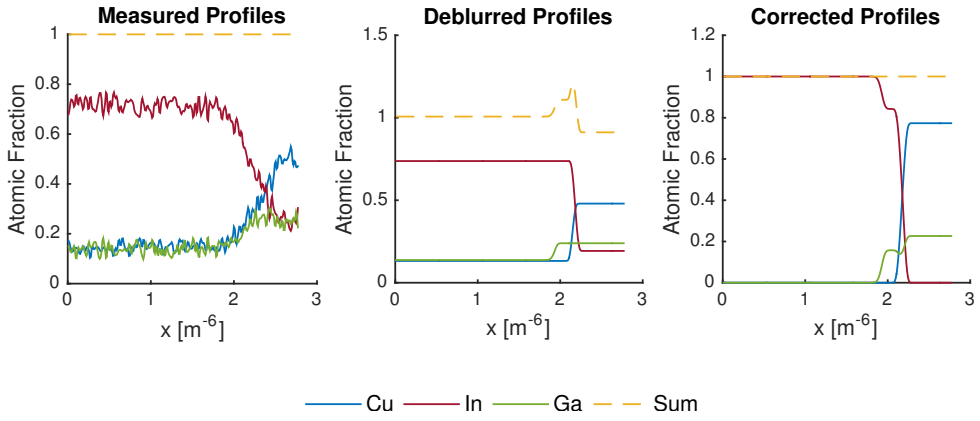

Figure 6.3: Data from Figure 6.2 deblurred with $\boldsymbol{\alpha} = (0, 0, 10^{-8}), \beta = 1, \gamma = 10^{-6}$ and a Gaussian response function with standard deviation $0.23\mu m$ and reflexive boundary conditions. Afterwards, the base level noise was removed and the profiles were normalized.

[15] Although we did not work this out in Chapter 3, one could try to work out an optimization scheme that looks for the *best* parameters, i.e. the parameters which minimize the error between the true and deblurred profiles (in some appropriate norm).

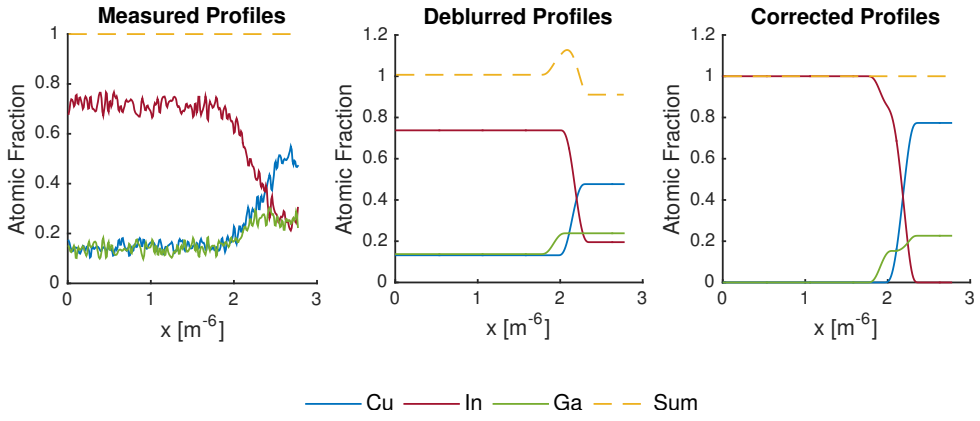

Figure 6.4: Data from Figure 6.2 deblurred with $\boldsymbol{\alpha} = (0, 0, 10^{-6}), \beta = 0, \gamma = 10^{-6}$ and a Gaussian response function with standard deviation $0.23\mu m$ and reflexive boundary conditions. Afterwards, the base level noise was removed and the profiles were normalized.

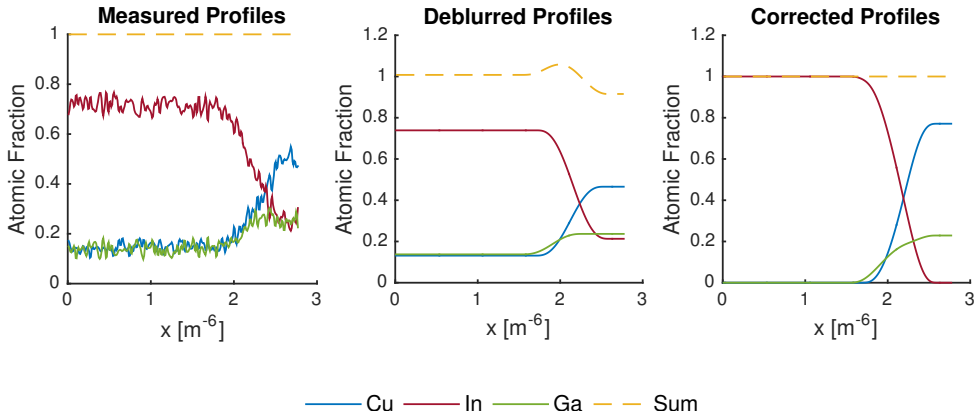

Figure 6.5: Data from Figure 6.2 deblurred with $\boldsymbol{\alpha} = (0, 0, 10^{-4}), \beta = 1, \gamma = 10^{-6}$ and a Gaussian response function with standard deviation $0.23\mu m$ and reflexive boundary conditions. Afterwards, the base level noise was removed and the profiles were normalized.

The question is now: which profile resembles best the real physical situation? And the problem is that we do not know. But more on that later.

For now, remember how initially the gallium and indium were assumed to form a uniform mixture on top of the layer of copper? And how we were hoping to see it migrate towards the copper? In the above figures we see that there is no gallium to be found on top. All the gallium has already diffused towards the copper (and most likely reacted with it). And all of this happened *before* the first measurement. And that is a problem for the Den Broeder method because, as we saw in Chapter 2, this method relies on the assumption that both ends of the diffusion couple remain unaffected during the diffusion process. Clearly we have a problem here: we have lost our sense of time. The gallium could have migrated towards the copper in a few minutes, a second or perhaps even milliseconds. As one can image, this has tremendous influences on the interdiffusion coefficients.

That is why it was decided to turn the problem around as follows. We assume that initially there is a uniform mixture of copper and gallium on top of which there is a layer of pure indium. And then perhaps the EDX measurements should be interpreted as showing an interdiffusion process that starts from this new initial profile. If we assume this to be the case, then at least we are in a position where the methods discussed in Chapter 2 can be applied. Let us see where this leads us.

6.2 Extracting Average Interdiffusion Coefficients

To extract interdiffusion coefficients, we first convert the atomic fraction profiles from Figures 6.3, 6.4 and 6.5 into concentration profiles. Under the assumption of constant partial molar volumes, the concentrations C_i and atomic fractions N_i are related by

$$C_i = \frac{N_i}{V_{\text{mol}}} = \frac{N_i}{\sum_{i=1}^{n} N_i V_i},$$

where V_i are the (constant) partial molar volumes (see equation (2.1)). Because we are dealing with a system of three components, we can only hope to find *average* interdiffusion coefficients using the algorithm presented in Subsection 2.3.3. For each of the three concentration profiles, we divided the domain into three regions over which we computed the average interdiffusion coefficients. Because we are dealing with a ternary system, there are two independent components and hence $2^2 = 4$ interdiffusion coefficients to be found in each region. We labeled them $D_{11}, D_{12}, D_{21}$and D_{22} respectively. The results are presented in the next three figures. On the left the concentration profiles are shown and on the right the extracted average interdiffusion coefficients. In each case we choose gallium to be the independent component.[16]

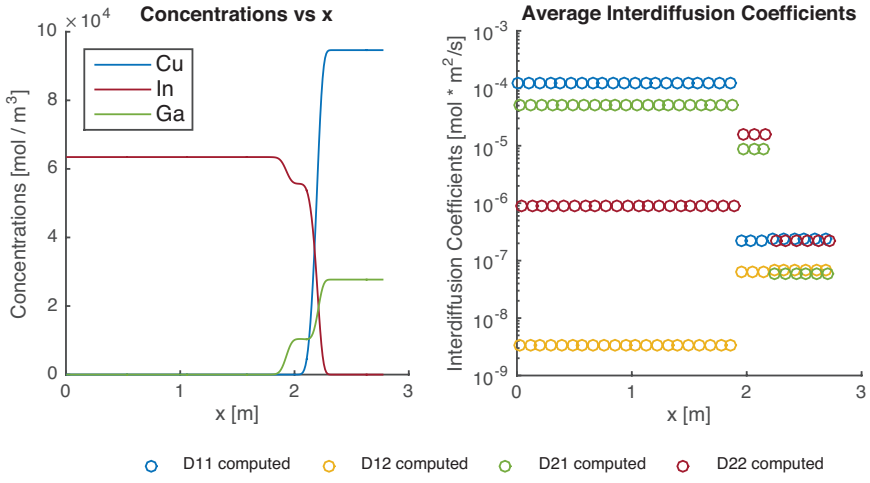

Figure 6.6: (Left) Concentration profile related to Figure 6.3.
(Right) Average interdiffusion coefficients computed over three regions with gallium chosen as independent component.

[16] As remarked in Chapter (2), choosing different dependent components gives rise to different interdiffusion coefficients. Relationships between the different (average) interdiffusion coefficients can be found in Chapter 9.1 of [43].

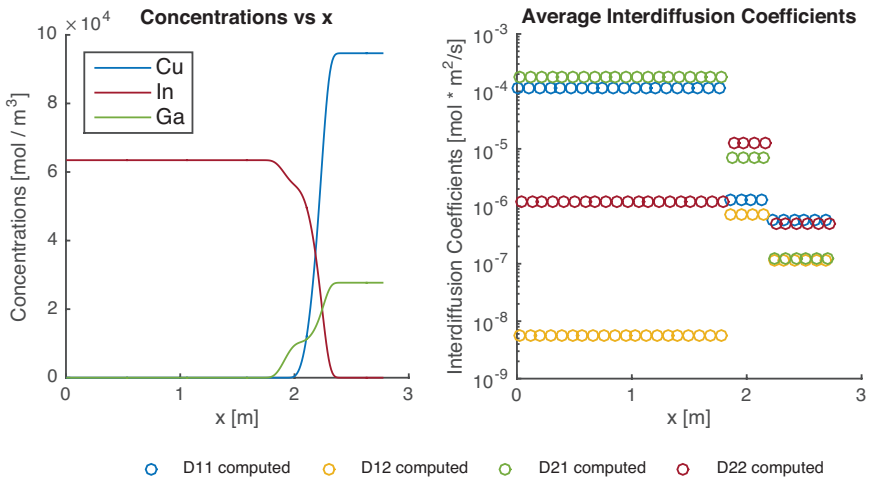

Figure 6.7: (Left) Concentration profile related to Figure 6.4.
(Right) Average interdiffusion coefficients computed over three regions with gallium chosen as independent component.

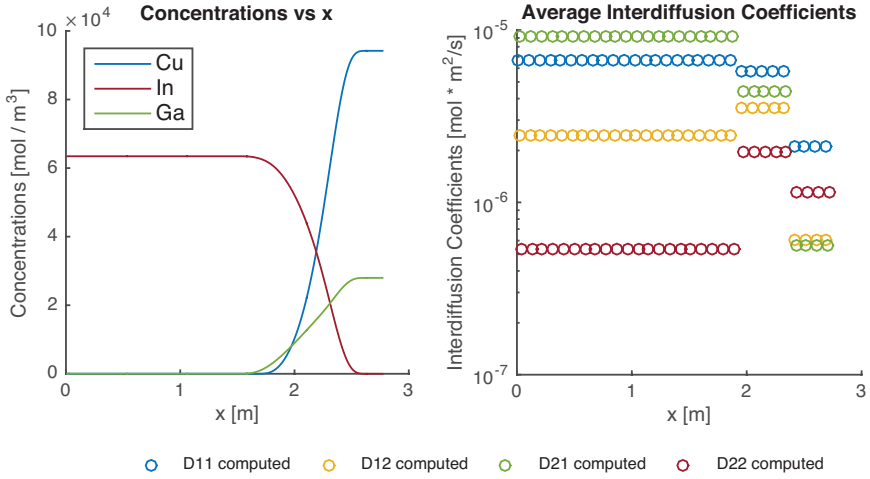

Figure 6.8: (Left) Concentration profile related to Figure 6.5.
(Right) Average interdiffusion coefficients computed over three regions with gallium chosen as independent component.

We observe that some of the coefficients are similar for the three profiles while others differ by one or more orders of magnitude. Also, if we look at the computed values for D_{22}, we see that in Figure 6.6 the values in the left region are higher than in the right region. In Figure 6.8 the opposite is true. The method used for extracting the average interdiffusion coefficients appears to be sensitive to the parameters used for deblurring (as expected). On top of that, we have also discussed in Chapter 2 that for ternary systems one cannot blindly trust the computed average interdiffusion coefficients to be the correct ones. We suspected that different sets of interdiffusion coefficients can give rise to a similar concentration profile at the time of measurement t^* (they may diverge before or afterwards). Moreover, to find concentration dependent interdiffusion coefficients to be used in the numerical method described in Chapter 4, it would make sense to fit concentration-dependent interdiffusion functions to the computed average interdiffusion coefficients. But it is difficult to say without *a priori* knowledge if one should do a constant, linear or quadratic fit or perhaps a piecewise linear fit. And doing the right fit may be important: as mentioned in Chapter 2.3, the diffusion coefficient matrix should always be positive definite. If it is not (and this may be the results of not fitting correctly), the diffusion equations may become unstable and the numerical solutions produced with the scheme from Chapter 4 may blow up.

We conclude that while both the deblurring methods and the methods for extracting interdiffusion coefficients have their respective uses, when combined their shortcomings and uncertainties 'multiply'. This makes it extremely difficult to get reliable results and to test them, especially in the case of thin films diffusion experiments where it is much more difficult to do accurate measurements than in bulk diffusion experiments. One may wonder if it would be possible to repeat the thin film experiments in a bulk setting to get more reliable results. Unfortunately, bulk diffusion and thin film diffusion may be governed by different types of diffusion mechanisms (grain boundary diffusion may be more pronounced in the case of thin film diffusion, leading to much higher diffusion coefficients). Results from one type of experiment cannot be directly translated to the other [24].

7 Summary and Suggestions for Future Work

7.1 Summary

- In Chapter 2 we explored solid-state diffusion. Fick's law, the more general Onsager's transport equations and the importance of choosing reference frames for expressing fluxes were discussed. Moreover we discussed and implemented a direct method for extracting concentration-dependent interdiffusion coefficients from a single concentration profile of a diffusion couples: the Boltzmann-Matano method and the more refined Den Broeder method. For diffusion couples consisting of two (atomic) components we were able to recover the interdiffusion coefficients to great accuracy in our test cases. We also discussed the theoretical and practical problems associated with interdiffusion in systems with more than two components. To obtain the interdiffusion coefficients at a specific composition one needs to prepare different diffusion couples whose diffusion paths intersect. In practice, especially when dealing with thin films, this may not always be possible to achieve. To overcome this issue, we discussed a method proposed in literature to derive average interdiffusion coefficients from the concentration profile of a single diffusion couple. The test cases showed us that the results should be treated with great care though. The method may results in the 'wrong' coefficients that can still be used to recover to good accuracy concentration profiles (at the specific time of measurements) from which they were derived. We believe the problem in this respect is that different sets of diffusion coefficients may give rise to (before and afterwards they may diverge). On top of that, even if one obtains the correct average interdiffusion coefficients, it is difficult to say how one should properly fit concentration-dependent interdiffusion functions to these average interdiffusion coefficients.

- In Chapter 3 we presented methods to deblur experimentally obtained concentration / atomic fraction profiles that suffer from blur. We saw how simply inverting the blurring operator, assuming it is known, causes blow-up and one needs to apply filters to prevent this blow-up. We formulated the image deblurring problem as a minimization problem and discussed two types of filters: a generalized Tikhonov filter and a Total Variation filter. Moreover, presented algorithms to solve the minimization problem. Numerical tests confirm the theoretical predictions that when deblurring with a Tikhonov filter one can nicely recover smooth edges but no sharp edges. On the other hand, using the Total Variation filter one can nicely recover sharp edges while smooth edges are problematic. We tried to combine both methods but there will always be a trade-off between the two results. Furthermore, because in real situations one usually does not know what the exact blurring operator looks like we presented some possibilities of what blurring operators could look like.

- In Chapter 4 we described the precursor model that is part of a larger model being developed at TNO/Solliance. We saw that the model is basically a system of coupled, non-linear diffusion reaction equations with no-flux boundary conditions. A numerical scheme was presented to solve this system of equations. The scheme consists of a finite-volume discretization in space together with a semi-implicit time integration method. The numerical scheme was implemented in the larger TNO/Solliance model and is seen to be a factor 400 faster than the old numerical scheme used at TNO/Solliance. Moreover, the new scheme can properly handle concentration-dependent diffusion coefficients and it can easily be scaled to include new components or intermetallic phases. The scheme was also used in generating concentration profiles for testing the methods presented in Chapters 3.

- The precursor model described in Chapter 4 was subjected to a rigorous mathematical analysis in Chapter 5. We first presented a weak formulation of the precursor model. Then,

assuming the problem to be linear, we were able to prove existence and uniqueness of (weak) solutions using the so called Method of Rothe. This method employs implicit Euler time discretization with a time step Δt to reduce the full space/time-dependent problems to a sequence of space-dependent (elliptic) problems. Using standard tools from functional analysis we could show existence and uniqueness of solutions to these space-dependent problems. Then, in the next step these solutions are patched together to form a solution approximation on the whole space/time domain. By proving *a priori* estimates we could ensure that the resulting solution approximation converges to a limit as $\Delta t \to 0$. We were able to show that the limit obtained this way is a (weak) solution to the original linear problem. Next, we went back to the non-linear case but then for a single component only. We used the *Kirchoff Transform* to remove the non-linearity from the flux term at the expense of introducing non-linearities in the time derivative and the source term. The latter were not too difficult to deal with though. In fact, we were able to follow similar steps as in the linear case to show existence of a solution. Showing uniqueness turned out to be problematic though and is left as an open problem. Showing existence and uniqueness for the full non-linear (system) case is also left as an open problem.

- In the final chapter, Chapter 6, we applied the tools developed in previous chapters to real data. Precursors were prepared at TNO/Solliance and using *cross-section Energy Dispersive X-Ray Spectroscopy* atomic fraction profiles of the different atomic components in the precursor were prepared. Because the layers of interest are really thin (a few μm), the limited resolution of the EDX caused the atomic fraction profiles to appear blurred (and as always in measurements noise is present as well). Using the method developed in Chapter 3 we deblurred the resulting atomic fraction profiles. A problem in this respect is that the deblurring method requires several input parameters and different parameters give rise to different deblurred profiles. And of course we do not know what the 'true' profiles look like (otherwise there would be no need to deblur). In most deblurring applications one is satisfied if something, say a license plate, that was previously unreadable can be read after the deblurring. In our case, the profiles are deblurred to prepare them for extracting interdiffusion coefficients using the methods described in Chapter 2. Unfortunately, for a process like diffusion, the difference between a sharp edge and a smooth edge can be quite large and hence different. That is, the diffusion coefficients may be highly sensitive to the way in which the profiles are deblurred. Our tests indeed reveal that the average interdiffusion coefficients obtained from the different profiles (that are supposed to represent the same true) can differ by orders of magnitude. And, even if we would be able to recover the 'true' profiles with the deblurring methods, there is still the problem associated with the reliability of the method itself. So, there is a lot of uncertainty and results are difficult to verify. We also had to be 'creative' to get the data to qualify as resulting from a diffusion couple (as defined in Chapter 2) because the diffusion processes at much faster rates than initially expected. Most of the discussed problems are related to the fact that we are dealing with thin films which are in general difficult to measure. While the developed tools may help to get additional insight into the growth process of the CIGS absorber layer in the two-step process, the results should not be trusted blindly. The developed tools may also prove to be useful for other projects at TNO and/or Solliance.

7.2 Suggestions for Future Work

- In Chapter 3 we have been dealing with minimization problems. When trying to solve such problems it is important to know wether a (global) minimizer even exists and if it does wether the minimizer is unique. If there is a unique minimizer, we would like to know whether the iterative scheme used to update solution approximations eventually converges

to the minimizer. For both the Tikhonov and Total Variation filters separately literature is available on these issues. It would also be interesting to work on such problems when combining the Tikhonov and Total Variation filters.

- Instead of trying to approach the deblurring problem as a minimization problem the following approach can be tried. It is assumed that a sharp interface should be present between the soda-lime glass and molybdenum in the deblurred images. Then, by comparing gradients in glass / molybdenum profiles against the gradients for the other components, one may be able to work out a scheme that determines how sharp or smooth the edges in the edges of the other profiles should be. Perhaps such a method could be combined with the minimization schemes developed in this thesis as well.

- The methods used to extract interdiffusion coefficients from concentration profiles are direct methods. Perhaps it would also be possible to formulate the identification of such parameters as a minimization problem. I can imagine that - especially for finding concentration-dependent interdiffusion coefficients - this can be really challenging. The benefit may be that one is not restricted to working with so called diffusion couples.

- Instead of trying to model the precursor using 'real' interdiffusion coefficients and chemical reaction rates it may beneficial to model the physical processes in terms of in terms of 'slow' processes and 'fast' processes. If the diffusion of one component is seen to be many orders of magnitude faster than the diffusion of some other components, it may not be necessary to know the real diffusion coefficients but only that one is much higher than the other (and similarly for chemical reactions). Such an approac can be paired with mathematical asymptotic analysis.

- The spatial discretization scheme proposed in Chapter 4, i.e. the finite volume scheme, is expected to be second order accurate. On the other hand, the time integration scheme proposed, i.e. the semi-implicit Euler scheme, is only first order accurate. Depending on the needs of TNO/Solliance it may be worth it to work out higher order time discretization schemes as well. For example, the time integration schemes discussed Chapter 4 can be combined to obtain a second order Crank-Nicolson scheme.

- The mathematical analysis has been restricted to the precursor model so far. In the second step of the so called two-step process, i.e. during the selenization of the precursor, the geometry of the physical domain changes in time due to the absorption of selenium. Including the changing geometry into the problem may give rise to some interesting mathematics.

A Working Principles of Solar Cells

In this appendix we will explore the working principles of solar cells. In the most general sense solar cells operate by converting energy emitted by the sun in the form of light into electrical energy that can power our electrical devices. To understand what is really going on under the hood, we first have to look more closely at what light actually is. Then we will look at how light can give off its energy to materials. As we will see, this requires us to zoom in to the subatomic level where some really strange things are happening.

This chapter is not intended to give complete and mathematical precise description of the matters at hand. It only serves as a gentle and intuitive introduction so that the reader can get a feeling and an appreciation for the working principles of solar cells. The reader is assumed to have a basic knowledge of classical mechanics, electromagnetism, probability theory and differential equations. Sections 1.1 and 1.2 cover material that should be familiar to anyone who took physics and chemistry in high school. Section 1.3 is basically a short introduction to quantum mechanics and relies more on mathematical formulations than other subsections. Finally, in subsection 1.4 we dive into the subject of semiconductor physics and work towards a model of a simple solar cell. The main source used in writing this chapter was the text book *University Physics* by Young and Freedman [60]. Other used sources are *Quantum Mechanics: An Introduction for Device Physicists and Electrical Engineers* by Ferry [20], *Physics of Solar Cells - From Principles to New Concepts* by Würfel [58], *Nanoelectric Devices* by Park, Park and Hwang [42] and finally *Lecture Notes on Quantum Mechanics* by Greensite [23].

A.1 Light

A.1.1 Light as Particles

The classical way to envision light is to think of it as being colored particles which emerge from a certain source - like the sun. These particles travel through space and when they hit an object, they either get absorbed, reflected or pass through the object. This interpretation of light can explain a lot everyday things. For example, when you are looking at a an object - say a cup of coffee - what actually happens is that light particles - coming from somewhere - reflect off the cup and enter your eye. The eye transmits this 'signal' to your brain and your brain turns it into an image of a cup of coffee. When an object is perceived as black, it basically means all light particles hitting the object are absorbed and none reflect into your eye. When an object is perceived as green, it means the green light particles are being reflected while all others are being absorbed. White objects reflect all light.

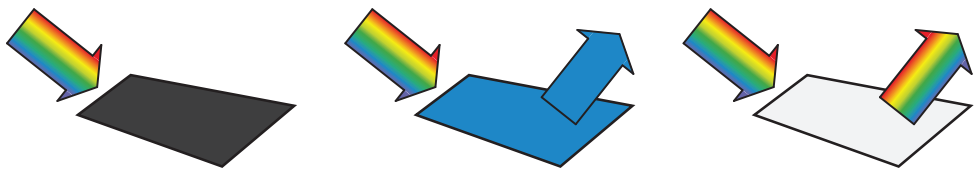

Figure A.1: Colors as reflections of specific light particles. Black objects absorb all incoming light particles while objects reflect them.

Now if you wear a black t-shirt on a sunny day, you will feel much warmer than if you had chosen to wear a white shirt instead. This must have something to do with light because - as we just

learned - the black shirt absorbs light rays while the white shirt does not. In fact, this suggests that light rays contain energy. The black shirt absorbs light from the sun and the energy in this light is transferred as heat to your body.

This model of light being particles is simple, elegant and it explains everyday things pretty well. Naturally, one may start to wonder if there even is anything to say about light that can not be explained by light being particles. It wasn't until the beginning of the nineteenth century that people had to accept that our particle model is not complete. Let us see why.

A.1.2 Light as Waves

Picture yourself in a room with a lamp on a table. Cover this lamp with a light absorbing box (a black box) and drill two tiny holes in one side of the box close to one another. Now turn off the light in the room, turn on the lamp and look at the wall of the room facing the two holes. What do you expect to see? Two dots of light, right? The only light particles that can escape from the box are the ones passing in a straight line from the source through the holes you drilled. Because there are two holes, you expect to see two dots on the wall.

In 1803, the British scientist Thomas Young conducted an experiment similar to what was just described. But when he looked at the wall he saw a pattern like in Figure A.2:

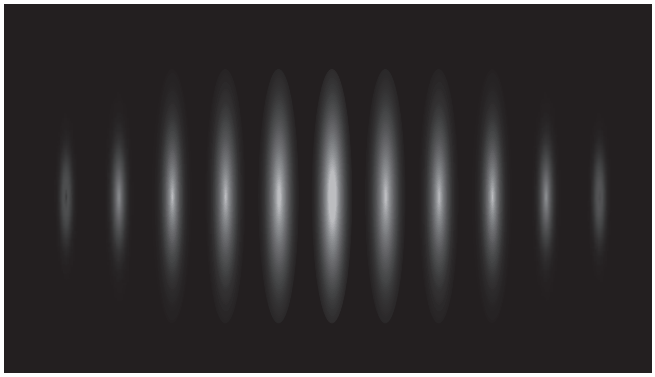

Figure A.2: Light Pattern observed by Young.

What's going on here? Luckily, Thomas Young had an explanation. He had observed this kind of behaviour before when experimenting with water. Or more specificaly, with water *waves*. Say you throw a stone in a pond. The moment the stone hits the water, it will start to ripple. That is, water waves emerge from where the stone hits the water and they propagate outwards in all directions. A water wave can be thought of as a disturbance or displacement from an equilibrium position. Now image throwing two stones into the pond instead of just one. There will be two sources of waves. What happens when the waves caused by one source meet with the waves caused by the other? As it turns out, the waves behave according to the so called *superposition principle*: the two waves become one and the new wave's displacement is given by the sum of the displacements of the two initial waves. If both waves are at their peaks when they meet, the new wave will have an even higher peak than the individual waves. This is called *constructive interference*. However, if one is at its peak while the other is at its trough, the waves will (partially) cancel each other out. This is called *destructive interference.*

Now how is this related to our experiment? Assume some water waves emerging from a source travel in a particular direction. After a while they hit a wall which only has a small opening in it. What will remain of the waves passing through this small hole? It turns out - and this can be explained by something called Huygens Principle - that behind the wall it will look as if the opening in the wall is a new source of waves. In particular, the waves are able to crawl around

corners. None of this is specific to water: other kinds of waves show similar behaviour. Think of sound - that is, pressure waves propagating through air. Because of sound being waves we are able to 'hear around corners'.

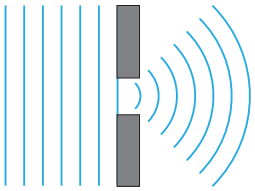

Figure A.3: Huyghen's Principle visualised.

If we change the setup of the above experiment so that the wall has two openings, there will be two 'sources' of waves behind the wall and the resulting waves will start interfering. If we place a wall to the right of the openings, then intensity patterns will develop on this wall that are strikingly similar to what Thomas Young saw when he looked at his wall. He had found definite proof that light comes in waves. But what kind of a wave is this thing called light then? We call something a *water wave* because it is a disturbance of the water level traveling *through* the water. Similarly, a wave created in a rope is is a disturbance travelling through that rope. So, what is it that light waves are disturbing then? As it turns out, light waves are oscillating electric and magnetic fields propagating at the speed of light. We also call them electromagnetic waves and their behaviour can be described by Maxwell's equations, after the Scottish scientist James Maxwell.

The weird thing about electromagnetic waves in comparison with other kinds of waves is that they don't require a medium to travel through. They can even travel through a vacuum. That is very good: otherwise light - or as we now know, electromagnetic waves - generated by the sun would not be able to reach us here on earth.

Now we know that light comes in waves. In general, waves have a frequency f and wavelength λ that are related to the speed of the wave - in this case the speed of light in a vacuum, denoted by c - in the following way:

$$c = \lambda f.$$

If we know the frequency of a wave, we can figure out the wavelength and vice versa because c is constant. Not all electromagnetic waves are visible to the human eye. We can only see waves with wavelength between approximately 400 and 700 nanometers. This is illustrated below.

Figure A.4: Part of the Electromagnetic Spectrum visualised.

The wave model of light can explain everything that the particle model of light can, and then some more. But this is not the end of the story, as we will see in the next section.

A.1.3 Light as Photons

At the end of the nineteenth century it was discovered that many metals emit little particles called electrons (much more on those later) when light shines on them. These electrons are normally bound to the metallic material by electric forces. By gaining enough energy they are able to escape from the material. Apparently, the electrons can get this energy from absorbing light. This also suggests that light must contain energy. The idea that light contains energy is perfectly compatible with the wave model of light. It may be hard to picture how this works for light, but if you think of a tidal wave, it is not such a weird idea that it contains energy! It is also a reasonable assumption that the energy in the wave is proportional to the amplitude of the wave. A huge tidal wave may destroy buildings, but a small ripple in a pond is not going to do much. As it turns out, one can derive from Maxwell's equations that the *intensity* - that is, the energy per unit time per unit surface area - of an electromagnetic wave is proportional to the (square of) the amplitude. In doing so, one also sees that the intensity is not dependent on the frequency of the light waves. So, if we go back to the photoelectric experiments, it is expected that increasing the intensity of the light incident on the metal increases the emission rate of electrons. Why? Electrons need energy to escape from the material. Light contains energy. The higher the amplitude, the higher the energy transfer rate. Hence more electrons should be able to absorb enough energy to escape. The frequency of the light waves should not be relevant. Sounds reasonable, right?

Now here's the weird thing: the photoelectric experiments showed a completely different behaviour. If light of a low frequency - say red or infrared light - was used, no electrons would be emitted. Increasing the intensity made no difference. On the other hand, when using high frequency light - like ultraviolet - they would observe a high electron emission rate, even at low intensities.

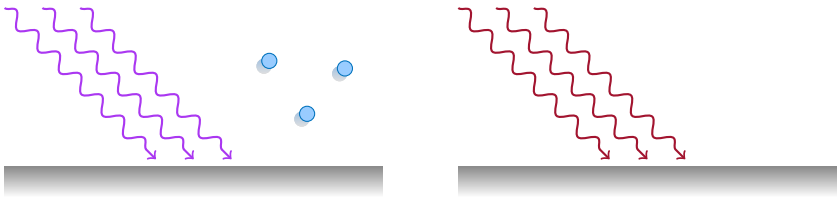

Figure A.5: Photoelectric Effect. High frequency light (purple) incident on a material causes emission of electrons (blue). No electrons are emitted when lower frequency light (red) is used.

What's going on here? Could it be the case that our wave model of light is also wrong, or at least incomplete? It was Albert Einstein who suggested a new model for light, the photon model, in 1905. He proposed that light comes in little packages of energy called photons and that the energy E contained in a photon is frequency-dependent:

$$E = hf,$$

where h is an incredibly small number (with units of $J \cdot s$) called Planck's constant. If this is indeed the case, the photoelectric effect can be properly explained as follows. Electrons need a minimum amount of energy to 'jump' out of the material. If they absorb a photon with enough energy, i.e. a photon with high enough frequency, they will be able to make the jump. If a low frequency photon is absorbed, the electron will not be able to make the jump and fall back to its original position (or energy level). It can certainly try to jump again by absorbing a new photon, but it will only escape if a photon with high enough frequency is absorbed.

According to the wave model, a high frequency and a low frequency wave should carry the same energy as long as the waves have equal amplitude. On the other hand, the photon model

is saying that high frequency photons should carry more energy than low frequency ones. To bring these two ideas together, we must assume that the high frequency wave contains less (but more energetic) photons than the low frequency one. But if we talk about there being 'more' or 'less' photons in light, it feels like we are dealing with a ray or particle model again, even though the photons are assumed to have a frequency (and hence a wavelength). In fact, photons do turn out to be particles. There are even devices - so called photonmultipliers - which can detect single photons. But the problem with seeing light as particles is that it is not possible to explain behaviour like interference. That is why we needed the wave model in the first place. But with the wave model, we run into problems with the photoelectric effect, which requires a more particle-like model. How do we get out of this circle? The answer to this question is remarkable: we accept both models at the same time. This is referred to as wave-particle duality. Sometimes we need one model, sometimes the other. One is not more true or better than the other. It just happens to be - if we understand correctly - that light has a dual nature. [17] And as we will see later, the story does not end here!

Now that we are more familiar with the concept(s) of light, let us start looking at how we can transform the energy contained in light into electricity. Devices which can do this are called *photovoltaics*. To understand photovoltaics, we need to get a better understanding of *atoms* first, the building blocks of everything around us.

[17] In a certain way, this proposed duality solution is unsatisfactory. How do you know up front whether to consider light as particles or as waves? And how does nature itself know how to distinguish between the two cases. In fact there is an model of light which solves this duality problem. This is the so called 'path integral formulation' as developed by the American scientist Richard Phillips Feynman. It is based on quantum physical theories which I don't want to discuss at this point. A gentle introduction to this model can be found in Feynman [21]

A.2 Atoms

A.2.1 Smallest Particles

Suppose you cut a piece of some material in half again and again until you are holding a 'smallest particle' which cannot be cut in half anymore. Ancient Greek philosophers referred to this *idea* of smallest particles as *atoms*. But it wasn't until the 19th century that people started finding scientific results and arguments which could back up this idea of atoms. Along the way, numerous different kinds of atoms with different kinds of properties were discovered and categorized in the so called periodic table. As of today, the periodic table contains 118 different atoms.

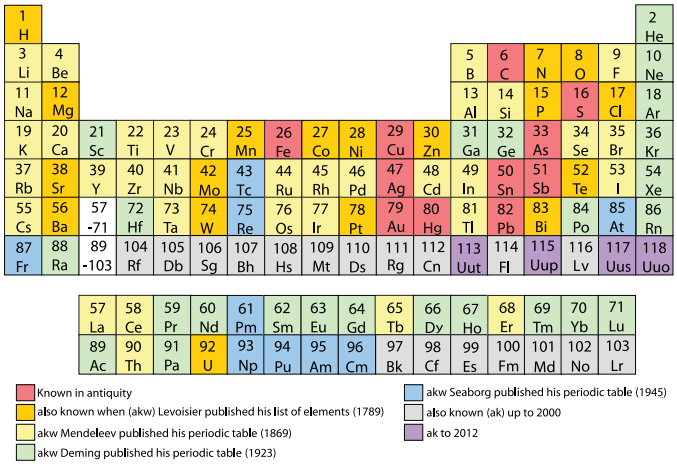

Figure A.6: Periodic Table
{https://upload.wikimedia.org/wikipedia/commons/3/3d/Discovery_of_chemical_elements.svg - By Sandbh (Wikimedia Commons.) [CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons}

The smallest and lightest atom known is the hydrogen atom (the *H* in the upper left corner of the periodic table). Since atoms are supposed to be the building blocks of all other matter, we should be able to conclude that nothing can be smaller and lighter than hydrogen atoms. But then in 1897 the British scientist Joseph Thomson made an interesting discovery with so called *cathode rays*. What he did was the following. Two plates of conducting material were placed inside of a vacuum glass tube. One of the plates had an opening. The plate with the opening is given a positive charge and is referred to as the *anode*. The other plate is given a negative charge and is called the *cathode*. Under a high enough voltage between the cathode and the anode, a stream of particles started to flow from the cathode to the anode. These particles were not visible to the naked eye, but they made the glass behind the opening in the anode glow.[18] Joseph Thomson was able to measure the charge-to-mass ratio. This ratio was huge, meaning that either the charge of the particles was huge, or the mass of the electrons was really small. Thomson went for the latter and figured out that the mass of the electrons more than a thousand times smaller than the mass of hydrogen atoms.[19] Furthermore, he found out that the particles emitted by the negatively charged electrode were the same for different kinds of cathodes. That is, cathodes made out of different atoms emitted the same kind of particles. All of this lead Joseph Thomson to conclude that atoms themselves must be built out of even smaller particles: *subatomic particles*. The subatomic particles emitted by the cathodes are what we now call *electrons*. Because electrons have a negative electric charge - that is why flow from the cathode

[18] Couple this setup with magnets to bend the trajectories of the particles and you have a television screen!

[19] Later experiments in which the charge of these particles had been measured confirm that Thomson was on the right track here.

to the anode - and because atoms themselves are electrically neutral, there must be something to balance the negative charge of the electrons. Thomson's idea was that atoms are like muffins: little raisins representing the negatively charged electrons surrounded by positively charged cake.

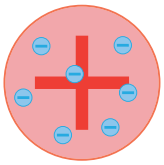

Figure A.7: Plum Pudding Model. The electrons (blue) are like negatively charged raisins in a positively charged cake (red).

A.2.2 Rutherford Model

The British scientist Ernest Rutherford and two of his students, Hans Geiger and Ernest Marsden, decided to test Thomson's model of the atom in 1910. To do so, they fired a beam of so called alpha particles in a straight line towards a thin metal foil. They knew that the alpha particles were positively charged and had a mass approximately 7300 times the mass of a single electron. When the alpha particle beam passes through the foil, the alpha particles interact with the electrons and the positively charged 'cake' of the atoms making up the metal foil. This interaction could influence the paths followed by the alpha particles. However, because the alpha particles are much heavier than the electrons, it was not expected that the electrons in the metal foil would have a serious impact on the pathways of the alpha particles. Also, the positive and negative charge inside the atom is more or less evenly distributed in Thomson's model. Hence the electrical fields inside the atoms should be small and no serious impact was expected from this interaction either. The results from the experiment were completely different though. Sometimes, as expected, the paths of the alpha particles would continue in a straight line, maybe deflected by a few degrees, after passing through the thin foil. However, in a few cases, the alpha particles would come straight backward! To quote Rutherford himself:

> *"It was quite the most incredible event that has ever happened to me in my life. It was almost as incredible as if you fired a 15-inch shell at a piece of tissue paper and it came back and hit you."* (Source: http://en.wikiquote.org/wiki/Ernest_Rutherford)

Apparently, the Thomson model of the atom must be flawed. The results of the experiment made Rutherford believe that, instead of the positive charge being distributed throughout the whole atom, the positive charge is concentrated in a tiny volume called the nucleus. Then, when an alpha particles is directed at this nucleus, it would feel a strong repulsive force when coming close to the nucleus because both are positively charged. This could explain the large angle of deflections sometimes observed. Because the nucleus is so tiny, in most cases the alpha particles would simply fly past the nucleus without much interaction. Note that because the mass of electrons is negligible compared to the mass of atoms, most of the mass of atoms should be contained inside of this nucleus.

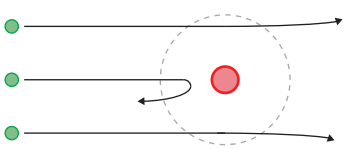

Figure A.8: Deflection of alpha particles (green). The alpha particle that is aimed at the small nucleus almost reverses direction. The paths of the other alpha particles are hardly influenced.

While the idea of an atom consisting of electrons and a nucleus may be able to explain the Rutherford experiment, it does raise new questions. Most importantly, since the nucleus is postively charged and the electrons are negatively charged, they are attracted to one another by electrostatic forces. One would expect to electrons to collapse into the much heavier nucleus. But then atoms would be of the size of their nuclei while experiments (like the Rutherford experiment) show that atoms are roughly 100.000 times larger than their nuclei! So, what prevents the electrons from collapsing into the nucleus then? Rutherford argued that electrons are actually orbiting the nucleus, just like planets orbit the sun (where the attractive force is then due to gravity). This model - the planetary model of atoms - is illustrated below.

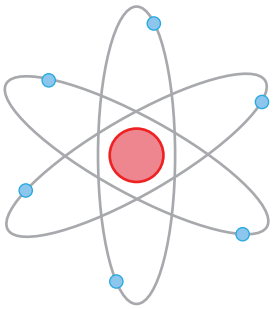

Figure A.9: Rutherford's planetary model of the atom.

There is a serious problem with this planetary model though. Why? We know that electrons are charged particles. Furthermore, orbiting particles are always accelerating: even though their speed may stay the same, their direction continuously changes. Maxwell's Equations tell us that accelerating charges radiate electromagnetic waves. Hence an electron in orbit must be constantly emitting electromagnetic waves. These waves contain energy and this energy must come from the electron. In other words, the electron is losing energy. But then the orbit of the electron must decrease over time (larger orbits correspond to larger energies). That is, the electron should be spiralling towards the nucleus. But the idea of a planetary model was to given an explanation for the fact that electrons do *not* collapse into the nucleus.

A related problem is the following. As the orbit of the electron becomes smaller and smaller, its angular velocity increases and hence it must be accelerating at faster rate. As a result, the frequency of the emitted electromagnetic waves will increase as the orbit becomes smaller. Because the electron loses energy in a continuous fashion, we would expect to see a continuous spectrum of electromagnetic waves being emitted. However, experiments show something completely different. As a simple example, think of neon lighting, which is characterized by light of a single color. Neon lighting is generated by atoms inside of a glass tube. For the particular case of neon atoms, an orange glow can be generated. Other atoms result in other colors. More generally, scientists discovered that atoms can only be made to emit electromagnetic waves at certain discrete frequencies. The frequencies are different for different kinds of atoms. But continuous spectra of emitted frequencies are never observed.

A.2.3 Bohr Model

It seems like we must reject Rutherford's planetary model of the atom. But then we are left with nothing since the Thomson model lead to problems as well. Fortunately, the Danish scientist Niels Bohr came with a possible solution. He proposed that electrons do orbit the nucleus in a certain sense, but that they can only have certain *discrete amounts of energy*, corresponding to different orbits. An electron can then jump from one orbit - or better, *energy level* - to a higher energy level by absorbing just the right amount of energy. Here the right amount of energy is of

course the energy difference between the two energy levels. Similarly, an electron can jump to a lower energy level by losing just the right amount of energy.

Now is the right time to go back to our discussion of photons. According to Einstein, light can be interpreted as coming in little energy packages called photons. When discussing the photoelectric effect, we mentioned how electrons can *absorb* photons. Electrons can also *emit* photons. The latter should not be surprising: electrons are charged particles, Maxwell's equations tell us that accelerating charged particles emit electromagnetic waves, and electromagnetic waves and photons are kind of the same thing. Note that even though we have said nothing about solar cells so far, it certainly feels as if we are getting a bit closer by discussing the interaction between photons and electrons. We will later see that this interaction is indeed crucial to the working of solar cells.

If we put together our discussion on energy levels and photons together, we conclude that electrons can jump to higher energy levels by absorbing photons and they release photons when dropping to lower energy levels. For example, if an atom has energy levels E_1 and E_2 (with $E_1 < E_2$) then an electron with energy level E_1 can jump to E_2 by absorbing a photon having energy $E_2 - E_1$. As we saw before, the energy of a photon can be expressed as hf, with f the frequency, so the electron needs to absorb a photon with frequency

$$f = \frac{E_2 - E_1}{h}$$

to make the jump. Similarly, when dropping from E_2 to E_1, the electron emits a photon of this same frequency f. If it is indeed the case that electrons in an atom can only have certain discrete energy levels, it follows directly that the electrons can only absorb / emit photons of certain discrete frequencies. Furthermore, since nature has a tendency to minimize energy, there must be some lowest energy state, a *ground state*, to prevent the electrons from falling into the nucleus by emitting photons. Electrons in energy states higher than the ground state are said to be in *excited states*. As we just saw, electrons can be *excited* by photons.

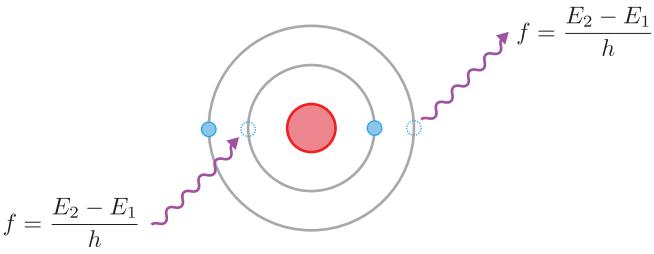

Figure A.10: Bohr Model and the interaction between photons and electrons. On the left a photon is absorbed. On the right a photon is emitted.

So that explains the discrete emission spectra and stability of the atoms. But why would nature only allow electrons to have discrete energy levels? Also, the model is not compatible with classical mechanics since, from a classical mechanics point of view, charges in orbit (like the orbiting electrons) should lose energy. So, maybe it is the case that our ideas about nature, and in particular classical mechanics, are flawed at the (sub)atomic level.

A.3 Quantum Mechanics

A.3.1 Particles as Waves

Remember the wave-particle duality of light? We said that sometimes, light behaves as waves, sometimes it behaves as particles. It was the French scientist Louis de Broglie who suggested in

1924 that maybe all matter, and in particular electrons, posses this kind of duality. This sounds counter-intuitive, but, as we will soon see, accepting matter to posses wave-like properties solves the problems encountered with previous atomic models. Furthermore, if matter can indeed behave like waves, we might be able to observe things like constructive and destructive interference for matter as well. And in fact we can. Already in 1927 were scientists able to conduct experiments which could only be explained by interference of electron waves.

If we accept the idea that matter - and when we talk about matter it will usual be electrons - has wave-like properties, then we should be able to say something about frequencies and wavelengths. Before we do so, remember that the frequency f of a photon is related to its energy E by

$$f = \frac{E}{h},$$

and that the relation between frequency and wavelength λ is given by

$$c = \lambda f.$$

Photons turn out to have momentum as well. This might feel counterintuitive since light is massless. But - and this is a consequence of the theory of special relativity as developed by Albert Eintein - mass, momentum and energy are all related to one another by the equation

$$E^2 = (mc^2)^2 + (pc)^2.$$

Here p denotes the momentum of a particle. For photons, $m = 0$ and it follows that

$$p = \frac{E}{c} = \frac{hf}{c} = \frac{h}{\lambda}.$$

This shows that photons have momentum indeed. Furthermore, we see that the wavelength and energy of a photon can be expressed as

$$\lambda = \frac{h}{p} \quad \text{and} \quad E = hf = \frac{ch}{\lambda}$$

respectively. Louis de Broglie proposed that the above two relations should hold for *any* particle, not just photons. It is important to note here that wave-like properties of matter only manifest themselves at length scales of the wavelength of matter. Since h, and therefore λ, is incredibly small in everday units, we usually don't notice the wave-like properties of matter. We certainly don't see ourself dissapearing because of destructive interference all of a sudden!

A.3.2 Schrödinger Wave Equation

Another thing about waves is that they are described by so called *wave functions*, that is, functions which satisfy some appropriate set of *wave equations*. For electromagnetic waves, the appropriate equations are in fact the Maxwell equations. For water waves, the structure of the wave equation is a bit different. As it turns out, the appropriate wave equation for matter waves is the so called *Schrödinger wave equation*.[20] For a single particle moving in a one-dimensional space, it is given by:

$$i\hbar\frac{\partial\Psi}{\partial t}(x,t) = -\frac{\hbar^2}{2m}\frac{\partial^2\Psi}{\partial x^2}(x,t) + U(x)\Psi(x,t).$$

Here Ψ is the wave function, m is the mass of the particle under consideration, $\hbar = h/2\pi$ and U is some potential energy function. Now say that we have been able to solve this partial differential equation for Ψ. Because of the complex number i in the equation, the solution Ψ itself can

[20] See Chapter 3.1 of [23] for an 'intuitive' derivation.

(and in fact must) take on complex values. This raises the question as to what this Ψ actually represents. To answer this question, remember that for electromagnetic waves we said that the intensity of the wave is proportional to the square of the amplitude. The analogue for complex 'square of the amplitude' is to take the square of the norm, i.e. $|\Psi|^2 = \Psi^*\Psi$. Then, if Ψ is the wave function for some particle, we say that the intensity of the particle is described by $|\Psi|^2$. But what is that, the 'intensity of a particle'? The most intuitive interpretation is a probabilistic one: $\int_a^b |\Psi(x,t)|^2 dx$ should be interpreted as the *probability* that the particle can be found between a and b at time t. In other words, the intensity $|\Psi|^2$ could be seen as a *probability density*. This interpretation does require the wave function to be normalized. That is, we should require that our wave functions satisfy

$$\int_{\mathbb{R}} |\Psi(x,t)|^2 dx = 1$$

at all times t.

Note that this description of matter is fundamentally different from what we are used to from classical mechanics. Indeed, in classical mechanics we describe particles by saying where they are and where they are going. In quantum mechanics, we describe particles in terms of wave functions and these wave functions can at most give us probabilities of the particle being somewhere. Now one can imagine the wave function for a particle to become localized in the sense that the particle can only be found with positive probability in a very tiny region. Then we can say 'for sure' that the particle is in that tiny area. And if we could visualize the wave function, we could say where the particle is going by looking in which way the wave if propagating. While that is true, nature places a fundamental limit on how far we can go in this respect. I'm talking about the so called Heisenberg uncertainty principle here. This principle says that is impossible to find simultaneously the position *and* the momentum of a particle to arbitrary precision. More precisely, if Δx denotes the uncertainty in the position of a particle and Δp the uncertainty in the momentum, then, no matter what,

$$\Delta x \Delta p \geq \frac{h}{4\pi}.$$

As a consequence, the more we localize the position in space where a particle might be found, the more difficult it becomes to say where it might be going next and vice versa. We are going to need this uncertainty principle in a moment.

To get a better physical understanding for the complicated looking Schrödinger wave equation itself it is convenient to first set the potential energy function U equal to zero. In other words, we are going to look at a *freely* moving particle first. I claim that the function

$$\Psi(x,t) = e^{ipx/\hbar} e^{-iEt/\hbar},$$

satisfies the Schrödinger wave equation with $U \equiv 0$. Indeed, by working out the partial derivatives we see that

$$i\hbar \frac{\partial \Psi}{\partial t}(x,t) = i\hbar \frac{\partial}{\partial t}\left[e^{ipx/\hbar} e^{-iEt/\hbar}\right] = E e^{ipx/\hbar} e^{-iEt/\hbar} = E\Psi(x,t)$$

and

$$-\frac{\hbar^2}{2m}\frac{\partial^2 \Psi}{\partial x^2}(x,t) = -\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2}\left[e^{ipx/\hbar} e^{-iEt/\hbar}\right] = \frac{p^2}{2m} e^{ipx/\hbar} e^{-iEt/\hbar} = \frac{p^2}{2m}\Psi(x,t).$$

The term $p^2/2m$ corresponds to the kinetic energy of the particle. Since $U \equiv 0$, the particle cannot have any potential energy. Hence its total energy is precisely its kinetic energy. That is,

$$E = p^2/2m$$

and we see that the Schrodinger wave equation is satisfied.[21]

If we do not assume the potential energy U to be zero, we can still try to see what happens when we plug in this solution. What happens is that the Schrödinger wave equation reduces to the following form:

$$E\Psi = (K + U)\Psi.$$

This suggests that the Schrödinger wave equation is the quantum mechanical way of saying that total energy E of a particle is given by the sum of its kinetic energy K and its potential energy U.

In general, the function $\Psi(x,t) = e^{ipx/\hbar}e^{-iEt/\hbar}$ will not satisfy the Schrödinger wave equation for non-zero U. However, it is still useful to look for solutions of the form

$$\Psi(x,t) = \psi(x)e^{-iEt/\hbar}.$$

The reason is that the Schrödinger wave equation then transforms into the time-independent form

$$E\psi(x) = -\frac{\hbar^2}{2m}\frac{\partial^2\psi}{\partial x^2}(x) + U(x)\psi(x). \tag{A.1}$$

Note that (A.1) is in fact an eigenvalue problem for the operator $-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2} + U$. We have to solve (A.1) simultaneously for *eigenvalues* E and corresponding *eigenfunctions* $\psi(x)$. If $\psi(x)$ is an eigenfunction with corresponding eigenvalue E, then $\Psi(x,t) = \psi(x)e^{-iEt/\hbar}$ is a solution to the full time-independent Schrödinger wave equation. Now observe that

$$\int |\Psi(x,t)|^2 dx = \int |\psi(x)e^{-iEt/\hbar}|^2 dx = \int |\psi(x)|^2 dx.$$

This tells us that the probability of finding the particle in a specific region in space is constant in time. The wave function $\Psi(x,t) = \psi(x)e^{-iEt/\hbar}$ is then said to be a *stationary state*.

A.3.3 Particle in a Box Model for Hydrogen Atoms

Remember that we were trying to find a model to describe how atoms work. Let us see if all this quantum mechanical stuff got us any closer. To this end, consider a hydrogen atom. A hydrogen atom is the simplest of all atoms. It consists of a nucleus and just one electron. The electron is bound to the nucleus by electrostatic forces. We model this by saying the that nucleus sets up a potential well for the electron. The potential well is described by a potential U which we suppose to be zero inside the domain $(0, L)$ (for some $L > 0$) and infinite outside of this domain. In other words, the electron is trapped inside the box $(0, L)$. We could imagine the nucleus to be positioned at $x = L/2$. Note that this model is an extreme oversimplification of reality, but it makes calculations easy and as we will see it will already lead to interesting results. If we assume the electron to have mass m and to be in a quantum state of energy E, then we can then write the system to be solved as:

$$\begin{aligned}
E\psi(x) &= -\frac{\hbar^2}{2m}\frac{\partial^2\psi}{\partial x^2}(x) \quad \text{for } x \in (0,L),\\
\psi(0) &= 0,\\
\psi(L) &= 0,\\
\int_{\mathbb{R}} |\psi(x)|^2 dx &= 1 \quad \text{for all } t \geq 0.
\end{aligned}$$

[21] The equation $E = p^2/2m$ also tells us that p is constant. Then $\Delta p = 0$ and the Heisenberg uncertainty principle tells us that Δx must be infinite. Indeed, from the fact that $|\Psi(x,t)|^2 = 1$, we see that the particle can be found anywhere in space with equal probability. Also note that the wave function cannot be normalized in this case!

If ψ is a solution to this system, then $\Psi(x,t) = \psi(x)e^{-iEt/\hbar}$ solves the time-dependent problem.

To actually solve the system, note that we are dealing with a linear ordinary differential equation. Its characteristic equation is given by

$$\frac{\hbar^2}{2m}\lambda^2 + E = 0.$$

Solving for λ gives

$$\lambda = \pm\sqrt{-\frac{2mE}{\hbar^2}} = \pm i\sqrt{\frac{2mE}{\hbar^2}}.$$

From theory on linear ODE's we know that the general solution takes the form

$$\psi(x) = C_1 \exp\left(i\sqrt{\frac{2mE}{\hbar^2}}x\right) + C_2 \exp\left(-i\sqrt{\frac{2mE}{\hbar^2}}x\right),$$

where C_1 and C_2 are two (complex) constants. By plugging in the boundary condition $\psi(0) = 0$ we find that $C_2 = -C_1$. Hence

$$\begin{aligned}
\psi(x) &= C_1 \exp\left(i\sqrt{\frac{2mE}{\hbar^2}}x\right) - C_1 \exp\left(-i\sqrt{\frac{2mE}{\hbar^2}}x\right) \\
&= C_1\left[\cos\left(\sqrt{\frac{2mE}{\hbar^2}}x\right) + i\sin\left(\sqrt{\frac{2mE}{\hbar^2}}x\right)\right] - C_1\left[\cos\left(\sqrt{\frac{2mE}{\hbar^2}}x\right) - i\sin\left(\sqrt{\frac{2mE}{\hbar^2}}x\right)\right] \\
&= 2C_1 i\sin\left(\sqrt{\frac{2mE}{\hbar^2}}x\right).
\end{aligned}$$

To satisfy the boundary condition $\psi(L) = 0$, we must have that

$$\sqrt{\frac{2mE}{\hbar^2}}L = n\pi \quad \text{for some } n \in \mathbb{Z}.$$

We will come back on this *very important* issue in a moment. For now, assume the condition is satisfied for some $n \in \mathbb{Z}$ and let ψ_n denote the associated wave function. The value of C_1 can be found by using the requirement that ψ_n be normalized. One finds that $|C_1| = \sqrt{1/2L}$ and we might as well choose $C_1 = \sqrt{1/2L}$ (choosing C_1 to be a different complex number with the same norm only introduces a phase shift in the solution). Our time-independent solution can now be written as

$$\psi_n(x) = \sqrt{\frac{2}{L}} i\sin\left(\frac{n\pi}{L}x\right).$$

We sketch a few solution in Figure A.11

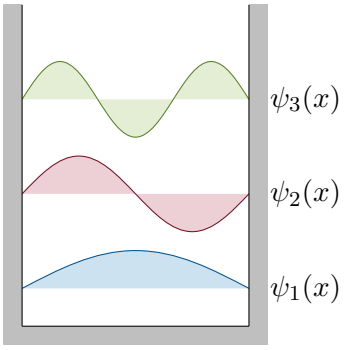

Figure A.11: Wave functions for a particle in a box.

We see that ψ_n is basically a sine wave with $n-1$ nodes. Let Ψ_n denote the time-dependent solution associated with ψ_n. Sketching Ψ_n is a bit more difficult since it can have both real and imaginary parts at the same time (whereas ψ_n only has an imaginary part). Both the real and the imaginary parts form standing waves. [22] This is completely different from the classical mechnical way of viewing a particle in a box: a particle bouncing around with some definite momentum.

Anyway, the real reason I wanted to work out the solution to the particle-in-a-box system is that it shows where the discrete energy levels in atoms come from. Remember that one of the conditions that had to be fulfilled was that

$$\sqrt{\frac{2mE}{\hbar^2}}L = n\pi \quad \text{for some } n \in \mathbb{Z}.$$

In other words, the system can only be solved for an electron having one of the energies

$$E_n = \frac{n^2\pi^2\hbar^2}{2mL^2}, \quad n \geq 0.$$

The energy levels $\{E_n\}_{n\geq 0}$ form a discrete set, as is needed to explain the discrete emission of atoms. In the Bohr model of the atom, this had stated explicitly but we had no clue where it came from. But if we work within a quantum mechanical framework, where all matter is described in term of wave functions, the discrete energy levels present themselves automatically.

But wait. So far we have only fully solved the time-independent equation. What about the full time-dependent equation? Maybe we were a bit too quick too conclude that only discrete energies are allowed. Luckily, as it turns out, the stationary states Ψ_n form a basis for the space of solutions to the Schrodinger wave equation.[23] That is, any solution Ψ can be written as a linear combination of the stationary states:

$$\Psi(x,t) = \sum_{n\geq 0} \alpha_n \Psi_n(x,t) = \sum_{n\geq 0} \alpha_n \psi_n(x) e^{-iE_n t/\hbar} \quad (\alpha_n \in \mathbb{C}, x \in (0,L), t \geq 0).$$

But what about the energy levels of such general solutions? In order to say something about this, we will show that the ψ_n are mutually orthogonal functions first. Indeed, from the double angle formulas one can derive that $\sin(a)\sin(b) = \frac{1}{2}\left(\cos(a-b) + \cos(a+b)\right)$. Using this relation we then see that for $n \neq m$

$$\begin{aligned}
\int_{\mathbb{R}} \psi_n(x)\psi_m(x)dx &= -\frac{2}{L}\int_0^L \sin\left(\frac{n\pi}{L}x\right)\sin\left(\frac{m\pi}{L}x\right)dx \\
&= -\frac{1}{L}\int_0^L \left(\cos\left(\frac{(n-m)\pi}{L}x\right) - \cos\left(\frac{(n+m)\pi}{L}x\right)\right)dx \\
&= -\frac{1}{L}\left[\frac{L}{(n-m)\pi}\sin\left(\frac{(n-m)\pi}{L}x\right)\right]_0^L - \frac{1}{L}\left[\frac{L}{(n+m)\pi}\sin\left(\frac{(n+m)\pi}{L}x\right)\right]_0^L \\
&= 0.
\end{aligned}$$

Using the mutual orthogonality condition together with the fact that each ψ_n is normalized, we

[22] The standing waves are in visualised in B, C and D of http://en.wikipedia.org/wiki/Particle_in_a_box#mediaviewer/File:InfiniteSqua (see B, C and D).

[23] Interestingly enough, for my Bachelor's thesis [53] I proved a spectral theorem for unbounded self-adjoint operators. At the time I had had no clue why anyone would care. Now an application has presented itself.

see that

$$\begin{aligned}\int_{\mathbb{R}} \Psi^*(x,t)\Psi(x,t)dx &= \int_{\mathbb{R}} \left[\sum_{n\geq 0} \alpha_n^* \psi_n^*(x) e^{iE_n t/\hbar}\right] \left[\sum_{n\geq 0} \alpha_n \psi_n(x) e^{-iE_n t/\hbar}\right] dx \\ &= \int_{\mathbb{R}} \left[\sum_{n\geq 0} |\alpha_n|^2 |\psi_n(x)|^2\right] dx \\ &= \sum_{n\geq 0} \left[|\alpha_n|^2 \int |\psi_n(x)|^2 dx\right] \\ &= \sum_{n\geq 0} |\alpha_n|^2.\end{aligned}$$

Because our wavefunctions should be normalized, it follows that $\sum_{n\geq 0} |\alpha_n|^2$ should be equal to 1.

A fundamental postulate of quantum mechanics says that anything that is *observable* is represented by a (Hermitian) operator acting on the (Hilbert) space of states. Examples of such observables are position, momentum and energy. When discussing the Heisenberg uncertainty principle, we saw that on a quantum scale particles don't have an exact position or momentum. The only thing we can hope to find by measurement is the *expectation value* of such an observable. In general, if F is an observable with corresponding operator $\hat{F}$ acting on some state Ψ, then its expectation value (or, the observed value) is given by [24]

$$\left\langle \hat{F} \right\rangle = \int_{\mathbb{R}} \Psi^*(x,t) \hat{F} \Psi(x,t) dx.$$

In our case, the energy operator $\hat{E}$ is given by $\hat{E} = i\hbar \frac{\partial}{\partial t} = -\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2} + U$. The expectation value

[24] How does this relate to the expectation value known from probability theory? Remember that if ρ is the probability density for some continuous random variable X then the expected value $\langle X \rangle$ is given by $\langle X \rangle = \int_{\mathbb{R}} x\rho(x)dx$. If we interpret X as an operator which multiplies by x, then this could be written as

$$\langle X \rangle = \int_{\mathbb{R}} (X\rho^{1/2})(x) = \int_{\mathbb{R}} \left(\rho^{1/2}\right)^* (x)(X\rho)(x)dx$$

(note that ρ is real and positive so taking square roots is fine and $\rho = \rho^*$). We see that the given definition of the expectation value of an observable is a generalization of the expectation value for random variables known from probability theory.

$\langle E \rangle$ is then found to be

$$\begin{aligned}
\langle E \rangle &= \int_{\mathbb{R}} \Psi^*(x,t) \left[-\frac{\hbar^2}{2m} \frac{\partial^2}{\partial x^2} + U \right] \Psi(x,t) dx \\
&= \int_0^L \Psi^*(x,t) \left[-\frac{\hbar^2}{2m} \frac{\partial^2 \Psi}{\partial x^2}(x,t) \right] dx \quad (U \equiv 0 \text{ inside the box}) \\
&= \int_0^L \left[\sum_{n \geq 0} \alpha_n^* \psi_n^*(x) e^{iE_n t/\hbar} \right] \left[\sum_{n \geq 0} -\alpha_n \frac{\hbar^2}{2m} \frac{\partial^2 \psi_n}{\partial x^2}(x) e^{-iE_n t/\hbar} \right] dx \\
&= \int_0^L \left[\sum_{n \geq 0} \alpha_n^* \psi_n^*(x) e^{iE_n t/\hbar} \right] \left[\sum_{n \geq 0} \alpha_n E_n \psi_n(x) e^{-iE_n t/\hbar} \right] dx \\
&= \int_0^L \left[\sum_{n \geq 0} |\alpha_n|^2 E_n |\psi_n(x)|^2 \right] dx \quad (\text{orthogonality condition}) \\
&= \sum_{n \geq 0} \left[|\alpha_n|^2 E_n \int_0^L |\psi_n(x)|^2 dx \right] \\
&= \sum_{n \geq 0} |\alpha_n|^2 E_n. \quad (\psi_n \text{ is normalized}).
\end{aligned}$$

Recall that the expectation value of a discrete random variable X is given by

$$\sum_x P(X = x) x.$$

Here $P(X = x)$ is the probability that X takes on the value x and $\sum_x P(X = x) = 1$. Comparing this with the expression $\sum_{n \geq 0} |\alpha_n|^2 E_n$ for the expected energy reveals that, for each $n \geq 0$, we can interpret $|\alpha_n|^2$ as being the probability of observing the energy E_n. Then it follows that energies other than the $\{E_n\}$ are *never observed*. So, even though electrons can be in superpositions of states with energies, when doing measurements - and the experiments which showed discrete spectra for atoms count as measurements in this respect - we only observe them to have energies from the set $\{E_n\}$. Finally we can say that our quantum mechanical model has correctly predicted that only discrete energies are allowed for hydrogen atoms.

So far we have not excluded the case $n = 0$. Note that for $n = 0$ the energy E_n of the electron is zero. But is it actually possible for the electron to have zero energy? The answer is no. Remember that inside the box, the potential is zero so all of the energy of the electron is kinetic energy. That is, $E = p/2m$. If $E = 0$, then $p = 0$ as well and we have no uncertainty with respect to the momentum of the electron. But then the Heisenberg uncertainty principle tells us that the uncertainty with respect to the position of the electron should be infinite. But that cannot be the case, since our electron is confined to a finitely sized box. Hence the electron always has some non-zero energy. In the lowest energy state, the ground state, the electron has an energy of E_1. If we go back to the Bohr model again, we had to assume the existence of such a *ground state*. It could not be further justified. But, again, within quantum mechanics, this result comes automatically.

Note that our model of the hydrogen atom was extremely simplified. A better model would have to be three-dimensional to begin with. Also, the potential energy funtion should not be some infinite square well but it should describe the electric field set up by the positively charged nucleus. And we would also have to take into account the wave-like behaviour of the nucleus and other possible interactions between the nucleus and the electron. But these things only make the mathematics more complicated. In the end, just like for our simple model, one will find that the

electron can only have discrete energies within the atom. Unlike the above case, there is only a finite amount of energy levels the electron can have within the atom (outside of the atom the electron can have any energy, just like a free electron). And as it turns out, the energy levels predicted by quantum mechanics match to high accuracy with energy levels found in experiments with emission spectra and such. In the end, the crazy assumption that all matter acts as waves may not be so crazy after all.

A.3.4 Quantum Numbers and Pauli Exclusion Principle

In the above discussion we have only considered the hydrogen atom. The hydrogen has only a single electron. As the amount of electrons in an atom increases, the complexity of setting up and (numerically) solving the appropriate Schrödinger wave equations increases as well. But again, discrete energy levels are found. In the case of a hydrogen atom, we could describe the allowed states with just one number n, the so called *principal quantum number*. In general, the allowed states are described by four quantum numbers: the principal quantum number n, the angular quantum number l, the magnetic quantum number m and the spin quantum number s. The principal quantum number n determines the energy of the electrons and it also tells us something about the *size* of the volume where an electron may be found (with high probability). This volume is referred to as the *orbital* of an electron.[25] The angular quantum number l says something about the *shape* of the orbital and the magnetic quantum m says something about the *orientation* of the orbital.

Since nature tends to minimize (free) energy, one would expect all the electrons to be in the lowest energy state. This turns out not to be possible. Nature places a limit on how many electrons can share an orbital. We refer to this as the *Pauli exclusion principle*. And really, it is not such a weird principle. It is like filling a bag with marbles. The marbles try to minimize their gravitional potential energy by going to the bottom of the bag. But they can't go all to the bottom. The marbles will start to pile up in higher and higher energy levels. For electrons in an atom, it turns out that each orbital can only be occupied by at most *two* electrons. We use the quantum spin number s to distinguish between two electrons in the same orbital. This means the number s can in fact only take on two different values. The quantum state of an electron can now be uniquely specified by the four numbers (n, l, m, s). Electrons in a quantum state for which n is maximal are said to be *valence electrons*. Generally speaking these electrons are most likely to be found at a greater distance from the nucleus than the other electrons and they have the highest energies. It is the valence electrons that participate in the bonding of atoms. We will see how that works in the next section.

First, let us summarize what we have seen so far. Experiments have shown that electrons can jump between discrete energy levels by absorbing and emitting photons. One can guess that this interaction between electrons and photons is crucial to the understanding of solar cells. But classical mechanics failed in giving us a proper understanding of the underlying physics. Following the wave-particle duality of light, it was proposed that matter might posses a wave-particle duality as well, leading us into the realm of quantum mechanics. The wave-like properties of small particles like electrons are governed by the Schrödinger wave equation. By solving this wave equation for a simplified model of a hydrogen atom we were able to predict the discrete energy levels as observed in experiments.

[25] For the particle in a box, it can be seen from the sketches of the wave functions that the electron in the ground state $n = 1$ is most likely to be found in the middle of the box where the nucleus is also supposed to be. That is, the orbital is small. For higher energy states, the probability to be close to the nucleus decreases while the probability to be further away from the nucleus increases. In other words, the orbital becomes larger.

A.4 Band Theory and Semiconductor Devices

A.4.1 Splitting of Energy Levels

So far we have only discussed isolated atoms. But atoms tend to bond with other atoms to form new and larger structures, eventually forming everything around us. Since the allowed energy levels of electrons in atoms are fundamental for understanding the interaction between electrons and photons - which is in turn fundamental for understanding solar cells - we want to know what happens to the electrons as atoms bond into *solid materials*.

As an illustrative example of what can happen we start with two hydrogen atoms initially separated. Each of them has a single electron in the ground state described by a wave function $\Psi_i, i = 1, 2$. Both electrons are at the same energy level. This situation is depicted in the left side of Figure A.12. When the atoms get closer together, they start interacting with each other. This interaction needs to be taken into account when setting up the Schrödinger wave equation. The reader is referred to Chapter 2.7 of Ferry [20] for details on how this can be done. In the end, when solving the Schrödinger wave equation one finds that the original ground states have split into two new states: one with an energy lower than the original ground states and one with an energy lower than the original ground states. To get a better understanding of why this happens, consider the two wave functions Ψ_1 and Ψ_2 as illustrated in Figure A.12 again. As the atoms are brouht closer together, the wave functions of the two electrons start to overlap. The wave function describing the two electrons will be a mixture of the two original wavefunctions. There are basically two ways in which the mixing can occur. Either the two wave functions are *added* together, or one is *subtracted* from the other (and in both cases we have to do this with certain scaling constants to make sure that the final wave function is normalized). Let the final wave functions be denoted by Ψ_+ and Ψ_- respectively. They are sketched in the right of Figure A.12.

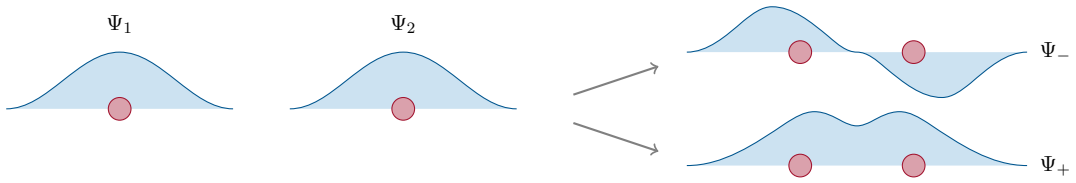

Figure A.12: Splitting of energy levels as two hydrogen atoms move closer together.

In the situation described by Ψ_+ the electrons prefer to be in the same area in between the two nuclei. In fact the electrons make the atoms stick together and we say that the electrons are in the *bonding state*. The energy level associated with the bonding state is *lower* than the energy level of the original ground state.

The situation described by Ψ_- is different. The two electrons are most likely to be found on the outside, one on the left and one on the right. It is as if the electrons are trying to pull the atoms apart. We will say that the electrons are in the *anti-bonding state*. The energy level associated with the anti-bonding state is *higher* than the energy level of the original ground state.

While originally the two electrons could only be in a single state (the ground state of their respective hydrogen atom) with a single energy level, we see that after bringing the hydrogen atoms together there are now two possible states for the electrons: the bonding and the anti-bonding state, each having their own energy level. The Pauli exclusion principle tells us that each of these states can hold at most two electrons.[26]

[26] Since each hydrogen atom has only one electron - and nature tends to minimize energy - the two bonding states will be filled when bringing two hydrogen atoms together. That is, it is energetically favorable for the hydrogen atoms to bond and form a H_2 molecule. For helium atoms that have two electrons, both the bonding and the

Let us generalise the above. Suppose we start with $2N$ hydrogen atoms. Initially, the atoms are separated and each electron is in the ground state of its own atom. All electrons have the same energy. In other words, the *system* has only one state. This state is filled with $2N$ electrons (note that this does not violate the Pauli exclusion principle since we are dealing with $2N$ atoms here, the state of the system can be filled with $2N$ electrons and not just two). Now we bring the atoms together. Like in the above example, the single state of the system will split into $2N$ different states. Half of them, N, will be bonding states and the energy associated with these states is lower than the energy of the original state. The other half will be anti-bonding states, where the energies are higher than before. Usually N is very large and the energy levels will be spaced so closely together that the splitting results in two *continuous energy bands*. The energy difference between the lowest anti-bonding state and the highest bonding state is called the *band gap*. Note that in our system, no states are allowed which have an energy in this band gap.

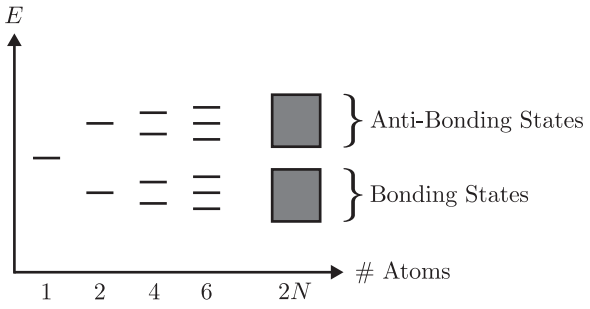

Figure A.13: Splitting of energy levels into energy bands.

For atoms other than hydrogen atoms the situation is more complex because then each atom has more than one electron and all these electrons will interact with one another. But as mentioned earlier, the outermost electrons, the valence electrons, are the ones that play the most important roles in bonding of atoms. Because valence electrons are characterized by having the same, maximal value for principal quantum number n they also have (more or less) the same energies. Just like with hydrogen atoms the energy levels associated with valence electrons will split into continuous energy bands as the atoms are brought closer together and there will be a band gap in between. In other words, when we only focus on valence electrons, what happens for general atoms is similar to what happens for hydrogen atoms.

So far we have said nothing yet about which of the states, the bonding or the anti-bonding states, will be occupied by the electrons. We have mentioned before that nature tends to minimize energy, and taking into account the Pauli exclusion principle, it is expected that the electrons will try to fill the bonding states first before filling up anti-bonding states. Generally speaking, this is indeed what happens but there are a few different flavors here. A useful concept in discussing this issue is the so called Fermi-Dirac distribution. It tells us for each energy level E - assuming this energy is allowed within the material - what fraction of the states with this energy is occupied at a given temperature T. More specifically, the Fermi-Direc distribution is given by

$$f(E) = \frac{1}{e^{(E-E_F)/kT} + 1}.$$

Here k is the so called Boltzmann constant. The energy E_F is the *Fermi energy* or the *Fermi level*. It represents the - possibly hypothetical - energy level that is half-occupied ($f(E_f) = 1/2$).

anti-bonding states are filled when bringing the two atoms closer together. From an energy point of view there is no reason for helium atoms to bond. And in fact they don't: helium is considered a noble gas.

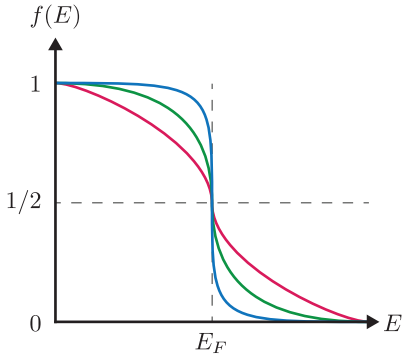

Figure A.14: Fermi-Dirac Function for three different temperatures. The blue line corresponds to the lowest temperature, the red line to the highest temperature.

Solid materials are classified into three different groups when it comes to energy band structures. They will be explored below. On a side note, it will be easier from now on to think of electrons as being small particles again. We needed their wave-like behaviour to get a better understanding of the origin of discrete energy levels and band gaps. Now that we do, we can more or less forget about the underlying framework and just take the energy levels and bands for granted. With this in mind, let's classify the different groups of solid materials.

A.4.2 Insulators, Conductors and Semiconductors

- Materials for which the band gap is relatively large are called *insulators*. At absolute zero temperature, all of the electrons will be in bonding states. No electrons will be in anti-bonding states. More specifically, the highest occupied energy band is completely filled. An energy band that is (almost) completely occupied with electrons is referred to as *valence band*. The next higher energy band is referred to as *conduction band*. If we increase the temperature of the insulator to non-zero temperatures, the electrons will gain thermal energy and some of them will be able to break free from their bonds and jump from the valence band into the conduction band.[27] Because the band gap is 'large', only a tiny fraction of the electrons will be able to do so. The Fermi level for an insulator is (more or less) halfway between the valence band and the conduction band. The reason is that each time an electron jumps into the conduction band, the valence band loses one. We want the Fermi-Dirac distribution to reflect this antisymmetric behaviour and that is achieved by placing the Fermi level in the band gap. Note that the Fermi level does not present a state that can actually be occupied by electrons!

- *Semiconductors* share the property with insulators that at absolute zero temperature they have a completely filled valence band and a completely empty conduction band. However, for semiconductors the band gap is much smaller compared to the band gap of insulators. It is much easier for the electrons to make the jump into the conduction band. In fact, the band gap is small enough that electrons can jump into the conduction band by absorbing photons coming from the sun. That is why semiconducting materials are key ingredients

[27] Even though nature tends to minimize energy, some electrons will still be in the conduction band. The reason is that nature in facts tends to minimize *free energy*. In situations of constant pressure this is the *Gibbs free energy* G given by $G = E - TS$. Here S is entropy of the system and T is the temperature. For $T = 0$, the energy G is minimized by minimizing E. As T increases, the entropy contribution becomes more and more important. The entropy associated with a semiconductor having some electrons in the conduction band is much higher than the entropy associated with a semiconductor having an empty conduction band because the former can occur in many different ways while the latter can occur in only one way. Therefore, even at room temperature, G will be minimized by having some electrons in the conduction band.

for solar cells. Just like for insulators, the Fermi level for a semiconductor is approximately halfway between the valence band and the conduction band.

- Lastly, *conductors* are materials for which the energy bands resulting from bonding and anti-bonding states overlap. So, strictly speaking there is no band gap. At absolute zero temperature the highest occupied energy band will be partly filled and this energy band is in fact a conduction band. The Fermi level lies within this conduction band.[28]

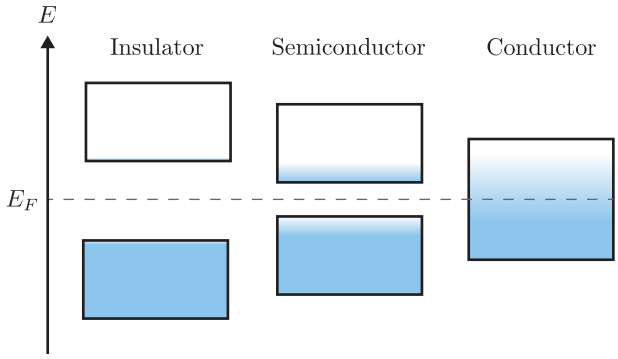

Figure A.15: Band gaps and Fermi levels of insulators, conductors and semiconductors at room temperature. The more 'blue' in a band, the more electrons there are present. A full band (traffic jam) is not good for conduction, nor is an empty band (empty road).

As the names suggest, insulators are not good at conducting electrical current while conductors are. But how exactly is this related to the energy bands? Remember that electric current is the flow of electric charge. In our case, the charge will be carried by the electrons. To get the electrons moving, we apply an electric field. This electric field is supposed to force the electrons to move within the material. But, for the atoms to move that would mean they have to go into different quantum states. For insulators, the valence band is almost completely filled. There is hardly any room for electrons to move into different quantum states within the valence band. It is as if the electrons are in a traffic jam. However, if an electron is able to jump into the desolated conduction band, it will be able to move around freely through the material because there are many states with similar energy levels available. But, like we said, it is really difficult for electrons in an insulator to make this jump. Hence there will be little to no current flowing. For conductors, because of the overlapping energy bands, there are always plenty of quantum states with comparable energies for an electron to jump in to. Hence even a small electric field can generate a large flow of current.

The conductivity of semiconductors is somewhere in between. What makes them special though is that their conductivity is highly sensitive to so called *impurities*. Let's have a look at that.

A.4.3 Doping of Semiconductors

An important semiconductor material is silicon. Each silicon atom has four valence electrons. Each of these electrons can participate in the bonding with one other atom. Given the right

[28] Remember how we said that the photoelectric effect was observed for most *metals*? With our current understanding of energy bands we can see why this is no coincidence. Electrons in the conduction band of a material are still bound to the material. To get them out of the material, the electrons need to get into an energy level that is above the conduction band. For insulators the energy from photons is not even enough to get the electrons into the conduction band, let alone get them out of the material. For conductors - and most metals are good conductors - electrons will be in the conduction band even without external energy sources. By absorbing energy from photons one could imagine the electrons in the conduction band to gain enough energy to break free from the material. And that is precisely the photoelectric effect.

conditions, the silicon atoms will organize themselves in such a way that each silicon atom forms bonds with four other silicon atoms. This can be depicted as follows.

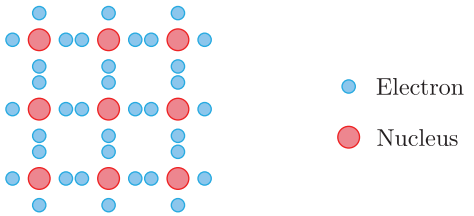

Figure A.16: Atomic arrangement in a perfect crystalline silicon semiconductor

We see that the atoms have arranged themselves in a highly order periodic structure that extends into all directions. We will refer to such structures as *crystalline* structures.

In Figure A.16, all of the electrons are in the bonding state. But we said that in semiconductors, even at room temperature, some of the electrons will be able to jump break free from the bonds and jump into the conduction band. When an electron has made such a jump, it leaves behind a *hole* in the valence band. A neighbouring electron in the valence band can easily jump in this hole (the energy levels are pretty much the same). In doing so, it leaves behind a hole itself and we could say that there is a hole travelling through the material. We could interpret such holes as *positive charge carriers*. Electrons are referred to as *negative charge carriers*. Both the electrons in the conduction band and the holes in the valence band contribute to the conductivity of a semiconductor. Note that the amount of electrons in the conduction band is equal to the amount of holes in the valence band. We refer to such a semiconductor as an *intrinsic semiconductor*.

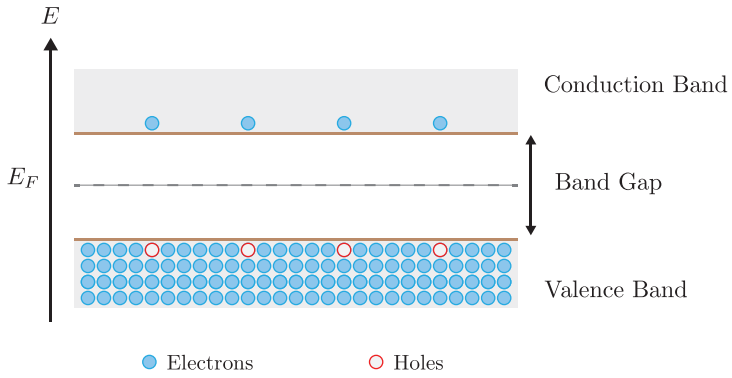

Figure A.17: A typical intrinsic semiconductor at room temperature. A small amount of electrons have made the jump into the conduction band, leaving behind holes in the valence band.

Suppose that instead of creating a material out of pure silicon, we add some *impurities* to the material through a process called *doping*. There are two different flavours here.

- First, we add a small amount phosphorus atoms to the mix. Phosphorus atoms have *five* electrons in their valence shell, one more than silicon atoms. At absolute zero temperature, only four of the five valence electrons can participate in the bonding of the material. What about the energy level of the remaining *donor electrons*? Such electrons do not participate in bonding so their energy is higher than the energy of electrons in the valence band. But, they are still bound to their respective nuclei. Electrons in the conduction band are only bound to the material itself, not to any particular atom within the material. Therefore, the energy level of the donor electrons will be below - but close to - the conduction band.

We call this energy level the *donor level* and the phosphorus atoms are called *donor atoms*. Note that the donor level lies within the previously forbidden band gap. Because electrons in the donor level are still bound to their respective nucleui, they hardly contribute to current flow when an electric field is applied. However, because the jump from the donor level to the conduction band is much smaller than the jump from the valence band to the conduction band, even at room temperature more or less all of these donor electrons will be able to make the jump in the conduction band. The introduction of donor atoms breaks the symmetry between the number of electrons in the conduction band and the number of holes in the valence band. As a result, the Fermi level is shifted towards the conduction band. It will be somewhere between the donor level and the bottom of the conduction band.

When donor electrons make the jump into the conduction band, they don't leave behind a hole in the valence band because the electrons were not participating in any bonding between atoms. They do *ionize* their donor atoms though. That is, the donor atoms were electrically neutral at first because the negative charges of the electrons and the positive charge of the nucleus cancelled out. But after losing an electron to the conduction band, the donor atoms are left with a positive charge. They are ionized. Overall the semiconductor is still electrically neutral though.

A semiconductor that has been doped with donor atoms is referred to as an *n-type semiconductor*. It has a lot of electrons in the conduction band that can contribute to current flow under the influence of an electric field. Note that, as mentioned before, even at room temperature some of the electrons in the valence band will be able to jump into the conduction band. In doing so, they leave behind holes. The holes can contribute to current flow as well. However, the amount of holes in the valence band is neglible compared to the amount of electrons in the conduction band for n-type semiconductors. We will say that the electrons are the *majority charge carriers* whereas the holes are the *minority charge carriers*.

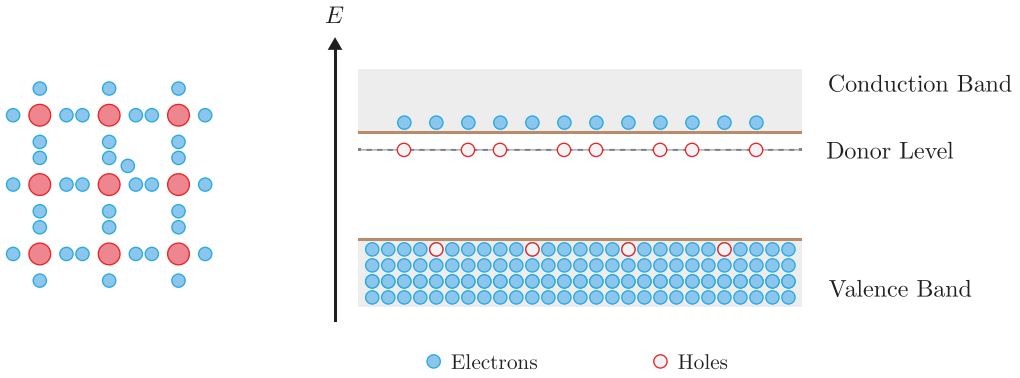

Figure A.18: (Left) Atomic configuration of an n-type semiconductor. One of the atoms has five valence electrons. (Right) Electron/hole configuration in an n-type semiconductor at room temperature. All of the electrons originally in the donor level have jumped into the conduction band (majority charge carriers). Only a few electrons have jumped from the valence band into the conduction band, leaving behind only a few holes (minority charge carriers).

- In a similar fashion, we can dope a silicon semiconductor with atoms that have only *three* valence electrons, like boron atoms. Then, even at absolute zero temperature, there will be holes in the valence band without there being electrons in the conduction band. We interpret the holes as positive charge carriers and we can associate an energy level with them. Like with donor atoms, the energy level of the holes lies within the band gap. This

time, it lies closer to the valence band and we refer to it is the *acceptor level*. The boron atoms itself will be referred to as *acceptor atoms*. The introduction of acceptor atoms breaks the symmetry between the number of electrons in the conduction band and the number of holes in the valence band. Only this time the Fermi level is shifted towards the valence band. It will be somewhere between the acceptor level and the top of the valence band.

When a hole is filled with an electron (like a valence electron), the respective acceptor atom is ionized: it has acquired a net negative charge. At room temperature, more or less all of the acceptor atoms will be ionized (but overall the semiconductor is still electrically neutral). We could interpret this as holes jumping down from their original energy level into the valence band, just like electrons from donor atoms in an n-type semiconductor jump into the conduction band.

A semiconductors that has been doped with acceptor atoms is referred to as a *p-type semiconductor*. It will have much more holes in the valence band than electrons in the conduction band. Therefore, the holes are the *majority charge carriers* in this case. The electrons, a few of which have been able to jump into the conduction band, are the *minority charge carriers* this time.

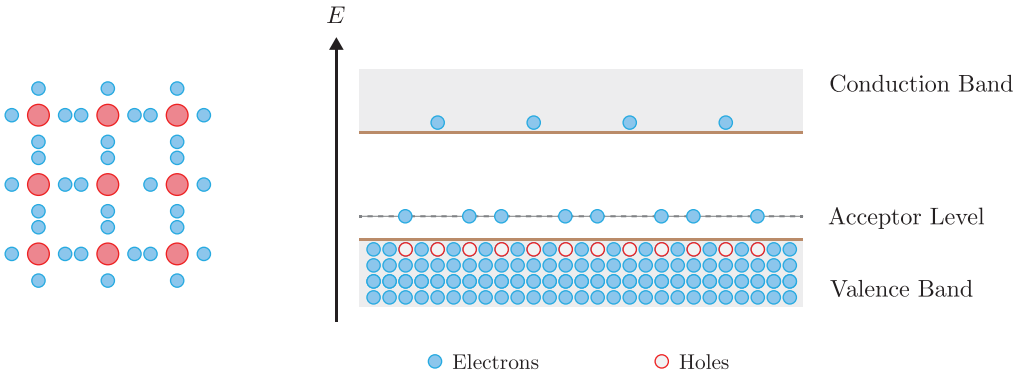

Figure A.19: (Left) Atomic configuration of a p-type semiconductor. One of the atoms has only three valence electrons. (Right) Electron/hole configuration in an p-type semiconductor at room temperature. All of the holes originally in the accepted level have falled into the valence band (majority charge carriers). Only a few electrons have jumped from the valence band into the conduction band (minority charge carriers).

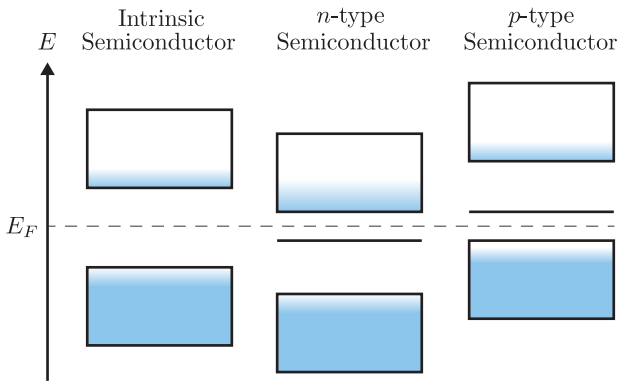

Figure A.20: The energy bands of intrinsic and doped semiconductors relative to the same Fermi level. The solid black lines represent the donor level (for the n-type semiconductor) and acceptor level (for the p-type semiconductor).

Through this process called doping the conductivity of semiconductors can be tuned to high

precision. And that is what makes them special. Insulators for example can be doped as well, but their band gap is too large to make them good conductors.

Now, remember that we said how electrons in the valence band of a semiconductor can jump into the conduction band by absorbing photons of high energy? In this process, energy contained in photons is transferred to energy contained in electrons, i.e. charged particles. That is getting close to something we could call a solar cell. But what happens if we connect an electric machine to a single semiconductor material that is illuminated? Will the machine work? No! In order to power the machine, we need the electrons and holes to flow as current through the machine under a voltage difference so that they can give off their energy. But to make the electrons and holes flow, *driving forces* needs to be present within semiconductor. Examples of such driving forces are gradients of gravitational potential that can act on the mass of electrons, chemical potential gradients that can act on the quantity of electrons and electrical potential gradients that can act on the charge of electrons. Since the mass of electrons is small, the gravitational force is neglible. We will also assume our device to be in thermal equilibrium. The only relevant driving forces will be gradients in chemical potential (think of concentration gradients) and gradients in electrical potential. But none of the semiconductors we have discussed so far, the intrinsic and the doped ones, have such net driving forces within them upon being illuminated. The only thing that will happen is that the excited electrons fall back into holes in the valence band after a bit (possibly by emitting a photons). That is, the electrons and holes *recombine*. Luckily, as we will now see, by combining doped semiconductors we can create a device in which driving forces will be present upon illumination that can send the excited electrons and their holes in different directions.

A.4.4 *p-n* Junctions

Suppose we join an n-type and p-type semiconductor together. The interface between the two semiconductors in such a device is referred to as a *p-n junction.* The electron rich side will be called the n-region while the side that is rich of holes will be called the p-region. Overall both the n-region and the p-region are - at least initially - electrically neutral. Now a few things will happen:

1. First of all, because there is a chemical potential gradient (think: concentration gradient) of electrons accross the p-n junction, electrons will diffuse from the n-region into the p-region. In a similar fashion, holes will diffuse from the p-region into the n-region. We see currents of majority charge carriers arising within the device. We will refer to these currents as *diffusion current*s. Once the majority charge carriers have crossed the junction, they suddenly become minority charge carriers and quickly recombine. After recombining, the charge carriers have become immobile - only thermal generation can make them mobile again but we assume the thermal generation rate to be small. As a result of the diffusion process and the recombination process, a region which has been depleted of mobile charge carriers is created around the p-n junction. This region is referred to as the *depletion region.*

2. The p-region that was electrically neutral initially has acquired a negative charge near the p-n junction by gaining electrons in the diffusion process. Similarly, the n-region has acquired a positive charge near the p-n junction by losing electrons in the diffusion process. As a result, there will be an electric field present in the depletion region. The electric field creates a potential energy barrier for the diffusing majority charge carriers. In other words, it will be more difficult for the majority charge carriers to diffuse across the depletion region.

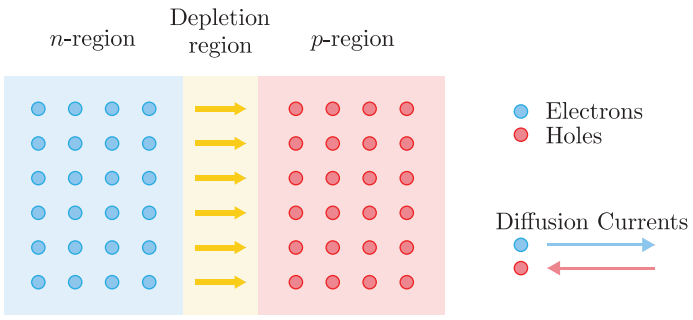

Figure A.21: Diffusion and subsequent recombination of charge carriers across the junction gives rise to a depletion region across which there is an electric field present.

3. The electrical potential barrier increases until an equilibrium situation is reached in which there are no diffusion currents. But what kind of equilibrium are we talking about? Since an electric field - that is, an electrical potential gradient - is present within the device, it is not in electrical equilibrium. Similarly because there is a concentration gradient - that is, a chemical potential gradient - present across the depletion region we can not say that the device is in chemical equilibrium either. The thing is, the electrical and chemical potentials are coupled (because in the end they both depend on the number of electrons or holes present) into a single potential called *electrochemical potential* and our device is in electrochemical equilibrium. Furthermore, as it turns out, this electrochemical potential coincides with the Fermi level that we are already familiar with.[29] Hence our device being in electrochemical equilibrium means that the Fermi level is constant throughout the device.

The equilibrium situation is sketched in Figure A.22 in an energy band diagram. Since charge carrier concentrations outside of the depletion region remain unaffected, the Fermi levels in the n-region and the p-region should be at the same relative levels as before (see Figure A.20). But then, in order for the Fermi level to be constant throughout the device, we need the energy levels in the n-region and the p-region to shift relative to one another. This is illustrated by the brown curves: the bottom one represents the highest energy level in the valence band while the top curve represents the lowest energy level in the conduction band. The difference in energy levels in the different regions is due to the electric field present within the device. Indeed, electrons in the n-region need additional energy before they can climb the potential barrier induced by the electric field and enter the p-region (think of an electron on the brown curve that needs to be rolled uphill). In a similar fashion holes in the p-region need additional energy before they can climb into the n-region.

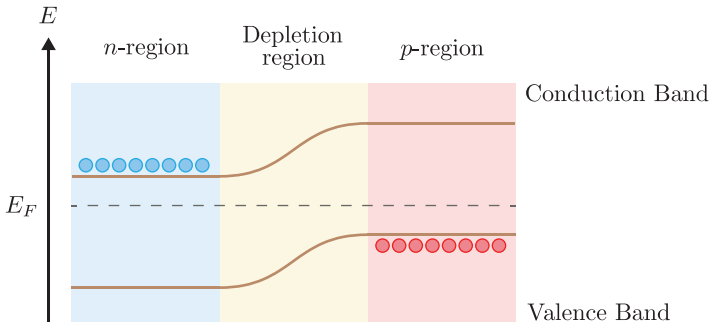

Figure A.22: Energy band diagram for a p-n junction in electrochemical equilibrium.

Even though there are electrical potential gradients (electric field) and chemical potential gradients

[29] See Chapter 3.4 of *Physics of Solar Cells* by Würfel [58] for more details.

(n-region is electron rich, p-region is hole rich) present within the material, there is no net driving force for the electrons and holes. The Fermi level is constant throughout the device. Connecting the two ends of the device with some conducting wire will *not* lead to a current flow of either electrons or holes.

Now we are finally going to have a look at what happens when we illumate our device.

A.4.5 Illuminated Semiconductors

In an n-type semiconductor the electrons are the majority charge carriers while the holes are the minority charge carriers. Now suppose that sunlight is incident on such an n-type semiconductor. The photons with an energy higher than the band gap will be able to excite electrons from the valence band into the conduction band, leaving behind holes in the valence band. We refer to this process as *photo-generation* of electron-hole pairs. Note that both the amount of majority and minority charge carriers increase under photo-generation. And in an *absolute* sense they do so by exactly the same amounts. But *relatively* speaking, the concentration of minority charge carriers will increase much more than the concentration of majority charge carriers. In fact, the concentration of minority charge carriers can increase by several orders of magnitude while the concentration of the majority charge carriers is hardly influenced. A similar thing happens for p-type semiconductors in which the holes are the majority charge carriers and the electrons are the minority charge carriers. Generally speaking, when sunlight is incident on a doped semiconductor, the concentration of minority charge carriers increases with several orders of magnitude while the concentration of majority charge carriers is hardly influenced. It is as if the sunlight is doping the material with minority charge carriers in each region.[30]

Previously we used a single Fermi level for both electrons and holes. When our device is illuminated this is no longer possible. Why not? Let us focus on the n-region first. As we just said, the concentration of electrons in the conduction band is hardly influenced so the Fermi level should be at (more or less) the same level as before. On the other hand, the concentration of holes in the valence band is increased by several orders of magnitude. That suggests the Fermi level should be shifted downwards. But we cannot shift it down and keep it at the same level at the same time! We solve this issue by introducing separate Fermi levels for the electrons and the holes.[31] Denote them by E_{F_e} and E_{F_h} respectively. With this notation, E_{F_e} should stay the same while E_{F_h} decreases upon illuminating the n-region. In a completely similar fashion it follows that E_{F_h} stays the same while E_{F_e} increases upon illuminating the p-region. We could say that there have always been separate Fermi levels for the electrons and holes respectively, but that in our discussion up to now, these two Fermi levels always overlapped.

We look at the energy band diagram for illuminated pn-junctions now.

A.4.6 Short-Circuit Current

First we assume that the ends of the device are connected by highly conductive wire. That is, we are looking at our device under *short circuit conditions*. Then, *if* there are driving forces present, current will be free to flow within the device.

[30] It should be noted here that the photo-generation of electron-hole pairs tends to disturb the equilibrium that was previously present within the device. We also know that electron-hole pairs can recombine. The recombination of photo-generated electron holes pairs can undo the effects of the photo-generated electron-hole pairs and bring back the device to its old equilibrium. If the recombination rate would be higher than the photo-generation rate, we would hardly notice any disturbance from equilibrium at all. However, as it turns out, the recombination process cannot keep up with the photo-generation process so they system will be pushed out of its old equibilibrium. See Chapter 3.5 and Chapter 3.6 of *Physics of Solar Cells* by Würfel [58] for more details.

[31] More precisely, we should introduce *quasi-Fermi levels* because the Fermi level is only well-defined in a situation of thermal equilibrium.

Now we illuminate the device. The discussion in the previous subsection tells us that E_{F_e} in the p-region increases while E_{F_e} in the n-region stays the same. Similarly, E_{F_h} increases in the n-region while E_{F_h} stays the same in the p-region. The energy band diagram must then look as in Figure A.23. In particular, we see that there is an electrochemical potential gradient present within the device. This will cause electrons to flow from (in this case) the right to the left while holes flow in the other direction. Assuming the device keeps on being illuminated, a steady current - the so called *short-circuit current* - will develop.

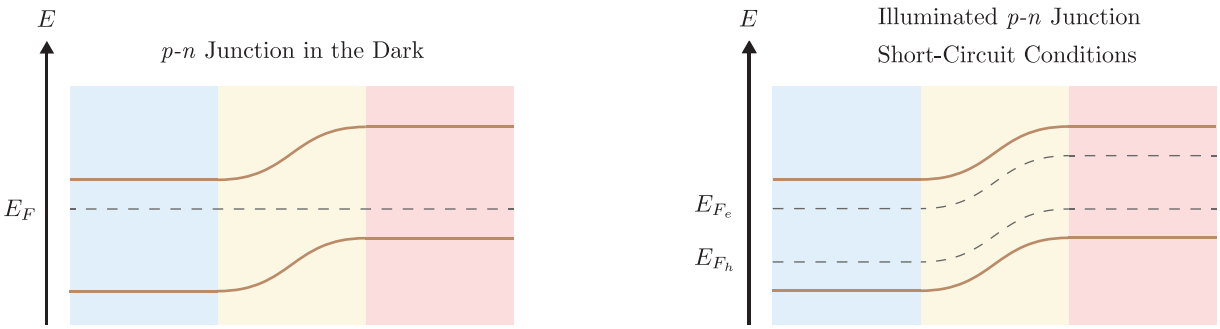

Figure A.23:
(Left) Energy band diagram for a p-n junction in the dark.
(Right) Energy band diagram for an illuminated p-n junction under open-circuit conditions.

While working with Fermi levels allows for clean and short arguments, it may be helpful to get a more physical picture of what is going on as well. In this respect, it should be noted that as the device is illuminated, the chemical potential gradients across the depletion region decrease. Because the chemical potential gradient previously balanced the electrical potential gradient across the depletion region, there will be a net driving force across the depletion region. Now imagine a photon being absorbed by an electron in the depletion region (or in the p-region but close enough to the depletion region that it can diffuse into the depletion region before recombining). The net driving force will sweep the electron across the depletion region into the n-region. Using the conducting wire, the electron can easily flow into the p-region. In the p-region the electron will quickly recombine with a hole and the process can repeat, resulting in an electron current. In a similar way a current of holes can arise in the opposite direction. Together these two currents make up the short-circuit current.

A.4.7 Open-Circuit Voltage

Next, we assume that the device is not connected to any electrical circuit. That is, we are looking at our device under *open circuit conditions*. Except for some transient behaviour no current can flow through the device in this setting.

As the sunlight starts illuminating the device, the discussion in Subsection A.4.5 again tells us that E_{F_e} in the p-region increases relative to E_{F_e} in the n-region and E_{F_c} in the n-region increases relative to E_{F_c} in the p-region. Initially, the energy band diagram will be the same as the one in Figure A.23 for short circuit conditions. But, because after the initial transient state no current can flow, the Fermi levels E_{F_e} and E_{F_c} must be constant throughout the device. It follows that the energy band diagram must look like the one in Figure A.24. In particular we see that the electrical potential barrier across the depletion region has decreased (the height-difference of the brown lines across the depletion region have decreased). The amount by which the electrical potential barrier has decreased compared with the non-illuminated equilibrium is referred to as the *open-circuit voltage*. It is the maximum voltage under which our device can operate.

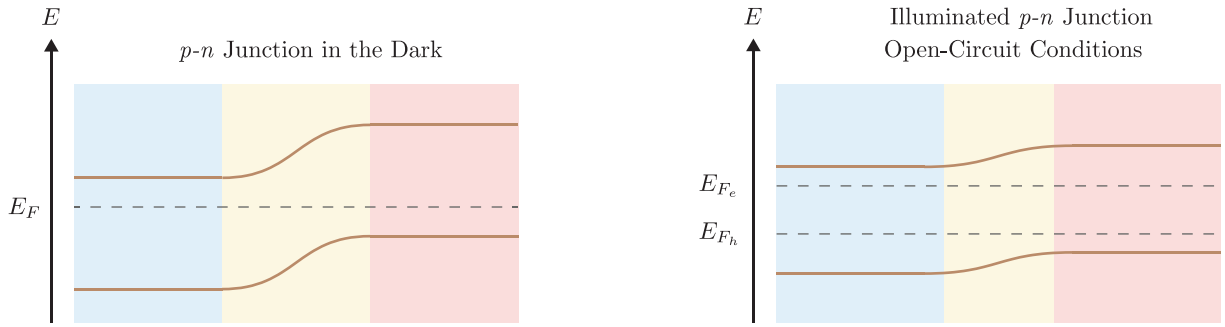

Figure A.24:
(Left) Energy band diagram for a p-n junction in the dark.
(Right) Energy band diagram for an illuminated p-n junction under open-circuit conditions.

A more physical explanation of what is going on is the following. As the the device is illuminated, the concentration gradients across the depletion region decrease. There will be a net driving force (the electrical potential gradient is now larger than the opposing chemical potential gradient) and electrons will flow from the p-region into the n-region. This time, because the ends of the device are not connected, the electrons cannot flow into the p-region to recombine and they accumulate in the n-region. For similar reasons holes accumulate in the p-region. The resulting built up of charge in each region reduces the electrical potential difference across the depletion region until a new electrochemical equilibrium is reached.

A.4.8 Solar Cells

Short-circuiting the device corresponds to connecting a load with zero resistance to the device. On the other hand, the open-circuit conditions correspond to connecting a load with infinite resistance to the device. Whenever a 'real' load having some finite resistance, for example a lightbulb, is connected to the circuit, there will be a trade-off between the open circuit voltage and the short circuit current. According to Ohm's law, the power delivered to the load is given by the product of the current and the voltage ($P = IV$). We see that we have created a solar cell: a device that can convert sunlight into electrical energy! Solar cells based on p-n junctions are the most simple types of solar cells. But most other types are based on the same underlying principles that have been discussed in this chapter.

An important parameter in optimizing the power output of solar cells is the band gap. A lower band gap means that more photons can excite electrons into the conduction band and hence a larger current can be generated. But, because electrons that have been excited into the conduction band will quickly fall down to the bottom of the conduction (by giving off energy as heat to the device for example), a lower band gap will result in a lower voltage output. There is a trade-off here but in most cases a higher band gap is beneficial because of high resistance losses otherwise. High efficiency solar cells have structures with multiple band gaps within a single cell so that a large spectrum of photons can be utilized at their full potential.

A.4.9 Overview

We do a quick recap of what has been dicsussed in this chapter. We started by trying to describe light and discovered that it has a wave-particle like duality. Most importantly for the discussion in the rest of this chapter was the idea that light consists of photons. Then we started looking at atomic models. We saw that atoms themselves are built out of a positively charged nucleus and negatively charged electrons. The electrons are only allowed to have discrete energy levels within an atom. We needed to dive into the realm of quantum physics to explain this bevahiour.

The discrete energy levels give rise to energy bands when bringing atoms together to form larger materials. The energy gap between the highest filled energy band (valence band) and the next higher energy band (conduction band) is called the band gap. Materials with a relatively small band gap are called semiconductors. Their conductivity can be tuned by doping them. Furthermore, electrons can jump from the conduction band into valence band by absorbing photons with energies higher than the band gap. In doing so, they leave holes behind that can be considered as positively charged particles. A single semiconductor will not function as a solar cell upon being illumated though. The electrons and holes will just recombine because there are no internal driving forces that separate them. However, by joining together two oppositely doped semiconductors a device is created that can deliver a current and a voltage upon being illuminated. Of course we have ignored a lot of details that are important to producing real solar cells, however, the general working principles of solar cells are now covered.

References

[1] H. W. Alt and S. Luckhaus. Quasilinear elliptic-parabolic differential equations. *Math. Z*, 183:311–341, 1983.

[2] S. A. Arrhenius. Über die Dissociationswärme und den Einflusß der Temperatur auf den Dissociationsgrad der Elektrolyte. *Zeitschrift für Physikalische Chemie*, 4(96-116), 1889.

[3] R. W. Balluffi, S. M. Allen, and W. C. Carter. *Kinetics of Materials*. John Wiley & Sons, New Jersey, 2005.

[4] L. Boltzmann. Zur Integration der Diffusionsgleichung mit variablem Diffusionskoeffizienten. *Wiedemanns Annalen*, 53:959–964, 1984.

[5] J. B. Brady. Reference frames and diffusion coefficients. *American Journal of Science*, 275: 954–983, 1975.

[6] J. Čermák and V. Rothová. Concentration dependence of ternary interdiffusion coefficients in Ni3Al/Ni3Al–X couples with X = Cr, Fe, Nb and Ti. *Acta Materialia*, 51(15):4411–4421, 2003.

[7] T. F. Chan and P. Mulet. On the convergence of the lagged diffusivity fixed point method in ôŔšłtotal variation image restoration. *SIAM J. Numer. Anal.*, 36(2):354–367, 1999.

[8] K. Cheng, W. Chen, D. Liu, L. Zhang, and Y. Du. Analysis of the Cermak-Rothova method for determining the concentration dependence of ternary interdiffusion coefficients with a single diffusion couple. *Scripta Materialia*, 76:5–8, 2014.

[9] J. Crank and G. S. Hartley. Some fundamental definitions and concepts in diffusion processes. *Transactions of the Faraday Society*, 45:801–818, 1949.

[10] L. S. Darken. Diffusion, mobility and their interrelation through free energy in binary metallic systems. *Transactions of the Metallurgical Society of AIME*, 175(1):184–194, 1948.

[11] K. M. Day. *Analysis of Interdiffusion and Diffusion Paths in Multicomponent Systems*. PhD thesis, Purdue University, 2007.

[12] M. A. Dayananda. Analysis of multicomponent diffusion couples for interdiffusion fluxes and interdiffusion coefficients. *Journal of Phase Equilibria and Diffusion*, 26(5):441–446, 2005.

[13] M. A. Dayananda and C. W. Kim. Zero-flux planes and flux reversals in Cu-Ni-Zn diffusion couples. *Metallurgical Transactions A*, 10A:1333–1339, 1979.

[14] M. A. Dayananda and Y. H. Sohn. Average effective interdiffusion coefficients and their applications for isothermal multicomponent diffusion couples. *Scripta Materialia*, 35(6): 683–688, 1996.

[15] M. A. Dayananda and Y. H. Sohn. A new analysis for the determination of ternary interdiffusion coefficients from a single diffusion couple. *Metallurgical and Materials Transactions A*, 30A:535–543, 1999.

[16] S. R. de Groot and P. Mazur. *Non-Equilibrium Thermodynamics*. Dover Publications Inc., New York, 2011.

[17] R. T. DeHoff and N. Kulkarni. The trouble with diffusion. *Materials Research*, 5(3):209–229, 2002.

[18] F. J. A. Den Broeder. A general simplification and improvement of the matano-boltzmann method in the determination of the interdiffusion coefficients in binary systems. *Scripta Metallurgica*, 3:321–326, 1969.

[19] L. C. Evans. *Partial Differential Equations*. American Mathematical Society, Providence, Rhode Island, 2010.

[20] D. K. Ferry. *Quantum Mechanics: An Introduction for Device Physicists and Electrical Engineers*. Institute of Physics Publishing, 2001.

[21] R. P. Feynman. *GED - The Strange Theory of Light and Matter*. Princeton University Press, Princeton, 1985.

[22] A. Fick. On liquid diffusion. *Poggendorffs Annalen*, 94:59–86, 1855.

[23] J. Greensite. Lecture notes on quantum mechanics, 2003. URL http://www.physics.sfsu.edu/~greensit/book.pdf.

[24] A. M. Gusak. *Diffusion-controlled Solid State Reactions Diffusion-controlled Solid State Reactions in Alloys, Thin Films, and Nano Systems*. Wiley-VCH, Weinheim, 2010.

[25] P. C. Hansen. A Matlab package for analysis and solution of discrete ill-posed problems. Version 4.1 for Matlab 7.3, 2008. URL http://www.imm.dtu.dk/~pcha/Regutools/RTv4manual.pdf.

[26] P. C. Hansen. *Discrete Inverse Problems: Insight and Algorithms*. SIAM, 2010.

[27] M. Hillert. *Phase Equilibria, Phase Diagrams and Phase Transformations - Their Thermodynamic Basis*. Cambridge University Press, Cambridge, 2007.

[28] K. Jäger, O. Isabella, A. H. M. Smets, R. A. C. M. M. van Swaaij, and M. Zeman. Solar energy: Fundamentals, technology, and systems. Technical report, Delft University of Technology, 2014.

[29] S. K. Kailasam, J. C. Lacombe, and M. E. Glicksman. Evaluation of the methods for calculating the concentration dependent diffusivity in binary systems. *Metallurgical and Materials Transactions A*, 30A:2605–2610, 1999.

[30] J. S. Kirkaldy and D. Young. *Diffusion in the Condensed State*. Maney Pub, 1988.

[31] E. O. Kirkendall and A. D. Smigelskas. Zinc diffusion in alpha brass. *Transactions of the Metallurgical Society of AIME*, 171:130–142, 1947.

[32] J. G. Kirkwood, R. L. Baldwin, P. J. Dunlop, L. J. Gosting, and G. Kegeles. Flow equations and frames of reference for isothermal diffusion in liquids. *The Journal of Chemical Physics*, 33(5):1505–1513, 1960.

[33] O. Krehel, A. Muntean, and P. Knabner. On modeling and simulation of flocculation in porous media. In Valochi [52], pages 1–8.

[34] D. Laughlin and K. Hono. *Physical Metallurgy - Volume I*. Elsevier, Amsterdam, 2014.

[35] A. Mannheim. Modelling the selenization process of Cu-In-Ga precursors. Technical report, TNO/Solliance, 2014.

[36] C. Matano. On the relation between the diffusion-coefficients and concentrations of solid metals (the nickel-copper system). *Japanese Journal of Physics*, 8:109–113, 1933.

[37] H. Mehrer. *Diffusion in Solids - Fundamentals, Methods, Materials, Diffusion-Controlled Processes*. Cambridge University Press, Cambridge, 2007.

[38] M. K. Ng, R. H. Chan, and W. C. Tang. A fast algorithm for deblurring models with neumann boundary conditionsa fast algorithm for deblurring models with neumann boundary conditions. *SIAM Journal on Scientific Computing*, 21(3):851–866, 1999.

[39] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, Berlin, 2000.

[40] L. Onsager. Reciprocal relations in irreversible processes I. *Physical Review*, 37:405–426, 1931.

[41] L. Onsager. Reciprocal relations in irreversible processes II. *Physical Review*, 38:2265–2279, 1931.

[42] B. G. Park, S. W. Hwang, and Y. J. Park. *Nanoelectric Devices*. Pan Stanford Publishing, Singapore, 2012.

[43] A. Paul, T. Laurila, V. Vuorinen, and S. V. Divinski. *Thermodynamics, Diffusion and the Kirkendall Effect*. Springer, Berlin, 2014.

[44] P. Perona and J. Malik. Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(7), 1990.

[45] I. S. Pop. Parabolic Equations. Technical report, Eindhoven University of Technology, 2012.

[46] I. S. Pop and B. Schweizer. Regularization schemes for degenerate Richards equations and outflow conditionsichards equations and outflow conditions. *Mathematical Models & Methods in Applied Sciences*, 21:1685–1712, 2011.

[47] I. S. Pop, M. Sepúlveda, F. A. Radu, and O. P. V. Villagrán. Error estimates for the finite volume discretization for the porous medium equation. *Journal of Computational and Applied Mathematics*, 234:2135–2142, 2010.

[48] E. Rothe. Zweidimensionale parabolische Randwertaufgaben als Grenzfall eindimensionaler Randwertaufgaben. *Mathematische Annalen*, 102:650–670, 1930.

[49] T. Roubíček. *Nonlinear Partial Differential Equations with Applications*, volume 153 of *International Series of Numerical Mathematics*. Birkhäuser Verlag, 2000.

[50] L. I. Rudin, S. Osher, and E. Fatemi. Non-linear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60:259–268, 1992.

[51] L. E. Trimble, D. Finn, and A. Cosgarea. A mathematical analysis of diffusion in binary systems. *Acta Metallurgica*, 13:501–507, 1965.

[52] A. J. Valochi, editor. *Proceedings of XIX International Conference on Water Resources*, 2012. CMWR 2012, Urbana-Champaign IL, USA, June, 17-22, 2012, Urbana-Champaign IL: University of Illinois at Urbana-Champaign.

[53] O. van der Heide. Spectral theorem for unbounded self-adjoint operators on a Hilbert space. Bachelor's thesis, University of Utrecht, 2013.

[54] F. J. J. van Loo. Multiphase diffusion in binary and ternary solid-state systems. *Progress in Solid State Chemistry*, 20:47–99, 1990.

[55] C. R. Vogel. *Computational Methods for Inverse Problems*. SIAM, 2002.

[56] C. R. Vogel and M. E. Oman. Iterative methods for total variation denoising. *SIAM Journal on Scientific Computing*, 171(1):227–238, 1996.

[57] C. Wagner. The evaluation of data obtained with diffusion couples of binary single-phase and multiphase systems. *Acta Metallurgica*, 17(2):99–107, 1969.

[58] P. Würfel. *Physics of Solar Cells - From Principles to New Concepts*. Wiley-VCH, Weinheim, 2005.

[59] A. E. Yagle. Regularized matrix computations. 2005. URL http://web.eecs.umich.edu/~aey/recent/regular.pdf.

[60] H. D. Young and R. A. Freedman. *University Physics*. Addison-Wesley, Boston, 2011.

[61] E. Zeidler. *Applied Functional Analysis - Applications to Mathematical Physics*. Springer-Verlag, Berlin, 1991.