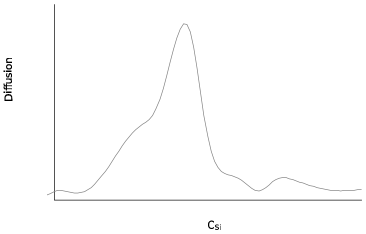
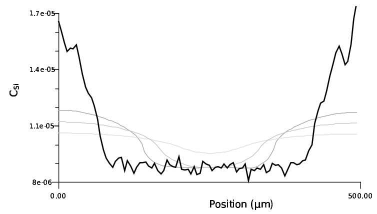
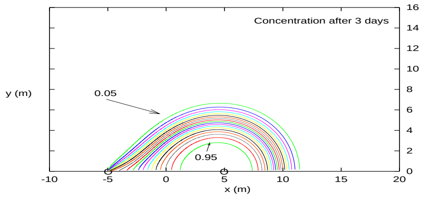
On numerical methods for direct and inverse convection-diffusion problems

Benny Malengier







Promoter: Roger Van Keer Date: 31.05.2006

Proefschrift ingediend tot het behalen van de graad van Doctor in de Wetenschappen: Wiskunde

Ghent University
Faculty of Engineering
Department of Mathematical Analysis
Research group for numerical functional analysis
and mathematical modelling – *NfaM*^{2}

UNIVERSITEIT GENT

Acknowledgements

People want to achieve many things during their lifetime. For me, obtaining my PhD was one of them. It was a long and hard work, that now comes to an end. It was a work I enjoyed very much. Having worked first in a private software company, one appreciates the possibilities offered at the university to constantly improve oneself, to learn new things and to be creative.

A lot of my gratitude therefore goes out to my promoter, Prof. Roger Van Keer, for accepting me into his research group $NfaM^2$, for bringing me into contact with interesting and actual topics of applied mathematics (mainly with an ecological motivation) and numerical analysis, for the fruitful discussions and for his guidance and stimulations. Also many thanks to Dr. Denis Constales, Prof. Marián Slodička and Prof. Jozef Kačur, a regular visitor of $NfaM^2$, for giving me a helpfull hand when necessary.

And of course, many thanks to my lovely wife, and fellow researcher, Dr. M.C. Ciocci. Who knows what I would be doing by now without her. Thank you, Cristina, for this wonder I still can't fully grasp.

From this thesis on, I will walk along new paths, explore different scientific wonders. But the knowledge and experience which I gained in the course of my PhD-studies will always be part of me. I hope some of it can be passed to you, the reader.

Benny Malengier.

Table of Contents

Acknowledgements **i**

1 Introduction **1**

1.1 Overview and main results 2

1.2 Nederlandse samenvatting 7

I Direct Problems **15**

2 Advection dominated diffusion problem **17**

2.1 Mathematical tools 19

2.1.1 Operator splitting 19

2.1.2 The Riemann problem 23

2.1.3 Finite volume methods 30

2.2 The dual-well as a practical example 33

2.2.1 Physical background 33

2.2.2 Flow field of the dual-well 35

2.2.3 Mathematical model for contaminant transport 41

2.3 Solution of advection dominated diffusion in a rectangle 45

2.3.1 Numerical approximation of (2.44) 45

2.3.2 Solution of the general transport problem 47

2.3.3 Solution of the diffusion part 49

2.3.4 Benchmark solution: the case $\alpha_T = 0 = D_0$ 50

2.3.5 General remark on operator splitting 52
2.4 Advection dominated diffusion and equilibrium adsorption 52
2.4.1 Numerical approximation of (2.61) 53
2.4.2 Solution of the nonlinear transport problem 55
2.4.3 Solution of the projection problem 57
2.4.4 Solution of the nonlinear diffusion problem 57
2.5 Convergence of the numerical method in 2D 60
2.5.1 Semi-discrete method 62
2.5.2 Aim 63
2.5.3 Fully discrete method 64
2.5.4 The hyperbolic step and the projection 67
2.5.5 The parabolic step 74
2.5.6 The three steps combined 82
2.5.7 Existence 86
2.5.8 Boundary conditions and operator splitting 93
2.6 Numerical experiments 96
2.6.1 Without adsorption 96
2.6.2 Mass balance and advective inflow flux 110
2.6.3 Adsorption 114

3 A degenerate diffusion problem in 1D 119

3.1 Annealing diffusion as a practical example 120
3.1.1 Mathematical model of diffusion annealing 120
3.1.2 Moving interface 122
3.1.3 Analytical solution 124
3.1.4 Numerical approximation 125
3.2 Physical experiments 127
3.2.1 Experimental procedure 128
3.2.2 Coating composition and formation 128
3.2.3 Diffusion annealing 128
3.3 Numerical experiments 132
3.3.1 Power type diffusion 132

3.3.2 B-spline diffusion 137

II Inverse Problems 141

4 Inverse problems in annealing diffusion 145

4.1 The adjoint problem 147

4.1.1 Deduction 147

4.1.2 Numerical approximation 150

4.1.3 Computation of $\nabla_p \mathcal{F}$ 150

4.2 Convergence for $\partial_\tau \psi - a(x,\tau)\partial_x^2 \psi = f(x,\tau)$ 151

4.2.1 Discretization 153

4.2.2 Weak solution 157

4.2.3 A priori estimates 158

4.2.4 Convergence 168

4.2.5 Convergence of the ϵ-regularization 175

4.3 Discrete time measurements 178

4.4 Numerical experiments 179

4.4.1 Experiment 1: Adjoint method 179

4.4.2 Experiment 2: Levenberg-Marquardt 181

5 Inverse problems for the dual-well experiment 183

5.1 Subsurface parameter identification 183

5.2 Use of the Levenberg-Marquardt method 185

5.3 Use of the adjoint equation method 186

5.3.1 Deduction of the adjoint system 187

5.3.2 Numerical approximation 191

5.4 Numerical experiments 192

5.4.1 Without adsorption 192

5.4.2 With adsorption 197

III On a practical groundwater flow problem 199

6 A practical groundwater flow problem 201
6.1 Physical background: Tòth's regional flow problem 201
6.2 Tòth's regional flow model 204
6.3 General mathematical model 206
6.4 Analytical solution 208
6.5 Approximation of the formal solution and numerical experiments 214
6.6 Numerical approximation methods 218
6.6.1 Finite element algorithm 218
6.6.2 Infinite element algorithm 220
6.7 Conclusion 224

A Notations, concepts and auxiliary results 227
A.1 Basic definitions, identities and inequalities 227
A.2 Theorems 229

B Basic concepts in groundwater flow modeling 231
B.1 The subsurface 231
B.2 Nomenclatura 232
B.3 Darcy's law and hydraulic conductivity 233
B.4 Mass conservation equation 235
B.5 Dupuit-Forchheimer flow 236
B.6 Contaminant transport 237
B.7 Adsorption 239
B.8 Radioactive decay 241

C Basic facts on numerical methods for inverse problems 243
C.1 Introduction 243
C.2 The penalty function 244
C.3 Abstract framework 245
C.4 Optimization: Gradient based methods 246
C.4.1 Steepest descent method 246

C.4.2 Conjugate gradient method 247
C.4.3 Newton methods and the practical Gauss-Newton method 248
C.4.4 Trust region globalization and the practical Levenberg-Marquardt method 250
C.5 Gradient determination methods 251
C.5.1 Finite differences 252
C.5.2 Adjoint or costate methods 252

List of Symbols 255

List of Figures 259

List of Tables 265

Chapter 1

Introduction

This thesis consists of 3 parts, spanning 5 chapters, and 3 appendixes. Elliptic, parobolic, and hyperbolic partial differential equations are considered, the emphasis being on numerical modeling. Our aim was to consider real world problems, determine the techniques needed to solve them and tackle open mathematical problems related to them. This resulted in the treatment of various subjects ranging from flow problems to adsorption experiments.

When considering real world problems, the emphasis is usually not on the study of the solution itself, but on the use of the solution to reach a better understanding of the world around us, or to build/construct better tools. This directly leads to the field of inverse problems. It should therefore be no surprise that we focussed on inverse problems. Two physical situations were considered: determining properties of the subsurface in relation to contaminants in the groundwater flow, and determining the effective diffusion during the construction of steel alloys.

These problems are both of diffusion type, although of very different nature: one being convection-diffusion, with dominant convection and combined with adsorption, the other being pure diffusion.

Diffusion, coupled with flow or other processes, is a long standing subject within classical physics and yet a thoroughly modern one. Looking at the current science and technology scene, one cannot help but be impressed that applications as diverse as biodiffusion across cell membranes and dopant diffusion in semiconductors can be understood from a few similar basic rules. However, the time and length scales involved can differ enormously, and extra processes like adsorption, reaction, etc. can complicate the matter further. Many approximation methods have been developed to study diffusion, each with its strong and weak points. Many mathematical problems related to their development are still open and many questions unanswered.

Our first task was the construction of the correct physical models for the problems at hand. Secondly, we determined suitable approximation methods to obtain solutions of the models. At last, but not least, we concentrated on two open mathematical problems: convergence of the operator splitting method for a nonlinear convection-diffusion equation, and existence of a solution for a degenerate, variable coefficient, convection-diffusion-reaction problem, that arises in the inverse problem of a degenerate nonlinear diffusion equation.

In addition, we solved the Tòth's regional flow problem without the simplifications and approximations considered by other authors. This is a groundwater flow problem, of which the solution is needed to deduce contaminant flow. It can be reduced to a steady state diffusion type problem. We could construct a semi-analytical solution and evaluate it numerically.

1.1 Overview and main results

Part I, Direct problems Two model problems are described and appropriate approximation methods are constructed.

The first problem is related to groundwater quality. Groundwater pollution is becoming an ecology endangering problem caused by economic activities and human lifestyle, not only in the industrial countries, but also in the developing countries. It can lead to environmental disasters and water shortages: an explosive combination due to the ever increasing world population. In the near future, many contaminated sites will have to be cleaned, and unspilt aquifers will need protection. To help this fight, mathematical modeling is of fundamental importance. Models can forecast future contamination situations, calculate the impact of certain remedies and provide a better understanding of the processes involved.

A major difficulty with the subsurface is that overall observation is not possible. The entire structure must be extrapolated from a limited number of discrete observations. A field worker can only chose some specific points at which to study the underground, mostly by taking ground samples, or by placing a groundwater well. This often limits the possibilities of a mathematical model, as insufficient data for the calibration of complicated regional groundwater pollution models can undermine the validity of the decisions based on the model. The determination of model characteristics is therefore very important.

At the same time, the consequences of many processes, such as adsorption and infiltration of NAPLs (nonaqueous phase liquids, like benzene) are still unclear. Small scale tests and models need to be implemented to better understand these processes before they can be incorporated in regional models. Convergence in ideal settings has to be proved to give the methods credibility.

In Chapter 2 the dual-well problem is solved. The dual-well is a subsurface measurent technique, in which two wells are drilled, one pumping water into the subsurface, and one pumping out the same amount of water. A tracer is added to the water, and its concentration is measured at the outflow. The dual well is in essence a convection-diffusion problem with dominant convection. In Section 2.1 the mathematical tools are presented: operator splitting, the Riemann problem and finite volume methods. In Section 2.2 the dual-well model is deduced. The flow field is calculated, and the model for contaminant transport is deduced. Then, the approximation method, without and with adsorption is given. In Section 2.5 the approximate solution is proved to converge to a very weak solution. Finally, Section 2.6 provides numerical experiments validating the developed techniques.

The constructed mathematical model is novel in that it reduces the dual-well problem in $\mathbb{R}^2$ to a rectangular domain by a conformal mapping where the flow takes place along the verticals. The contaminant flow is governed by a convection-diffusion equation with dominant diffusion

$$\partial_t F(u) - v(y)\partial_y u - g(x,y)\nabla \cdot D(x,y)\nabla u = 0.$$

We apply operator splitting as an approximation method. Here, the convection is splitted from the diffusion. A Rieman solver is constructed for the transport part, and a finite volume method for the diffusion part. We show that the approximation is very effective and fast. Moreover, in the nonlinear setting when also adsorption is considered, a hybrid implementation of the relaxation method and Newton's method combined with the finite volume method, allows for taking large time steps and provides convergence in few iteration steps.

We define a very weak solution to the problem, see Definition 2.5.2. The novelty is that boundary conditions are taken into account in the definition itself. We prove convergence of the numerical method to this very weak solution in Theorem 2.5.1. The proof is based on Riesz-Frechet-Kolmogorov compactness arguments and techniques from Crandall and Majda [16]. One of the key points is the proof of boundedness of the total variation, which is achieved in 2D by combining the contributions of the different steps.

Finally, the performed experiments not only illustrate the validity of the approach, but also show how the different subsurface parameters influence the output of the dual-well experiment.

The second problem treated in Part I relates to power losses in electrical motors and transformators. These consist of iron cores which are made from steel alloys. For economical and ecological reasons one wants to produce and use alloys that optimize the inner workings of the motor or transformator. The minimization of power losses can be achieved with steels that have a higher Si (silicon) content than the nowadays produced ones. Industrial techniques are being developed towards this purpose, but much work remains in optimizing them and in understanding the physical processes involved.

One such physical process is the diffusion of Si into the steel matrix by a technique called annealing diffusion. The interaction with Al (aluminium) complicates the behaviour. Physical parameters for a ternary Si-Al-Fe combination are mainly unknown, but are needed in order to obtain solutions of a model for the time evolution of the alloy. A first step to obtain these parameters, avoiding large research investments, is the determination of the apparent diffusion in a simplified setting.

In Chapter 3 annealing diffusion is presented. In Section 3.1 the process is explained, and a mathematical model is given. This model is reduced to one with an apparent diffusion. Furthermore, a numerical approximation based on the method of lines is developed, both with and without the presence of a moving interface. In Section 3.2 physical experiments performed at LabMet of Ghent University are presented, and in Section 3.3 numerical experiments are given.

Our main, novel contribution is the set-up of the reduced model, and its approximation based on the method of lines. Furhermore, in the case of a moving interface, we show how the application of Landau's transformation allows for a straightforward determination of the time varying, a priori unknown position of the inferface.

Part II, Inverse problems The models developed in Part I depend on several parameters, which have to be known exactly to enable the use of the models on real-world examples. These parameters are diffusion coefficients, reaction rates, etc. In this thesis, the set-ups have been specifically constructed to allow the correct determination of some of the parameters: the purpose of the dual-well is the retrieval of subsurface parameters, and during the annealing diffusion extra measurements are performed to allow the determination of the diffusion coefficient.

We start in Chapter 4 with the treatment of diffusion annealing. First the cost functional, measuring the deviation of the measured values from the values obtained numerically for a choosen parameter set, is given. In Section 4.1 the adjoint or costate problem is developed. This is an auxilary PDE that allows to construct the gradient of the cost functional, which is needed to obtain the optimal parameter values. The adjoint equation is of degenerate convection-diffusion-reaction type. In Section 4.2, existence of solution of the adjoint equation is proved. At the end of the chapter, we provide an extension of the problem to discrete time measurements and provide numerical experiments illustrating the validity of the developed approximations.

Our main contribution is the construction of the costate problem for the specific setting of diffusion annealing. We obtain an adjoint equation of the form,

$$\partial_t u - a(x,t)\partial_x^2 u = f(x,t),$$

where there is a degenerate, time varying point $x = s(t)$, with $a(x,t) = 0$ for $x \geq s(t)$ and $\partial_x a(s(t),t) = c(t)$ with $-\infty \leq c(t) < 0$.

The technique to arrive at the adjoint equation is well established. However, the resulting adjoint equation seems not to have been considered so far. The difficulty concerns the fact that the diffusion can be strongly degenerate, i.e.

$$\partial_x a(s(t),t) = -\infty,$$

resulting in a pure reaction problem in a part of the domain. We are able to prove existence of a solution of the weak formulation, in two settings. First, when only a regularization of $a(x,t)$ around $x = s(t)$ is considered, Theorem 4.2.1. Secondly, when a vanishing viscosity solution is considered, Theorem 4.2.2. A numerical experiment on a model problem shows that the adjoint equation provides good results for the recovery of the parameters.

In Chapter 5 we consider the dual-well experiment. We give an overview of parameter identification techniques for the subsurface in Section 5.1. In Section 5.2 the Levenberg-Marquardt method is presented and in Section 5.3 the adjoint problem is developed, as well as a suitable numerical approximation. Finally, in Section 5.4, numerical experiments are given.

The interesting feature of the chapter is that we develop the adjoint problem for a convection-diffusion problem, where measurements are averaged values over a part of the boundary, the so called break through curve. This could seem to be a poor input for the adjoint equation. Indeed, the only way the experimentally measured concentration $u_B(t)$ of the tracer in the outflow well (as solution of the direct problem in terms of $u(x, y, t)$), appears in the adjoint equation (given in terms of $v(x, y, t)$), is through a boundary condition of the adjoint problem,

$$a\partial_y(g(x,y)v(x,y,t)) - bg(x,y)v(x,y,t) = c\int_{\Gamma_1} (u(\tilde{x}(s),\tilde{y}(s),t) - u_B(t)\,\mathrm{d}s, \text{ op } \Gamma_1.$$

One would expect this to lead to an ill-posed inverse problem, unsuitable for the determination of parameters. The contrary is true: numerical experiments give very good results. The concept of the dual-well turns out to be a powerfull methodology to determine subsurface parameters.

Part III, On a practical groundwater flow problem In solving the dual-well groundwater flow problem, we use the so-called Dupuit-Forchheimer approximation, which neglects vertical groundwater flow. While validating this approach, we considered Tòth's regional flow problem, a well known example in which vertical groundwater flow is very prominent. We discovered that although the model is wel described in the literature, the approximations used are based on strong simplifications. In the original paper [73], Tóth projects the intricate domain of the problem on a rectangle, whereas a more recent contribution [68] uses an infinitely deep basin. In Chapter 6 we reconsider the problem giving not only a semi-analytical solution, but also illustrating the use of an "infinite" element method as a possible approximation technique.

The outline of Chapter 6 is as follows. In Section 6.2 the original analytical solution of Tóth is presented; in Section 6.3, we state the improved mathematical model of our choice. An analytical solution involving infinite series is derived in Section 6.4. The semi-analytical approach is discussed in Section 6.5, and numerical results are compared with results from the literature. In Section 6.6, we briefly sketch a finite element approach and an infinite element approach. Furthermore, we demonstrate how the last method can be used for the case of a very deep basin.

The main merit of this chapter is that it gives different types of solution to an important problem, varying from (semi)-analytical ones to numerical solutions by finite and infinite element methods.

Appendices To make this thesis (partially) self-contained, we added 3 appendices. Appendix A reviews some standard mathematical tools from functional analysis. Appendix B deals with basic concepts in groundwater modeling, encountered in Chapters 2 and 6. Finally, Appendix C covers the basic aspects of numerical methods for inverse problems, as a background for Part II.

1.2 Nederlandse samenvatting: Numerieke methoden voor directe en inverse convectie-diffusie vraagstukken

Introductie Deze thesis bevat 3 delen, verspreid over 5 hoofdstukken, en 3 appendices. Elliptische, parabolische en hyperbolische partiële differentiaalvergelijkingen worden beschouwd. De nadruk ligt op numerieke modellering. Ons doel was het bestuderen van enkele practische probleemstellingen, het bepalen van de technieken nodig om deze op te lossen, en het beantwoorden van open wiskundige vragen die ermee gerelateerd zijn. Dit resulteerde in de behandeling van diverse onderwerpen, gaande van stromingsproblemen tot adsorptie-experimenten.

Bij vraagstukken met praktisch nut ligt de nadruk normaal niet op de studie van de oplossing van het vraagstuk zelf, maar op het gebruik van die oplossing om de fysische/chemische/... probleemstelling beter te begrijpen, en zo ons toe te laten betere materialen, technieken, enz. te ontwerpen. Dit leidt onmiddelijk tot inverse problemen. Het hoeft dan ook niet te verwonderen dat we focussen op twee inverse problemen: het bepalen van eigenschappen van de ondergrond in relatie tot contaminantentransport, en het bepalen van de effectieve diffusie gedurende de constructie van staallegeringen.

Deze problemen zijn beide van het diffusietype, alhoewel met erg verschillende eigenschappen: de ene is convectie-diffusie met dominante convectie en gecombineerd met adsorptie, de andere is een puur diffusieprobleem.

Diffusie, al dan niet gekoppeld met andere processen, is een goed bestudeerd thema binnen de klassieke fysica, maar terzelfdertijd een erg modern toepassingsgebied. Men kan enkel onder de indruk zijn van het feit dat toepassingen die zo divers zijn als biodiffusie tussen celmembranen en doperingsdiffusie in halfgeleiders, beschreven kunnen worden met enkele, gelijkaardige fysische basisregels. Niettemin kunnen de tijd- en lengteschalen sterk verschillen, en extra wisselwerkingen, zoals adsorptie, reactie, enz., kunnen de complexiteit sterk doen toenemen.

Veel benaderingstechnieken zijn ontworpen om diffusie te bestuderen, elk met sterke en zwakke punten, maar veel hierbij opduikende wiskundige problemen zijn nog onbeantwoord.

Onze eerste taak was de constructie van een correct fysisch model voor de vraagstukken. Daarna bepalen we steeds een geschikte benaderingstechniek om tot een oplossing te komen van het model. Uiteindelijk, concentreren we ons op twee open wiskundige vraagstukken: convergentie van de operatorsplitsingstechniek voor niet-lineaire convectie-diffusievergelijkingen, en existentie van een oplossing van een gedegenereerd convectie-diffusie-reactievraagstuk met variabele coëfficiënten, dat optreedt bij de studie van het inverse probleem van een ontaarde niet-lineaire diffusievergelijking.

Daarenboven lossen we Tòth's regionaal stromingsvraagstuk op, zonder de simplificaties en benaderingen die tot nu toe gebruikt werden in de literatuur. Dit is een grondwaterstromingsvraagstuk waarvan de oplossing kan gebruikt worden om contaminantentransport te bepalen. Het vraagstuk kan gereduceerd worden tot een stationair probleem van het diffusietype. We hebben een semi-analytische oplossing opgesteld en numeriek geëvalueerd.

Deel I, Directe vraagstrukken Twee specifieke problemen worden beschreven en gepaste benaderingstechnieken worden ontwikkeld.

Het eerste probleem is verbonden met de kwaliteit van het grondwater. Grondwatervervuiling is een groeiend probleem ten gevolge van economische activiteiten en onze levenstijl, niet enkel in de industriële landen, maar ook in de ontwikkelingslanden. Het kan leiden tot milieurampen en watertekorten: een explosieve combinatie ten gevolge van de stijgende wereldpopulatie. In de nabije toekomst zullen veel vervuilde sites opgeruimd moeten worden en zullen onbedoezelde aquifers moeten beschermd worden. Hierbij zullen mathematische modellen een fundamentele rol spelen. Modellen kunnen toekomstige situaties van vervuiling voorspellen, ze laten toe de impact van bepaalde remedies te onderzoeken, en bieden de mogelijkheid de betrokken processen beter te begrijpen.

Een belangrijke moeilijkheid betreffende de ondergrond is dat een grootschalige monitoring niet mogelijk is. De totale structuur moet geëxtrapoleerd worden vertrekkende van een beperkt aantal discrete metingen. Een veldwerker kan enkel specifieke plaatsen kiezen waar hij de ondergrond bestudeert, via een boring of door het opzetten van een boorput. Dit zorgt vaak voor beperkingen in de mogelijkheden van een wiskundig model omdat er onvoldoende data zijn voor de callibratie van complexe grondwaterpollutiemodellen. Dit ondermijnt dan beslissingen die genomen worden op basis van het model. Bijgevolg is het bepalen van karakterestieken van de ondergrond erg belangrijk.

Terzelfdertijd zijn de gevolgen van vele processen, zoals adsorptie en infiltratie van NAPLs (bv. benzeen), nog onduidelijk. Proeven op beperkte schaal en modellen moeten geïmplementeerd worden om tot een beter begrip te komen, alvorens deze processen kunnen opgenomen worden in regionale modellen. Verder moet minstens convergentie van de approximatiemethoden in ideale omstandigheden bewezen worden om de methoden credibiliteit te geven.

In *Hoofdstuk 2* wordt het doublet-probleem opgelost. De doublet is een meettechniek waarin twee bronnen geboord worden. De ene bron wordt gebruikt om water in de ondergrond te pompen (infiltratiebron), via de andere wordt dezelfde hoeveelheid water terug opgepompt (onttrekkingsbron). Een merker wordt aan het water toegevoegd, en de concentratie van de merker wordt gemeten in het opgepompte water.

De doublet leidt in essentie tot een convectie-diffusieprobleem met dominante convectie. In Sectie 2.1 worden de wiskundige methoden gepresenteerd die nodig zijn om een oplossing te bekomen: operatorsplitsing, het Riemann probleem, en eindige volumemethoden. In Sectie 2.2 wordt het doubletmodel afgeleid. Het snelheidsveld wordt berekend, en het model voor contaminantentransport wordt afgeleid. Daarna wordt de benaderingsmethode, met en zonder adsorptie, gegeven. In Section 2.5 wordt bewezen dat de benadering convergeert naar een erg zwakke oplossing. Uiteindelijk, in Sectie 2.6, valideren we de ontwikkelde technieken aan de hand van numerieke experimenten.

Het ontwikkelde wiskundige model is vernieuwend in de zin dat het de doublet reduceert van een probleem in $\mathbb{R}^2$ tot een rechthoekig domein via een conforme afbeelding, waar bovendien de stroming gebeurt langs de vertikalen. Het contaminantentransport wordt beheerst door een convectie-diffusievergelijking met dominante diffusie van de vorm

$$\partial_t F(u) - v(y)\partial_y u - g(x,y)\nabla \cdot D(x,y)\nabla u = 0.$$

We passen operatorsplitsing toe als benaderingstechniek. Hierdoor wordt de convectie gesplitst van de diffusie. Een Riemann methode wordt gebruikt voor het transportgedeelte, en een eindige volumemethode voor het diffusiegedeelte. We tonen aan dat de approximatie erg effectief en snel is. Daarenboven, in het niet-lineaire geval, wanneer ook adsorptie beschouwd wordt, implementeren we een combinatie van de relaxatiemethode en Newton's methode op basis van de eindige volumemethode. Dit laat toe grote tijdstappen te beschouwen, met convergentie over één tijdstap van de niet-lineaire approximatie na enkele iteraties.

We definiëren vervolgens een erg zwakke oplossing van het probleem, zie Definitie 2.5.2. Nieuw is dat de randcondities ingebouwd worden in deze definitie zelf. We bewijzen dan de convergentie van de numerieke methode naar deze erg zwakke oplossing, Theorema 2.5.1. Het bewijs is gebaseerd op Riesz-Frechet-Kolmogorov compactheidsargumenten en op technieken van Crandall en Majda [16]. Een van de kernpunten is het bewijs van de begrensdheid van de totale variatie, die bereikt wordt in 2D door het combineren van de bijdragen van de verschillende stappen.

De uitgevoerde experimenten illustreren niet enkel de geldigheid van de werkwijze, maar tonen ook hoe verschillede ondergrondparameters invloed hebben op de doubletmetingen.

Het tweede probleem dat behandeld wordt in Deel I, heeft verband met stroomverliezen in electrische motoren en transformatoren. Deze bestaan uit ijzerkernen gemaakt met staallegeringen. Omwille van economische en ecologische redenen beoogt men de produktie en het gebruik van legeringen die de werking van de motor of transformator optimalizeren. De minimisering van de stroomverliezen kan bereikt worden met staal dat een hogere Si (silicium) concentratie heeft dan het thans via massaproductie gemaakte staal. Met dit doel worden industriële technieken ontwikkeld, maar het optimalizeren hiervan en het begrijpen van de relevante fysische processen blijven belangrijk.

Een van deze fysische processen is de diffusie van Si in de staalmatrix via een techniek genaamd “diffusie annealing”, het opwekken van diffusie door het uitgloeien van het staal. Dit zou een standaard probleemstelling kunnen zijn, maar vanwege de interactie met Al (aluminium) treden complicaties op. Fysische parameters voor een ternair Si-Al-Fe systeem zijn nog onbekend, maar zijn toch nodig om tot oplossingen te komen in modellen die de tijdsevolutie van de legering pogen te simuleren. Een eerste stap in het bekomen van deze parameters, zonder grote onderzoeksinversteringen, is het bepalen van een schijnbare diffusie van een gereduceerd probleem.

In *Hoofdstuk 3* wordt diffusie-annealing beschreven. In Sectie 3.1 wordt het proces verklaard en een mathematisch model opgesteld. Dit model wordt dan gereduceerd tot een model met een schijnbare diffusie. Ook ontwikkelen we een numerieke benadering gebaseerd op de methode der lijnen, en dit zowel voor het geval met een bewegend interactievlak als zonder. In Sectie 3.2 worden fysische experimenten, uitgevoerd aan het laboratorium LabMet van de Universiteit Gent weergegeven. In Sectie 3.3 vermelden we enkele numerieke experimenten.

Onze voornaamste bijdrage is het construeren van een gereduceerd model en de benadering ervan via de methode der lijnen. Verder hebben we in het geval van een bewegend interactievlak aangetoond hoe het gebruik van Landau’s transformatie toelaat om de tijdsveranderlijke en a priori ongekende positie van dit vlak transparant te bepalen.

Deel II: Inverse vraagstrukken De modellen ontwikkeld in Deel I zijn afhankelijk van verschillende parameters die nauwkeurig moeten gekend zijn om toe te laten de modellen te gebruiken in realistische omgevingen. De parameters zijn bv. diffusiecoëfficienten, reactiesnelheden, enz. In deze thesis reflecteren de modellen werkwijzen die specifiek opgesteld zijn om sommige van de parameters te bepalen: het doel van de doublet is het bekomen van ondergrondparameters; gedurende de annealing-diffusie worden extra metingen uitgevoerd om de diffusiecoëfficient te bepalen.

We beginnen *Hoofdstuk 4* met de probleembeschrijving van diffusie-annealing. We stellen eerst de kostfunctionaal op. De kostfunctionaal is een maat voor de afwijking tussen experimenteel bepaalde waarden van de oplossing van het diffusieprobleem enerzijds en numeriek bekomen waarden van de oplossing corresponderend met een gekozen parameterstel anderzijds. In Sectie 4.1 wordt het duaal probleem opgesteld. Dit is een hulpvraagstuk dat toelaat om de gradient van de kostfunctionaal te berekenen. Deze gradient is nodig om de optimale parameterwaarden te vinden. De duale vergelijking is van het ontaarde convectie-diffusie-reactietype. In Sectie 4.2 bewijzen we existentie van een oplossing van de duale vergelijking. Op het einde van het hoofdstuk breiden we het probleem uit tot discrete tijdsmetingen, en beschouwen we numerieke experimenten die de bruikbaarheid van de ontwikkelde benaderingen aantonen.

De hoofdbijdrage bestaat uit het opstellen van het duaal probleem voor het diffusie-annealingvraagstuk. We verkrijgen een differentiaalvergelijking van de vorm

$$\partial_t u - a(x,t)\partial_x^2 u = f(x,t),$$

waarbij ontaarding optreedt in het tijdsveranderlijke punt $x = s(t)$, met $a(x,t) = 0$ voor $x \geq s(t)$ en met $\partial_x a(s(t),t) = c(t)$, waarbij $-\infty \leq c(t) < 0$.

De methode om een duaal probleem op te stellen is doorgaans goed gekend. In het huidig geval evenwel, is de resulterende duale vergelijking nog niet beschouwd in de literatuur. De vergelijking kan namelijk sterk ontaard zijn, $\partial_x a(s(t),t) = -\infty$, en overgaan in een puur reactieprobleem in een tijdsveranderlijk deel van het domein. We kunnen de existentie van een oplossing van de zwakke formulering van het probleem bekomen in twee gevallen. Ten eerste via een regularizatie van $a(x,t)$ rond $x = s(t)$, Theorema 4.2.1. Ten tweede wanneer een viscositeitsoplossing beschouwd wordt, Theorema 4.2.2. Een numeriek experiment van een model probleem illustreert dat de duale vergelijking goede resultaten oplevert voor de reconstructie van de parameters.

In *Hoofdstuk 5* beschouwen we het doublet-experiment. We geven een overzicht van methoden voor parameteridentificatie voor de ondergrond in Sectie 5.1. In Sectie 5.2 wordt de Levenberg-Marquardt-methode bondig in herinnering gebracht. In Sectie 5.3 wordt het duale probleem ontwikkeld, samen met een gepaste numerieke benadering. In Sectie 5.4 tenslotte, worden numerieke experimenten gegeven.

De voornaamste bijdrage van het hoofdstuk is dat we een duaal probleem uitwerken voor een convectie-diffusievraagstuk, waarbij de metingen gemiddelde waarden zijn over een deel van de rand van het domein, de zogenaamde doorsijpelingscurve. Dit lijkt beperkte informatie te zijn om een duaal probleem mee op te stellen. Inderdaad, de enige manier dat de experimenteel opgemeten concentratie $u_B(t)$ van de merker in het opgepompte water (als resultaat van het directe vraagstuk opgesteld in termen van $u(x,y,t)$), voorkomt in de duale vergelijking (opgesteld in termen van $v(x,y,t)$), is via een randconditie van het duale probleem,

$$a\partial_y(g(x,y)v(x,y,t)) - bg(x,y)v(x,y,t) = c\int_{\Gamma_1} (u(\tilde{x}(s),\tilde{y}(s),t) - u_B(t)\,\mathrm{d}s, \text{ op } \Gamma_1.$$

Men zou verwachten dat dit leidt tot een slecht gesteld invers probleem, waarmee het bijzonder moeilijk is parameters te identificeren. Het tegendeel is evenwel waar: via de numerieke experimenten worden zeer goede resultaten bekomen. Dit illustreert dat de doublet een efficiënte methode oplevert om parameters van de ondergrond mee te bepalen.

Deel III: Een practisch groundwaterstromings probleem Om het doublet-grondwaterstromingsprobleem op te lossen, gebruiken we de zogenaamde Dupuit-Forchheimer benadering, die vertikale grondwaterstroming verwaarloost. Gedurende de validatie van deze techniek, werden we geconfronteerd met Tòth's regionaal stromingsprobleem, een bekend voorbeeld waarin vertikale grondwaterstroming wél van belang is en zelfs een prominente rol speelt.

Alhoewel het model goed beschreven is in de literatuur, steunen de gebruikte oplossingsmethoden op sterke vereenvoudigingen. In het originele rapport, [73], projecteert Tóth het domein op een rechthoek, terwijl een meer recente bijdrage, [68], een oneindig diep domein gebruikt. In *Hoofdstuk 6* behandelen we het probleem en geven we niet enkel een semi-analytische oplossing, maar tonen we ook aan hoe de "oneindige" elementenmethode kan gebruikt worden als approximatiemethode.

Het hoofdstuk is als volgt ingedeeld. In Sectie 6.2 geven we de originele oplossing van Tóth weer. In Sectie 6.3 beschouwen we een verbeterd wiskundig model. Een analytische oplossing bestaande uit oneindige reeksen wordt afgeleid in Sectie 6.4. De semi-analytische benadering wordt uiteengezet in Sectie 6.5. De numerieke resultaten worden verder vergeleken met resultaten uit de literatuur.

In Sectie 6.6 wordt dan kort de toepassing van de eindige elementenmethode besproken, alsook van een gepaste oneindige elementenmethode. We tonen aan hoe die laatste kan gebruikt worden bij de modellering van diepe grondlagen.

Het voornaamste doel van dit hoofdstuk is om voor een belangrijk hydrogeologisch vraagstuk enkele uiteenlopende oplossingsmethodes voor te stellen, gaande van een (semi)-analytische methode tot numerieke methoden.

Appendices Om deze thesis (gedeeltelijk) op zichzelf staand te maken, hebben we 3 appendices toegevoegd. Appendix A geeft een overzicht van enkele standaard resultaten uit de functionaalanalyse. Appendix B beschouwt basisbegrippen in verband met grondwatermodellering, nuttig voor Hoofdstukken 2 en 6. In Appendix C worden basisaspecten van numerieke methoden voor inverse vraagstukken beschreven. Dit als achtergrond bij Deel III.

Part I

Direct Problems

Chapter 2

A nonlinear advection dominated diffusion problem in 2D

We consider advection-diffusion problems in a domain $\Omega \subset \mathbb{R}^d$ modeled by the variable coefficient equation

$$\partial_t \phi(u) + \nabla \cdot (\boldsymbol{v} u - \boldsymbol{D} \nabla u) = 0. \tag{2.1}$$

Here $\phi(u)$ is the retardation, $\boldsymbol{v}$ is the velocity field depending on the position, and $\boldsymbol{D}$ is the diffusion tensor, also depending on the position. We look for a solution $u(x,t)$, $x \in \Omega$ and $t \in (0,T) := I_t$ satisfying (2.1), along with an initial condition

$$u(x,0) = u^0(x), \quad x \in \Omega,$$

and boundary conditions of Dirichlet type

$$u(x,t) = u_D(x,t), \quad x \in \partial\Omega_D, \quad t > 0,$$

and Neumann type

$$\partial_n u(x,t) = u_N(x,t), \quad x \in \partial\Omega_N, \quad t > 0,$$

or Robin type

$$\partial_n u(x,t) + c(x,t) u_(x,t) = u_R(x,t), \quad x \in \partial\Omega_R, \quad t > 0,$$

where $\partial\Omega = \partial\Omega_D \cup \partial\Omega_N \cup \partial\Omega_R \cup \Gamma$, $\operatorname{meas}\Gamma = 0$.

These types of equations occur in porous media flow, heat transport in flowing water, propagation of epidemics or carrier transport in semiconductors. They are advection dominated because their so-called *global Péclet number*

$$\mathrm{Pe} := \frac{\|\boldsymbol{v}\|_\infty \mathrm{diam}(\Omega)}{\|\boldsymbol{D}\|_\infty}$$

is significantly larger than 1.

Many difficulties are encountered in the numerical approximation of advection dominated diffusion problems. This is due to the nature of many approximation methods, which have specific drawbacks. Duffision problems can efficiently be solved with variational approximations (Galerkin methods), but these fail dramatically when applied to hyperbolic problems, like pure advection problems. The reason is that hyperbolic problems entail discontinuities, and many methods break down under such circumstances. Hyperbolic problems can be solved with high-resolution finite volume methods, where appropriate numerical flux functions are used, but these are not as practical for diffusion problems. Hybrid methods have been developed over the past decades; for an overview see [43], Chapter 9, where StreamLine-Diffusion, Lagrange-Galerkin and Finite Volume methods are presented. See also [50], Chapter 7, for finite volume methods (there called generalized upwind difference schemes). These methods are in full development.

We have chosen a different approach. Instead of developing a method that solves the advection dominated diffusion problem, we intend to use the broad knowledge that exists on solving advection problems on one hand, and diffusion problems on the other. A technique called operator splitting makes this possible: the original problem is split in two: one purely hyperbolic problem, and one parabolic problem.

The number of existing numerical methods is huge. In the choice of methods, we were always guided by the practical example we wanted to solve, the dual-well experiment. We will consider methods which are suitable for the dual-well. We are aware that many other methods exsist, and that there might be better choices.

This Chapter starts with an overview of the mathematical tools needed to solve (2.1): the operator splitting method, Riemann solvers, and the finite volume method. In Section 2.2 we present the dual-well problem. The next two Sections apply mathamatical tools to the dual-well problem, first without and then with adsorption. Section 2.5 is devoted to the convergence of the numerical approximation. Finally, in Section 2.6 numerical experiments are given.

2.1 Mathematical tools

2.1.1 Operator splitting

Operator splitting is the technique of dividing a complicated differential equation into several simpler parts. The corresponding methods are called *operator splitting methods* or *fractional steps methods*. The splitting of a differential equation can be done according to the subtype of the problems: diffusion, reaction, advection, source terms. It can also be done according to the dimensions, so called *dimensional splitting*.

Dimensional splitting

Dimensional splitting is typically used to reduce mutidimensional hyperbolic problems to one dimensional problems. Consider the two dimensional conservation law

$$\partial_t u + \partial_x f(u) + \partial_y g(u) = 0, \quad u(x,y,0) = u^0(x,y). \tag{2.2}$$

Denote by $\mathcal{T}_t^{f,x} u^0$ the solution of

$$\partial_t v + \partial_x f(v) = 0, \quad v(x,y,0) = u^0(x,y),$$

where y is a passive parameter. Similarly, let $\mathcal{T}_t^{g,y} u^0$ be the solution of

$$\partial_t w + \partial_y g(w) = 0, \quad w(x,y,0) = u^0(x,y),$$

where x is a passive parameter. The idea of dimensional splitting is to approximate the solution of (2.2) at $t = n\Delta t$, $(n \in \mathbb{N}_0)$, by

$$u(x,y,n\Delta t) \approx \left[\mathcal{T}_t^{g,y} \circ \mathcal{T}_t^{f,x}\right]^n u^0(x,y). \tag{2.3}$$

The above is called a semi-discrete splitting method, as the operators $\mathcal{T}$ are considered to produce exact solutions of the PDE's. If the operators correspond to a numerical method approximating the solution of the corresponding PDE, we call the approximation a fully-discrete splitting method. It can be shown that the semi-discrete dimensional splitting produces a sequence of functions for $\Delta t \to 0$ that converges to a solution of (2.2) in a weak sense, and that it preserves stability, meaning that if v is a solution of a problem with different data,

$$\partial_t v + \partial_x \tilde{f}(v) + \partial_y \tilde{g}(v) = 0, \quad v(x,y,0) = v^0(x,y),$$

we have that

$$\|u(\cdot,t) - v(\cdot,t)\|_1 \leq \|u_0 - v_0\|_1 + Ct \max\left(\|f - \tilde{f}\|_{\mathrm{Lip}}, \|g - \tilde{g}\|_{\mathrm{Lip}}\right). \qquad (2.4)$$

For more details, see [27].

Convergence rate

It is important to give extra attention to the convergence rate. The intrinsic error of the dimensional splitting is of the order $\sqrt{\Delta t}$, see [27], Theorem 4.8. So one might think this method performs much worse than other numerical methods which are of first or higher order. This is not the case, as in the operator splitting method, the timestep Δt is not bounded by a CFL condition. The CFL condition in advection problems states that the timestep must relate to the space grid in such a way that in one step the flow only reaches a point of the space grid from points which are in the stencil of the numerical method. Thus, a method that calculates the value in x_i from values in x_{i-1}, x_i, x_{i+1} in the previous timestep, must have a timestep such that the flow field indeed only carries information from these neighboring points, i.e. $\boldsymbol{v}\Delta t \leq \Delta x$. Formally we introduce, see [48],

Definition 2.1.1 (CFL Condition). *A numerical method can be convergent only if its numerical domain of dependence contains the true domain of dependence of the* PDE, *at least in the limit as Δt and Δx go to zero.*

Note that the CFL condition is only a necessary condition for convergence and stability, and is not always sufficient.

In advection dominated problems the CFL condition makes computation very hard, as a small grid implies an extremely small timestep. The operator splitting does not have this limitation, so one can split at time steps up to 10-15 times the CFL numbers. This makes it a fast method, usefull in advection dominated problems, or in problems where the computations need to be fast, like inverse problems.

It must be noted that the numerical method used to solve the split problem, does have to satisfy the CFL condition. For pure hyperbolic problems however, special methods like front tracking, based on the Riemann problem, can be used. These methods do not have a CFL type of limitation. Also pure diffusion problems do not exhibit this limitation.

From the above discussion operator splitting techniques might appear to be less usefull in non advection dominated problems. However, even then they might be handy to split the original problem in less complicated parts. Naturally, the great advantage of being able to use a timestep which is much larger than in other numerical methods is not an argument anymore.

Logical operator splitting: diffusion

In logical operator splitting, the original PDE is split along logical parts: diffusion, advection, source terms, reaction, etc. Consider an advection diffusion equation and initial condition

$$\partial_t u + \sum_{j=1}^{m} \partial_{x_j} f_j(u) - \mu \Delta u = 0, \quad u(x,0) = u^0(x), \tag{2.5}$$

where $\Delta u = \sum_j \partial_{x_j^2} u$, and where μ is a constant. Denote by $\mathcal{T}_j(t)u^0$ the solution of

$$\partial_t v + \partial_{x_j} f_j(v) = 0, \quad v(x,0) = u^0(x),$$

and by $\mathcal{D}(t)u^0$ the solution of

$$\partial_t w = \mu \Delta w, \quad w(x,0) = u^0(x).$$

The idea of operator splitting is to approximate the solution of (2.5) at $t = n\Delta t$, $(n \in \mathbb{N}_0)$, by

$$u(x, n\Delta t) \approx \left[\mathcal{D}(\Delta t) \circ \mathcal{T}_m(\Delta t) \circ \ldots \circ \mathcal{T}_1(\Delta t)\right]^n u^0(x). \tag{2.6}$$

For this semi-discrete operator splitting, the convergence to a weak solution u of (2.5) can be proved, see [27].

In logical operator splitting, it is not necessary to solve the sub problems in the same timestep as the operator splitting timestep Δt. Let $\Delta t = l\Delta\tau$, $l \geq 1$ an integer. Then, another slightly different operator splitting method is to approximate the solution of (2.5) at $t = n\Delta t$ by

$$\tilde{u}(x, n\Delta t) \approx \left[\mathcal{D}(\Delta t) \circ \left[\mathcal{T}_m(\Delta\tau) \circ \ldots \circ \mathcal{T}_1(\Delta\tau)\right]^l\right]^n u^0(x). \tag{2.7}$$

So far, we described only the semi-discrete method. In the fully discrete method, the operator $\mathcal{D}(\Delta t)$, can again be approximated by a numerical method that uses several timesteps to approximate the diffusion over Δt.

Formal analysis for linear problems

In general, consider the linear PDE of the form

$$\partial_t u = (\mathcal{A} + \mathcal{B})u,$$

where $\mathcal{A}, \mathcal{B}$ may be differential operators. For simplicity suppose they do not depend explicitely on t. We have that

$$\partial_t^j u = (\mathcal{A} + \mathcal{B})^j u.$$

If $\mathcal{A}, \mathcal{B}$ depend on t we would have to use the product rule and terms like $\partial_t \mathcal{A}$ would also appear. As the operators do not depend on t we can write the solution at time $t = \Delta t$ using Taylor series as

$$\begin{aligned} u(x, \Delta t) &= u(x,0) + \Delta t(\mathcal{A} + \mathcal{B})u(x,0) + \tfrac{1}{2}(\Delta t)^2(\mathcal{A} + \mathcal{B})^2 u(x,0) + \dots \\ &= \sum_{j=0}^{\infty} \frac{(\Delta t)^j}{j!}(\mathcal{A} + \mathcal{B})^j u(x,0) \\ &= e^{\Delta t(\mathcal{A}+\mathcal{B})} u(x,0). \end{aligned} \tag{2.8}$$

Denoting by $\tilde{u}$ the solution obtained by the fractional step method, we have

$$\tilde{u}(x, \Delta t) = e^{\Delta t \mathcal{B}} e^{\Delta t \mathcal{A}} u(x,0).$$

Therefore, the *splitting error* is given by

$$u(x, \Delta t) - \tilde{u}(x, \Delta t) = \left(e^{\Delta t(\mathcal{A}+\mathcal{B})} - e^{\Delta t \mathcal{B}} e^{\Delta t \mathcal{A}} \right) u(x,0).$$

This can can be calculated by a Taylor series expansion, for $\tilde{u}$

$$\begin{aligned} \tilde{u}(x, \Delta t) &= \left(\boldsymbol{I} + \Delta t \mathcal{B} + \tfrac{1}{2}(\Delta t)^2 \mathcal{B}^2 + \dots\right)\left(\boldsymbol{I} + \Delta t \mathcal{A} + \tfrac{1}{2}(\Delta t)^2 \mathcal{A}^2 + \dots\right) u(x,0) \\ &= \left(\boldsymbol{I} + \Delta t(\mathcal{A} + \mathcal{B}) + \tfrac{1}{2}(\Delta t)^2(\mathcal{A}^2 + 2\mathcal{B}\mathcal{A} + \mathcal{B}^2) + \dots\right) u(x,0). \end{aligned}$$

Comparing with (2.8) yields

$$u(x, \Delta t) - \tilde{u}(x, \Delta t) = \tfrac{1}{2}(\Delta t)^2(\mathcal{A}\mathcal{B} - \mathcal{B}\mathcal{A})u(x,0) + (\mathcal{A}\mathcal{B} - \mathcal{B}\mathcal{A})\mathcal{O}((\Delta t)^3). \tag{2.9}$$

Hence, the splitting error depends on the *commutator* $\mathcal{A}\mathcal{B} - \mathcal{B}\mathcal{A}$ and is zero up only in the special case when the differential operators $\mathcal{A}$ and $\mathcal{B}$ commute. Over all $\frac{T}{\Delta t}$ time steps, we arrive at a method that is first order accurate, even if the subproblems are solved exactly. Notice that the order of accuracy obtained here only holds for smooth solutions. In the general case a lower order of accuracy will be observed.

The above form of operator splitting is sometimes called *Godunov splitting.* One can obtain second order accuracy with a slight modification called *Strang splitting.* It is given formally by

$$\overline{u}(x, \Delta t) = e^{\frac{1}{2}\Delta t \mathcal{A}} e^{\Delta t \mathcal{B}} e^{\frac{1}{2}\Delta t \mathcal{A}} u(x, 0).$$

Practically the same accuracy is observed. This is due to the fact that the coefficient of the $\mathcal{O}(\Delta t)$ term may be much smaller than the coefficient in the second-order term. Strang splitting is often used in dimensional splitting. For logical splitting it might be less favourable when also boundary conditions need to be taken into account: there is little use in diffusion of a contaminant when not yet the total influx over Δt has occured along the edges.

Objective

We will apply operator splitting to (2.1) and prove its convergence. The main difference with previously published results is the fact that we consider a retardation $\phi(u)$ and that we will consider numerical methods which solve the split problem in this setting. Thus, we don't lean upon the transformation to a new variable $w = \phi(u)$, which would bring the problem in a form like (2.2) or (2.5). For the semi-discrete method this distinction has no impact: results for w are transferable to results for $\phi(u)$, as the operators considered are exact solutions. For the fully discrete method the numerical method approximating the problem in w is different from the one in u.

Another difference with standard convergence proofs concerns the domain Ω. Most proofs are on infinite domains. We will consider a bounded domain, with appropriate boundary conditions. The correct splitting of the boundary conditions to the sub problems of the original problem is non-trivial, and has to be done with care.

We will prove the convergence of the fully discrete operator splitting in Sec. 2.5.

2.1.2 The Riemann problem

Having split the problem in an advection and a diffusion problem, we need appropriate numerical methods for each of them. We start with the hyperbolic problem. The most succesfull approximation methods are finite volume methods, especially the high resolution methods, see [48], and the front tracking method, see [27]. We will not use the finte volume method for the hyperbolic problem, but will use it for the parabolic problem, see Sec. 2.1.3.

We will apply the front tracking method, but in its most exact form: exact solution of the Riemann problem. In general, the front tracking method can solve all types of flux problems in a general way, based on a generalization of the Riemann problem. For the application we have in mind, it is possible to work with exact solutions. Therefore, the general front tracking method need not be applied. However, when more complicated fluxes must be considered, then all the results can be extended to encompass solutions obtained with the front tracking method.

We start with the definition for one dimensional conservation laws.

Definition 2.1.2 (The Riemann problem). *For conservation laws, the Riemann problem is the initial value problem*

$$\partial_t u + \partial_x f(u) = 0, \quad u(x,0) = \begin{cases} u_l & \text{for} \quad x < 0 \\ u_r & \text{for} \quad x \geq 0 \end{cases} \tag{2.10}$$

The reader is referred to [47, 71, 27] for a complete analysis of the Riemann problem. Here, we briefly recall the results.

Motivation

We rewrite Eq. (2.10) in nonconservative form

$$\partial_t u + f'(u)\partial_x u = 0, \tag{2.11}$$

and we define the *characteristic curve* as the solution of the differential equation

$$\frac{d}{dt}x(t) = f'\left(u\left(x\left(t\right),t\right)\right). \tag{2.12}$$

Proposition 2.1.1. *Along any characteristic curve defined by (2.12) the solution of (2.11) is constant.*

This follows from

$$\begin{aligned} \frac{d}{dt}u(x(t),t) &= u_1(x(t),t)\frac{d}{dt}x(t) + u_2(x(t),t) \\ &= u_1(x(t),t)f'\left(u\left(x\left(t\right),t\right)\right) + u_2(x(t),t) \\ &= 0. \end{aligned}$$

Here u_i denotes the partial derivative of u with respect to the i-th argument.

From (2.12) it follows that the characteristic curves will intersect for certain smooth initial conditions, at which moment the solution becomes a multi-valued function. This is illustrated in Fig. 2.1 where a convex flux function is given and the speed is drawn for a very sharp front (drawn as a shock, but here considered to be smooth with large value of the derivative). In Fig. 2.1 (ii-b) a muti-valued profile should develop. Physically, this is not acceptable. The solution for this dilemma is that the function is no longer smooth: shocks are formed. Therefore, shocks are essential in conservation laws, and the study of their basic form (the Riemann problem) is important.

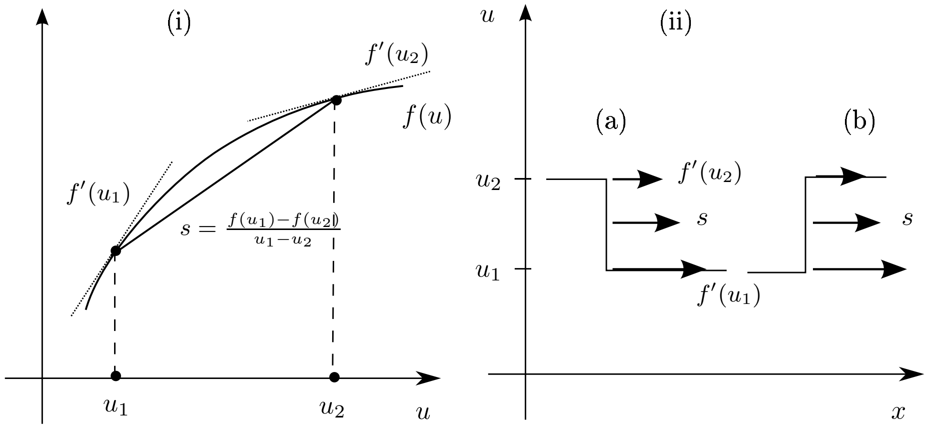


Figure 2.1: (i) Convex flux function $f(u)$ (ii) Two shocks with corresponding speeds

Weak solutions and acceptable shocks

If the solution $u(x,t)$ has shocks in $x = x_i$, then the original PDE (2.10) makes no longer sense. To include the discontinuous solutions we may consider ***weak solutions*** of the PDE.

Consider the problem

$$\partial_t u + \partial_x f(u) = 0, \quad u(x,0) = u^0(x), \quad x \in \mathbb{R}. \tag{2.13}$$

We define the set of *test functions*, C_0^1 as

$$C_0^1 = \{\phi \in C^1 : \{(x,t) \in \mathbb{R} \times [0,\infty) : \phi \neq 0\} \subset [a,b] \times [0,T]\}, \tag{2.14}$$

for some a, b, T. The functions in C_0^1 are said to have compact support in $\mathbb{R} \times [0,\infty)$. This fact is denoted by the subscript 0. We multiply (2.13) by $\phi \in C_0^1$ and integrate with respect to x from $-\infty$ to ∞, and with respect to t from 0 to ∞, to obtain

$$\int_0^\infty \int_{-\infty}^\infty [u\partial_t\phi + f(u)\partial_x\phi]\, dxdt + \int_{-\infty}^\infty u^0\phi^0\, dx = 0, \tag{2.15}$$

where $\phi^0 = \phi(x,0)$. All classical solutions will satisfy property (2.15), whereas all continuoulsy differentiable functions u satisfying (2.15) will also be classical solutions of (2.13). Also non differentiable functions might be solutions of (2.15). Therefore, we introduce

Definition 2.1.3. *If u statisfies (2.15) for all $\phi \in C_0^1$, u is said to be a weak solution of the initial-value problem (2.13).*

Having extended the definition of a solution to the discontinuous cases, we must specify which type of discontinuities are acceptable. We limit our interest to solutions which are smooth except across one or more curves in (x,t)-space, where they have jump discontinuities, which we call *shocks*. The following important result can be derived, see [27].

Proposition 2.1.2. *Let $C : x_C = x_C(t)$ be a smooth curve in the (x,t)-space accross which a weak solution u of (2.13) has a jump discontinuity. Let $P = (x_0,t_0)$, $t_0 > 0$, be any point of C, $s = \frac{dx_C}{dt}(t_0)$, and let u_l and u_r be the limit values of u from the left and the right of P, respectively. Then*

$$(u_l - u_r)\frac{dx_C}{dt} = f(u_l) - f(u_r). \tag{2.16}$$

The speed $s = \frac{dx_C}{dt}$ is the *speed of propagation of the discontinuity*. Eq. (2.16) is called the *jump condition* or the *Rakine-Hugoniot condition.* For the Riemann problem this implies that if the given jump is acceptable, the speed of the shock follows from (2.16). Indeed, the value of u cannot change left or right of the shock (it is the constant u_l and u_r respectively). However, it is important to note that physically unacceptable weak solutions will also satisfy the jump condition. To chose the acceptable solution one uses an *entropy condition.*

One of the most common entropy conditions is the so-called *viscous regularization*, where (2.13) is replaced by $\partial_t u + \partial_x f(u) = \epsilon \partial^2_{xx} u$, the so-called regularized equation.. Since in a physical situation there will always be some sort of dissipation, which the modeler neglected when writing the conservation law, we look for solutions (of the conservation law) that are the limit of the regularized equation as $\epsilon \to 0$. This is called the *vanishing viscosity solution*, given by

$$\partial_t u^\epsilon + \partial_x f(u^\epsilon) = \epsilon \partial^2_{xx} u^\epsilon, \quad u^\epsilon(x,0) = u^0(x), \quad x \in \mathbb{R}, \quad \text{as} \quad \epsilon \to 0. \tag{2.17}$$

The following entropy condition follows, [27].

Definition 2.1.4 (Entropy Condtion I). *The solution $u(x,t)$ of (2.15) containing a discontinuity propagating with speed s, is said to satisfy Entropy Condition I if*

$$s|k - u_l| < \operatorname{sign}(k - u_l)\left(f(k) - f(u_l)\right), \tag{2.18}$$

for all k stricktly between u_l and u_r.

In the case of a convex flux function f this condition reads as

$$f'(u_l) > s > f'(u_r). \tag{2.19}$$

This corresponds with Fig. 2.1, where (2.19) is satisfied for the case (ii-b), making this an acceptable shock, but not for case (ii-b).

The inequality (2.18) motivates another entropy condition, the *Kružkov entropy condition.* This is often more convenient as it combines the definition of a weak solution with that of the entropy condition.

Definition 2.1.5 (Kružkov Entropy Condition). *The solution $u(x,t)$ of (2.15) containing a discontinuity propagating with speed s, is said to satisfy the Kružkov Entropy Condition if*

$$\int\int \left(|u - k|\partial_t \phi + \operatorname{sign}(u - k)(f(u) - f(k))\partial_x \phi\right) \geq 0, \tag{2.20}$$

for all real constants k and all non-negative test functions $\phi \in C_0^\infty(\mathbb{R} \times (0,\infty))$.

Here C_0^∞ is the space of infinitely differentable functions with compact support in $\mathbb{R} \times (0,\infty)$. For the deduction of the Kružkov Entropy Condition, and the relationship with other entropy conditions, we refer to [27].

General solution of the Riemann problem

We now have all the ingredients to solve the Riemann problem. We can determine wether a shock is acceptable, and calculate its speed. If the shock is not acceptable, we have to follow the characteristics. A so-called *rarefaction* wave develops. Its form can easily be written down mathematically, but in practice it might be very hard to be determined. We look for a solution of the form $u = u(x,t) = w(x/t) = w(z)$, where $z = \frac{x}{t}$ is the only variable. Substitution in (2.10) leads to

$$-\frac{x}{t^2}w' + \frac{1}{t}f'(w)w' = 0 \quad \Rightarrow \quad z = f'(w).$$

If f' is strictly monotone, then $w = f^{-1}(z)$. In general f' needs to be replaced by a monotone function on the interval between $[u_l, u_r]$.

Definition 2.1.6 (Lower convex envelope). *The lower convex envelope $f_{\smile}$ of a function f in the interval $[a,b]$ is the largest convex function that is smaller than or equal to f in $[a,b]$, so*

$$f_{\smile}(u) = \sup\{g(u) |\ g \leq f \textit{ and } g \textit{ convex on } [a,b]\}.$$

For a function f, the lower convex envelope can be interpreted as an elastic rubber band stretched along but below f from $(a, f(a))$ to $(b, f(b))$,

The following proposition summerizes the results for the solution of the Riemann problem (2.10) in the case $u_l < u_r$.

Proposition 2.1.3. *In the case $u_l < u_r$, the solution of (2.10) is given by*

$$u(x,t) = w(z) = \begin{cases} u_l & \textit{for } x \leq f'_{\smile}(u_l)t, \\ (f'_{\smile})^{-1}(x/t) & \textit{for } f'_{\smile}(u_l)t \leq x \leq f'_{\smile}(u_r)t, \\ u_r & \textit{for } x \geq f'_{\smile}(u_r)t, \end{cases} \tag{2.21}$$

where $f_{\smile}$ denotes the lower convex envelope of f in the interval $[u_l, u_r]$, and $(f'_{\smile})^{-1}$ is the inverse of its derivative.

This proposition is illustrated in Fig. 2.2

If $u_l > u_r$, we can transform the problem to the case given above by the transformation $x \to -x$, from u_r to u_l. We need the lower convex envelope of $-f$, which is nothing else than the negative of the upper concave envelope $f_{\frown}$ from u_l to u_r. The latter is defined over the interval $[a,b]$ by

$$f_{\frown}(u) = \inf\{g(u) |\ g \geq f \text{ and } g \text{ concave on } [a,b]\}.$$

We have

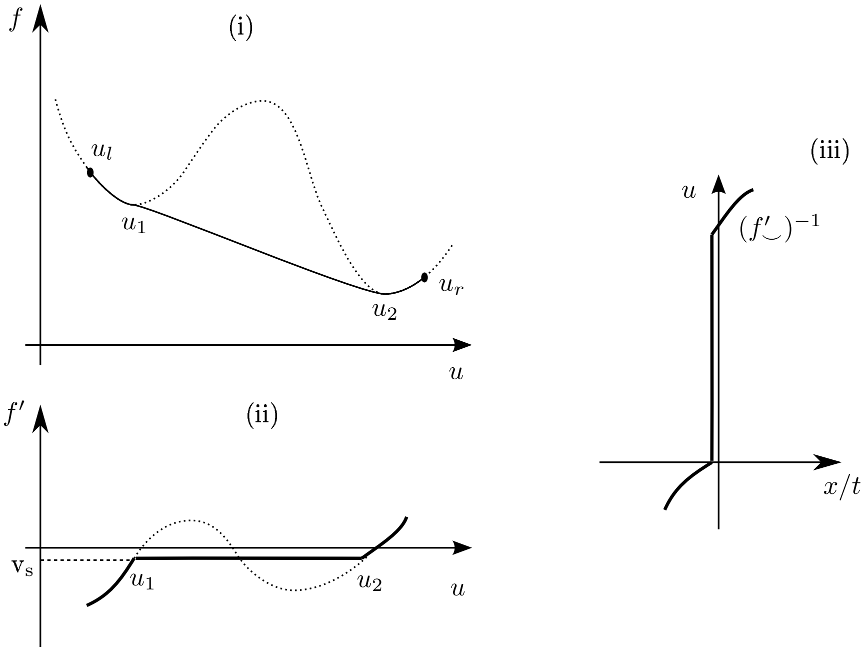


Figure 2.2: Determination of the solution of the Riemann problem. (i) Flux function f and it's lower convex envelope $f_\smile$. (ii) The derivatives, where $\mathrm{v_s} = \frac{f(u_1)-f(u_2)}{u_1-u_2}$. (iii) The solution $(f'_\smile)^{-1}(x/t)$, consisting of a shock and two rarefaction waves.

Proposition 2.1.4. *In the case $u_l > u_r$, the solution of (2.10) is given by*

$$u(x,t) = w(z) = \begin{cases} u_l & \textit{for } x \leq f'_\frown(u_l)t, \\ (f'_\frown)^{-1}(x/t) & \textit{for } f'_\frown(u_l)t \leq x \leq f'_\frown(u_r)t, \\ u_r & \textit{for } x \geq f'_\frown(u_r)t, \end{cases} \tag{2.22}$$

where $f_\frown$ denotes the upper concave envelope of f in the interval $[u_l, u_r]$, and $(f'_\frown)^{-1}$ is the inverse of its derivative.

Propositions 2.1.3 and 2.1.4 are valid as long as the envelope consists of a finite number of intervals where $f_\smile \neq f$ and $f_\frown \neq f$, alternating with intervals where f coincides with $f_\smile$ or $f_\frown$. This can be extended to the case where f is only *Lipschitz continuous*, see [27].

Front tracking

In the application, i.e. the dual-well problem, it will be possible to work with the exact Riemann solution, that is: the inverse $(f'_{\smile})^{-1}$ can be determined explicitely. This is in general not the case. One usally approximates f by a piecewise linear function f^δ. Then, the solution consists only of shocks which are known exactly. These shocks are tracked, hence the name *front tracking* for this method. It is important to show convergence of the solution obtained in this way, to the solution of the original problem. See [27] for more details.

The method can be extended to higher dimensions. A suitable procedure is to apply dimensional operator splitting.

2.1.3 Finite volume methods

The most popular difference methods are typically *finite difference methods* and *finite element methods*. However, there is a third popular discretization method, the *finite volume method*. Development of this method started already in 1960 (Forsythe and Wasow). It includes ideas from both finite difference and finite element methdos, and is therefore sometimes called a *generalized finite difference method*. For an extensive discussion see [43].

We start our treatment with the elliptic case. As mentioned before, the finite volume method is also very succesfull in discretizing hyperbolic equations. Therefore, it is often used in fluid flow problems.

Second order linear elliptic differential equation

Consider an equation of the form

$$Lu := -\nabla \cdot (\boldsymbol{K}\nabla u - \boldsymbol{c}u) + ru = f, \tag{2.23}$$

where $\boldsymbol{K} : \Omega \to \mathbb{R}^{d\times d}$, $\boldsymbol{c} : \Omega \to \mathbb{R}^d$ and $r, f : \Omega \to \mathbb{R}$. For simplicity we restrict ourselves to the case $r = 0$ and $d = 2$.

In order to derive the finite volume discretization, the domain Ω will be subdivided into M subdomains Ω_i, forming a *partition* of Ω, with each Ω_i open, simply connected, and polygonally bounded, and with $\Omega_i \cap \Omega_j = \varnothing$ $(i \neq j)$ and $\cup_{i=1}^M \overline{\Omega}_i = \overline{\Omega}$. These subdomains are called *control volumes* or *control domains.*

Next, we integrate (2.23) over each control volume Ω_i, and apply Gauss's divergence theorem, to get

$$\int_{\partial\Omega_i} \boldsymbol{\nu} \cdot (\boldsymbol{K}\nabla u - \boldsymbol{c}u)\; d\sigma = \int_{\Omega_i} f\; dx, \quad i \in \{1, \ldots, M\}, \tag{2.24}$$

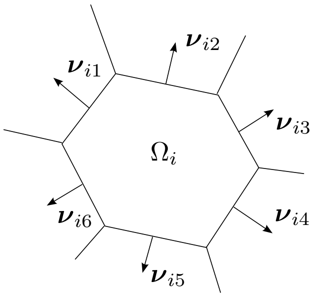


Figure 2.3: Control Volume for the finite volume method in 2 dimensions.

where $\boldsymbol{\nu}$ denotes the outer unit normal to $\partial\Omega_i$. As the control volumes are polygonally bounded, the left hand side can be rewritten as a sum of simple line integrals

$$\sum_{j=1}^{n_i} \int_{\Gamma_{ij}} \boldsymbol{\nu}_{ij} \cdot (\boldsymbol{K}\nabla u - \boldsymbol{c}u)\; d\sigma = \int_{\Omega_i} f\; dx, \quad i \in \{1, \ldots, M\}, \tag{2.25}$$

where n_i is the number of straight-line segments Γ_{ij} of the boundary of Ω_i, with normal $\boldsymbol{\nu}|_{\Gamma_{ij}} =: \boldsymbol{\nu}_{ij}$, a constant vector, see Fig. 2.3.

In the final step, the integrals appearing in (2.25) are appoximated. This can be done in many different ways, and so different final discretizations are obtained.

An important distinguishing condition between finite volume methods is the position of the unknowns with respect to the control volumes. On the one hand there are the *cell-centred* methods where the unknowns are associated with the control volumes (eg. a function value at some interior point). On the other hand there are the *cell-vertex* finite volume methods where the unknowns are located at the vertices of the control volumes.

The main advantages of the method are, see e.g. [43]:

- Flexible geometry of the domain and admissibility for unstructered grids.
- Simple assembling
- Conservation of certain laws can be garanteed locally.
- Easy linearization of nonlinear problems

- Simple discretization of boundary conditions

The drawbacks are

- Smaller range of application than finite element or finite difference methods.

- Difficulties in design of higher order methods.

- In higher spatial dimensions ($d \geq 3$), construction of general types of control volumes can be complex and time-consuming

- Difficult mathematical analysis (stability, convergence, ...).

In the dual-well problem that we will consider, small amounts of contaminant need to be tracked. Conservation of mass can be garanteed locally in finite volume methods. For this reason we will use finite volume methods. Accuracy will be obtained by reducing our spatial gridsize. This is a consequence of our choice for operator splitting.

A second important reason to use a finite volume method is the fact that we use a Riemann solver for the hyperbolic problem obtained in the operator splitting. Therefore, our initial condition for the diffusion problem will be a piecewise constant initial profile (certainly when front tracking is used, otherwise also rarefaction waves occur). This is compatible with a cell-centered finite volume method. The solution obtained by a cell-centered finite volume method shows again a piecewise constant profile, the ideal starting point for a general Riemann problem. So these two methods can be combined perfectly.

The integrals appearing in (2.25) will be approximated by central difference formulas, as discussed further.

Extension to parabolic differential equations and nonlinearity

The extension to parabolic differential equations is relatively easy. First, the spatial discretization is constructed. Then, a suitable time discretization is chosen. Typically, backward or implicit Euler will be used. This will give rise to a matrix equation that needs to be solved at all discrete time points.

Nonlinearity can be taken into account by many different methods, each of which depending strongly on the specific type of nonlinearity considered. We postpone this discussion to Section 2.4.4.

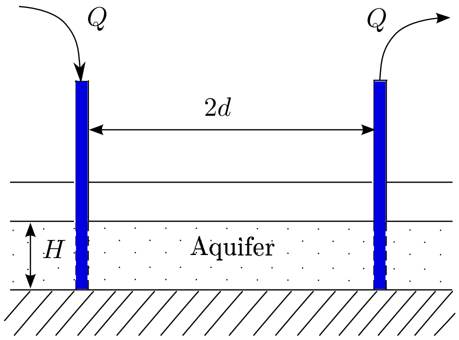


Figure 2.4: Aquifer with dual-well. One recharge well and one pumping well.

2.2 The dual-well as a practical example

The practical setup we want to model in detail is the dual-well experiment. For general terminology concerning groundwater flow and modeling, we refer to Appendix B. This Section is organized as follows. First, the physical background is given, demonstrating the relevance of the dual-well experiment. Next, the flow field is determined, and finally, the contaminant transport model is developed. The results obtained in this Section have been published jointly with Dr. D. Constales and Prof. J. Kačur in [15].

2.2.1 Physical background

The dual-well test, or doublet tracer test, is a field experiment consisting of an injection well and extraction well of equal strength. They are used for determining model characteristics, [20, 63, 69, 76] over a global scale typically not attainable in a laboratory. The scale of the experminents is typically 4 to 20 meters, see Fig. 2.4. When steady state conditions are achieved for the flow field, a pulse or step input tracer is introduced at the recharge well, and the break through curve of this tracer is monitored at the pumping well. The tracer can be a salt or a radioactive element (e.g. ^{131}I or ^{34}Br). Sometimes recirculation of the discharge water is employed.

If the aquifer is infinite, homogeneous and isotropic and when the effect of the natural flow velocity near the well can be neglected, it is easy to formulate a mathematical model for this system. The velocity distribution is the superposition of two radial flow velocity fields generated by a source and a sink. In [28], an analytical expression for the tracer concentration distribution in a dual-well test in integral form is given. It is found that dispersion in the extraction well mainly appears in the beginning of pumping, when the relative concentration is rather low. Already in 1971, Grove, [23], provided a program for calculating the dispersivity based on the interpretation of a dual-well test.

Determination of dispersivities can be done by modifying the longitudinal dispersivity, α_L, so that the model output agrees with the observed curve. If there are observation wells outside the line between the two wells, the transversal dispersivity, α_T, can be found by fitting the tracer concentration in the observation wells.

Many field tracer tests have been done in the literature. It has been found that the dispersivity values obtained by using mathematical models to interpret tracer injection tests are not constant, but depend on the scale of the test. This is not consistent with the original physical meaning of dispersivity. One explanation is that since the porous media in the field are all inhomogeneous and anisotropic, the larger the experiment scale is, the more heterogeneity is encountered, lifting the value of the dispersion. One way to handle this is the statistical theory of mass transport, see e.g. [18]. Another possible explanation is the use of an 'incorrect' model. If, for example, the mean flow changes over depth, a model with fixed mean flow will obtain unphysical dispersivities. Generally, this is the problem of using a two-dimensional model to a physical three-dimensional problem. When interpreting test data, we must carefully analyse the conditions of the test, and know the structure of the aquifer. Then, one can select the most suitable model.

One can argue that since the three-dimensional numerical models can take all practical conditions into account, it is most suitable for interpreting the results of field experiments. Nevertheless, we will deduce a new two-dimensional model for the dual-well test. First of all, a two-dimensional model is sensible in many cases, as the distances involved are still small (4 to 20m). Secondly, the model developed will have small numerical errors and will be stable and convergent. Moreover, it will be very time efficient, an important property for parameter identification. In Section 2.4 the model will be extended for nonlinear adsorption. This will be done in such a way that convergence can be shown. As far as we know, similar results don't exist in the literature.

As final argument, we point out the specific difficulties in groundwater modeling. The subsurface is a complex medium. Applying complicated models that depend on numerous parameters, with no idea of the value of these and no ability to validate them, has little use. Therefore, in commercial groundwater modeling, still many very basic models, like analytic element modeling, [24], are used which depend on few *global* parameters such as the dispersivity. A good estimation of these parameters with models based on in situ experiments, like the dual-well, is important. The model we develop is a very good approximation in a limited set of practical situations and willl be valuable in a larger set of environments (like non-homogeneous ones), if used properly.

2.2.2 Flow field of the dual-well

We consider an infinite, homogeneous and isotropic aquifer of height H, with two wells. The well at position $(-d, 0)$ is an extraction well with discharge rate (pumping rate) Q_1 (> 0), and the well at position $(d, 0)$ is a recharge well with pumping rate $Q_2 = -Q_1$, see Fig. 2.4. We can use the Dupuit-Forchheimer approximation, simplifying the model to two dimensions. This will be valid if the head gradients are not large, which can be garanteed with a dual-well. In radial coordinates (the well is situated at $r = 0$), the flow potential for the extraction well is then, see (B.11),

$$\Phi_1(r) = \frac{Q_1}{2\pi} \ln r + C_w,$$

where C_w is a constant that has to be determined from the boundary conditions. The flow equals $Q_r = -\partial_r \Phi_1(r)$ and the Darcy velocity is $q = -(1/h_{\text{eff}})\partial_r \Phi_1$, where $h_{\text{eff}} = \min(h, H)$, H the height of the aquifer and h the piezometric head (counted from the bottom of the aquifer).

The flow potential of the well doublet can be found by superposition of the two wells. In cartesian coordinates this leads to

$$\Phi(x, y) = \frac{Q_1}{4\pi} \ln \frac{(x+d)^2 + y^2}{(x-d)^2 + y^2} + \Phi_0, \tag{2.26}$$

where the constant Φ_0 must be determined from the boundary conditions. We will write the discharge rate Q_1 as Q. Eq. (2.26) satisfies the steady state equation

$$\Delta \Phi = 0, \tag{2.27}$$

and can be completely determined by prescribing the flow potential in a reference point (x_0, y_0), implying a value for the constant Φ_0. Note that when two head values, h_1, h_2, are given at different points, the value of Q and Φ_0 can be determined. In Fig. 2.5 the resulting flow potential is plotted, together with the flow lines which are perpendicular to the equipotential lines of Φ.

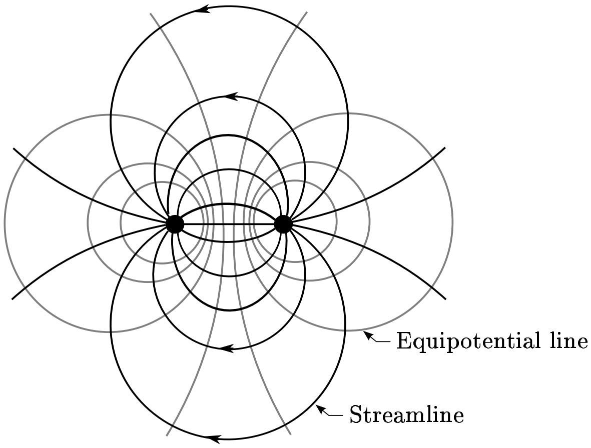


Figure 2.5: Flow potential and flow lines for a dual well, modeled by a point source and sink.

In the case that the regional flow in the aquifer can not be neglected, an extra term must be added, see Section B.5, giving

$$\Phi(x,y) = -Q_0 x + \frac{Q}{4\pi}\ln\frac{(x+d)^2+y^2}{(x-d)^2+y^2} + \Phi_0, \tag{2.28}$$

where Q_0 is the uniform flow field in the x-direction. So, for $Q_0 > 0$, there is regional flow from left to right. Only for sufficiently large Q, water injected at the recharge well will reach the pumping well. It can be shown that this is the case when $Q > \pi d Q_0$, see [24]. For contaminant transport, it seems more usefull to set up the two wells such that $Q_0 < 0$, which will always allow recirculation to occur.

In fact, solution (2.26) is not realistic in the neighborhood of the wells, since there the Dupuit-Forchheimer approximation is strongly violated. This can be solved by considering two wells separated by a distance D with given radii r_1, r_2, $(D > r_1 + r_2)$, and by prescribing the head values reached on their boundary under steady-state conditions. We use the notation $B_r(a,b)$ for the ball with radius r and center in (a,b), and $\partial B_r(a,b)$ for its boundary. Thus, we are considering the equation

$$\Delta\Phi = 0 \qquad \text{in } \Omega = \mathbf{R}^2 \setminus B_{r_1}(-d,0) \cup B_{r_2}(d+c,0), \tag{2.29}$$

where $D = 2d + c$, along with the boundary conditions $\Phi = \Phi_1$ on $\partial B_{r_1}(-d, 0)$ and $\Phi = \Phi_2$ on $\partial B_{r_2}(d + c, 0)$. This is the Dirichlet problem for an outer domain. Due to the symmetry along the x-axis, we solve (2.27) in the upper half-plane, see Fig. 2.6, with Dirichlet conditions on the half-circles and a homogeneous Neuman condition on the parts of the x-axis bordering the domain (because of symmetry). This problem can be solved efficiently using conformal mapping, and, especially, bipolar transformation, [52], that transforms (2.29) into a rectangle $\widetilde{\Omega} = [0, \pi] \times [v^{(1)}, v^{(2)}]$, see also Fig. 2.6.

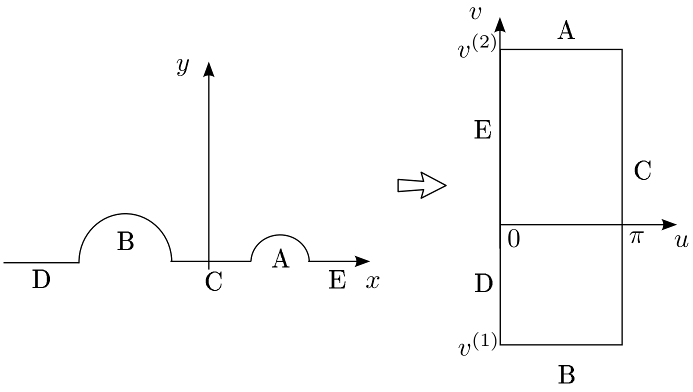


Figure 2.6: Boundary of the domain Ω in the (x, y) plane, and the domain $\widetilde{\Omega}$ in (u, v)-coordinates after bipolar transfomation. A is the injection well, B the extraction well.

Generally, the bipolar transformation is given by

$$x = \frac{\gamma}{2} \frac{\sinh v}{\cosh v - \cos u}, \quad y = \frac{\gamma}{2} \frac{\sin u}{\cosh v - \cos u}, \quad u \in [0, 2\pi), \quad v \in (-\infty, \infty), \tag{2.30}$$

where the value γ can be chosen. The transformation has the special property that curves with constant u or v are circles in the xy-space. This follows from the identities

$$x^2 + \left(y - \frac{\gamma}{2 \cot u}\right)^2 = \frac{\gamma^2}{4} \frac{1}{\sin^2 u} \tag{2.31}$$

$$\left(x - \frac{\gamma}{2 \coth v}\right)^2 + y^2 = \frac{\gamma^2}{4} \frac{1}{\sinh^2 v}. \tag{2.32}$$

Thus, a constant v-value leads to a circle with radius $\frac{\gamma}{2}\frac{1}{\sinh v}$ and center on the x-axis located in $x = \frac{\gamma}{2\coth v}$. We can choose two v values, $v^{(1)}$ and $v^{(2)}$, that will correspond with our given well radii, and we can choose γ and c, so as to center the corresponding circles in the correct position. This leads to the system

$$\begin{cases} 2d + c & = D \\ \sinh v^{(1)} & = -\frac{\gamma}{2r_1} \\ \sinh v^{(2)} & = \frac{\gamma}{2r_2} \\ \frac{\gamma}{2\coth v^{(1)}} & = -d \\ \frac{\gamma}{2\coth v^{(2)}} & = d + c. \end{cases}$$

Simplifying, we find that the unknowns $v^{(1)}$, $v^{(2)}$, d , c and γ follow from

$$\begin{cases} \sinh v^{(1)} & = -\frac{\gamma}{2r_1} \\ \sinh v^{(2)} & = \frac{\gamma}{2r_2} \\ d & = \frac{1}{2}\sqrt{\frac{2{r_1}^2D^2+{r_1}^4-2\,{r_1}^2{r_2}^2+{r_2}^4-2\,{r_2}^2D^2+D^4}{D^2}} \\ \gamma & = 2\sqrt{d^2 - r_1^2} \\ c & = D - 2d. \end{cases}$$

With these values, the bipolar transformation (2.30) will transform the domain Ω into the rectangle $[0,\pi] \times [v^{(1)}, v^{(2)}]$, where we need to solve the Laplace equation (2.27), see Fig. 2.6. Note that when $r_1 = r_2$, we have $d = D/2$ and $c = 0$. The Laplace equation under the bipolar transformation is given by

$$\frac{4}{\gamma}\left(\cosh v - \cos u\right)^2 \left(\frac{\partial^2\tilde{\Phi}}{\partial u^2} + \frac{\partial^2\tilde{\Phi}}{\partial v^2}\right) = 0, \quad (u,v) \in [0,\pi] \times [v^{(1)}, v^{(2)}]$$

where $\tilde{\Phi}(u,v) = \Phi(x,y)$. This reduces again to the Laplace equation, now in (u,v)-variables. In bipolar coordinates, the Laplace equation is separable. The solution is uniquely defined if on all boundaries a boundary condition (Dirichlet, Neumann, Robin) is given. In our approach, we have a homogeneous Neumann condition in $u = 0$ and $u = \pi$, and a Dirichlet condition at the inflow and outflow. The solution is given by

$$\tilde{\Phi}(v) = Av + B \tag{2.33}$$

where A and B are determined by the boundary equations

$$Av^{(1)} + B = \Phi_1, \qquad Av^{(2)} + B = \Phi_2. \tag{2.34}$$

In this way we obtain a simple exact solution of the flow problem in the domain $\tilde{\Omega}$ and, transforming back, in Ω. Here, the equipotential curves of Φ in Ω create the horizontal lines in $\tilde{\Omega}$ (parallel with the u-axis) and the streamlines, which are orthogonal to them, create the vertical lines, parallel with the v-axis.

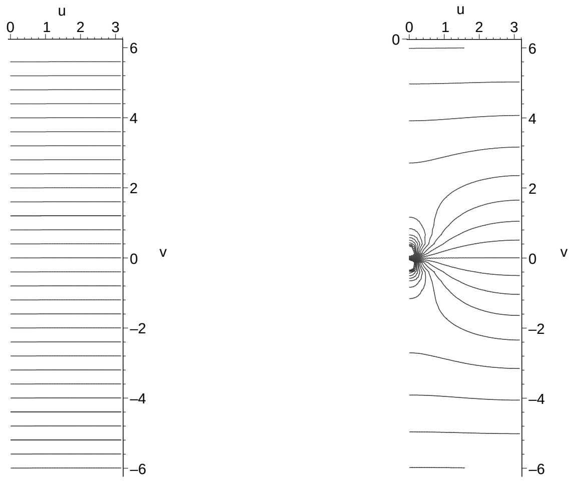


Figure 2.7: Flow equipotential lines for a dual well in uv-coordinates. Left: no regional flow. Right: regional flow aligned to dual-well flow, recirculation between the dual wells: flow from top to bottom.

The potential (2.26) is a good approximation of the exact potential Φ determined from (2.29). Our solution is identical to a dual well constructed from two point sources, where these sources are set in $(-\gamma/2, 0)$, $(\gamma/2, 0)$:

$$\Phi(x, y) = \frac{Q}{4\pi} \ln \frac{(x + \gamma/2)^2 + y^2}{(x - \gamma/2)^2 + y^2} + \Phi_0 = Av + B = \tilde{\Phi}(u, v),$$

under the given transformation. This allows for introducing the background groundwater flow, i.e. the flow present in the subsurface independent of the

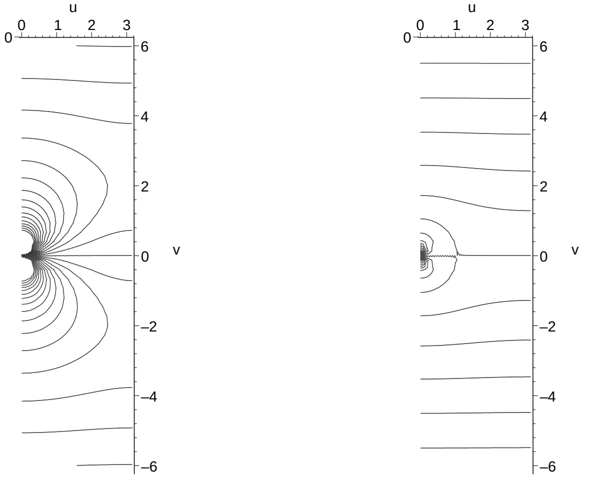


Figure 2.8: Flow equipotential lines for a dual well in uv-coordinates, regional flow against dual-well flow. Left: no recirculation between the dual wells, flow from top to infinity, and from infinity to bottom. Right: partial recirculation, part of top goes to bottom, and part goes to infinity.

wells, for which we suggest the general form

$$\begin{aligned} \Phi(x,y) &= \frac{Q}{4\pi}\ln\frac{(x+\gamma/2)^2+y^2}{(x-\gamma/2)^2+y^2}+\Phi_0-Q_0x \\ &= Av+B-C\frac{\gamma}{2}\frac{\sinh v}{\cosh v-\cos u} \\ &= \tilde{\Phi}(u,v). \end{aligned} \tag{2.35}$$

Here A, B and C must be determined from the given value of the flow potential in 3 points. Indeed, the head value will be no longer constant over the well boundaries for (2.35).

We know that the head h is related to the flow potential, see (B.9)-(B.10). Therefore, we can give the head values at inflow and outflow boundaries, and calculate Φ_1 and Φ_2 needed in (2.34). The curve $h(x,y)=H$ separates the confined and unconfined zone.

The seepage velocity $\mathbf{v}_s$ is given by

$$\mathbf{v}_s = -\frac{1}{h_{\text{eff}}\theta_0}\nabla\Phi, \tag{2.36}$$

where θ_0 is the porosity. Thus, the flow field is completely determined. We plot the equipotential lines in several set-ups. The flow lines will be perpendicular to them. In Fig. 2.7, Left, there is no regional flow. The injection well is at the top. The water flows from the top to the bottom, with flow lines parallel to the v-axis. In Fig. 2.7, Right, the regional flow is from $x = +\infty$ to $x = -\infty$, i.e., $Q_0 < 0$, so it goes from the injection to the extraction well. The resulting flow lines are from the top to the bottom, but around the point at infinity. There, we see the flow from $x = +\infty$, i.e. $(u, v) = (0, 0+)$, to $x = -\infty$, i.e. $(u, v) = (0, 0-)$.

In Fig. 2.8, the regional flow is from $x = -\infty$ to $x = +\infty$, and the type of flow depends on the strength of this flow field. To the right, we have the case of a strong regional flow, making circulation of water from the injection well to the extraction well impossible. To the left, the regional flow is weaker, and some of the water of the injection well can still reach the extraction well.

For the mathematical model below, we take $Q_0 = 0$, so $C = 0$. In practice, this means that the regional flow must be neglectible in comparison to the flow field of the dual-well. In this case, as will be seen, the simple form of the flow potential $\tilde{\Phi}$ can be exploited.

2.2.3 Mathematical model for contaminant transport

The transport equation for a contaminant/tracer has the form (see (B.14)),

$$\partial_t(h_{\text{eff}}C) = \nabla\cdot(h_{\text{eff}}\boldsymbol{D}\nabla C) - \nabla(h_{\text{eff}}\boldsymbol{v}C) + \frac{h_{\text{eff}}I}{\theta_0}, \tag{2.37}$$

where the porosity θ_0 is taken to be constant over the aquifer, and where $\boldsymbol{D}$ is the dispersivity tensor

$$D_{ij} = \{(D_0 + \alpha_T|\boldsymbol{v}|)\delta_{ij} + \frac{v_i v_j}{|\boldsymbol{v}|}(\alpha_L - \alpha_T)\}, \tag{2.38}$$

D_0 being the molecular diffusion and δ_{ij} the Kronecker symbol. Moreover,

$$\boldsymbol{v} = -\frac{1}{h_{\text{eff}}\theta_0}\nabla\Phi \tag{2.39}$$

and Φ is the solution of (2.29). The source term I models radioactive decay, adsorption, etc. At the moment we set $I = 0$. We will come back to this point in later Sections.

To use the analytical solution of the flow model in (2.37), we have to consider an unbounded domain Ω, which can cause many problems, in particular numerical errors. Thus we transform (2.37) into the domain $\tilde{\Omega}$ of the variables u and v, using the same transformation as for the flow potential. This was done first in [15]. It leads to complicated formulas and computations; we have relied on symbolic computation using the Maple package. Moreover, to get a good discretization, we must write the new governing equation in conservative form in the u and v variables. We briefly sketch the calculations, and then present the final results in the absence of source terms ($I = 0$).

A direct calculation leads to the following transformation for the derivatives,

$$\frac{\partial}{\partial x} = \frac{-2\sinh v \sin u}{\gamma}\frac{\partial}{\partial u} - \frac{2(\cos u \cosh v - 1)}{\gamma}\frac{\partial}{\partial v}, \tag{2.40}$$

$$\frac{\partial}{\partial y} = \frac{2(\cos u \cosh v - 1)}{\gamma}\frac{\partial}{\partial u} - \frac{2\sinh v \sin u}{\gamma}\frac{\partial}{\partial v}. \tag{2.41}$$

In the confined case ($h_{\text{eff}} = H$), using (2.39) and setting $\Phi(x, y) = \tilde{\Phi}(v) = Av + B$, we get

$$\begin{aligned} v_x &= 2\,\frac{(\cos(u)\cosh(v) - 1)\frac{d}{dv}\tilde{\Phi}(v)}{\theta_0 H \gamma} \\ v_y &= 2\,\frac{\sinh(v)\sin(u)\frac{d}{dv}\tilde{\Phi}(v)}{\theta_0 H \gamma} \\ |v| &= 2\sqrt{\frac{\left(\frac{d}{dv}\tilde{\Phi}(v)\right)^2(\cosh(v) - \cos(u))^2}{{\theta_0}^2 H^2 \gamma^2}}. \end{aligned}$$

Hence, (2.37) in the confined domain (where $h > H$, i.e., $h_{\text{eff}} = H$) yields:

$$\begin{aligned} \partial_t C &= \frac{4\lambda^2}{\gamma^3\theta_0 H}\left\{\partial_u\left[\left(D_0\theta_0 H\gamma + 2\alpha_T\lambda(\partial_v\widetilde{\Phi}(v))\right)\partial_u C\right] + \right. \\ &\quad \left. + \partial_v\left[\left(D_0\theta_0 H\gamma + 2\alpha_L\lambda(\partial_v\widetilde{\Phi}(v))\right)\partial_v C + \gamma(\partial_v\widetilde{\Phi}(v))C\right]\right\}, \end{aligned} \tag{2.42}$$

where $\lambda = \cosh v - \cos u$, and where $\partial_v\widetilde{\Phi}(v) = A$ from (2.33).

In the unconfined domain (where $h \leq H$, i.e., $h_{\text{eff}} = h(v)$), we use the relation between the head value and the flow potential, i.e. from (B.10),

$$\Phi(x, y) = \tfrac{1}{2}kh^2(x, y).$$

This allows to write (2.39) as

$$\boldsymbol{v} = -\frac{1}{h(x, y)\theta_0}\nabla(\tfrac{1}{2}kh^2(x, y)).$$

It then follows that

$$\begin{aligned} v_x &= \frac{2k}{\theta_0\gamma}(\cosh v \cos u - 1)\frac{\partial h(v)}{\partial v}, \\ v_y &= \frac{2k}{\theta_0\gamma}\sinh v \sin u \frac{\partial h(v)}{\partial v}, \\ |v| &= 2\sqrt{\frac{k^2 \frac{\partial h(v)}{\partial v}^2 (\cos u - \cosh v)^2}{\theta_0^2\gamma^2}}. \end{aligned}$$

Hence, (2.37) reads as

$$\begin{aligned} \partial_t C &= \frac{4\lambda^2}{\gamma^3\theta_0 h(v)}\left\{\partial_u\left[\left(D_0\theta_0 h(v)\gamma + 2\alpha_T\lambda(\partial_v\widetilde{\Phi}(v))\right)\partial_u C\right] + \right. \\ &\quad \left. + \, \partial_v\left[\left(D_0\theta_0 h(v)\gamma + 2\alpha_L\lambda(\partial_v\widetilde{\Phi}(v))\right)\partial_v C + \gamma(\partial_v\widetilde{\Phi}(v))C\right]\right\}, \end{aligned} \tag{2.43}$$

where we used $\partial_v\tilde{\Phi}(v) = kh(v)\partial_v h(v)$.

For simplicity, we shall write (2.42), (2.43) in the form

$$\partial_t C = g\{\partial_u(a\partial_u C) + \partial_v(b\partial_v C)\} + G\partial_v C, \quad \text{in } \Omega \tag{2.44}$$

where g, a, b and G are known functions depending on u and v:

$$g = \frac{4\lambda^2}{\gamma^3\theta_0 h_{\text{eff}}(v)}, \ \lambda = \cosh v - \cos u, a = D_0\gamma\theta_0 h_{\text{eff}}(v) + 2\alpha_T\lambda A$$

$$b = D_0\gamma\theta_0 h_{\text{eff}}(v) + 2\alpha_L\lambda A, \ G = A\gamma g,$$

and where Ω is a rectangle in the (u, v)-domain.

We consider the inflow boundary condition

$$C = C_0(t) \text{ on } \Gamma_1, \tag{2.45}$$

the symmetry boundary conditions

$$\partial_u C = 0 \text{ on } \Gamma_2 \cup \Gamma_4, \tag{2.46}$$

and the outflow boundary condition

$$\partial_v C = 0 \text{ on } \Gamma_3, \tag{2.47}$$

where $\Gamma_1 := (0,\pi) \times \{v = v^{(2)}\}$, $\Gamma_2 := \{0\} \times (v^{(1)}, v^{(2)})$, $\Gamma_3 := (0,\pi) \times \{v^{(1)}\}$ and $\Gamma_4 := \{\pi\} \times (v^{(1)}, v^{(2)})$. We consider the homogeneous initial condition

$$C((u,v),0) = 0. \tag{2.48}$$

The function $C_0(t)$ is the prescribed concentration at the inflow, which we choose to be constant, $C_0(t) = C^0$, or pulse shaped.

The outflow boundary condition follows from the assumption that the contaminant concentration inside and outside the well lateral wall are approximately equal during extraction,

$$\boldsymbol{n} \cdot \nabla C(x,y) = 0 \quad \text{on } \Gamma_3,$$

$\boldsymbol{n}$ being the normal unit vector on Γ_3. This is transformed to (2.47) using

$$n_x = \frac{\cosh v \cos u - 1}{\cos u - \cosh v}, \quad n_y = \frac{\sin u \sinh v}{\cos u - \cosh v},$$

and (2.40)-(2.41).

The inflow boundary condition allows for diffusive flux. As the molecular diffusion can usually be neglected, this diffusive flux follows from the molecular dispersion. For large molecular dispersion this will contribute to an extra amount of contaminant mass or, in the case of a pulse, a mass loss due to the injection well. This might be questionable, as molecular dispersion is a mechanical process and cannot cause diffusion against the flow direction, or from the well into the subsurface. Therefore, some authors, [49], suggest the advective boundary condition

$$\boldsymbol{n} \cdot (C(t)\boldsymbol{v} - \mathbf{D}\nabla C(t)) = C_0(t)\boldsymbol{n} \cdot \boldsymbol{v}, \text{ on } \partial B_{r_2}(d+c,0), \tag{2.49}$$

with $\boldsymbol{n}$ being the normal unit vector on $\partial B_{r_2}(d+c,0)$. This BC is of Cauchy type and expresses the fact that a fixed amount of contaminant mass will be present in the subsurface, independent of the dispersion coefficients. In the case that the flow potential is given by (2.33), (2.49) reduces to

$$\left(D_0\theta_0 h_{\text{eff}} + \frac{2\alpha_L\lambda}{\gamma}(\partial_v\widetilde{\Phi}(v))\right)\partial_v C + (\partial_v\widetilde{\Phi}(v))C = (\partial_v\widetilde{\Phi}(v))C_0(t), \text{ on } \Gamma_1, \tag{2.50}$$

taking into account (2.39) and (2.40)-(2.41) and

$$n_x = -\frac{\cosh v\cos u - 1}{\cos u - \cosh v}, \quad n_y = -\frac{\sin u \sinh v}{\cos u - \cosh v}.$$

Note that (2.50) will only differ significantly from (2.45) for D_0 or α_L sufficiently large.

2.3 Solution of advection dominated diffusion in a rectangle

We have reduced the contaminant transport problem of the dual-well well to a convection-diffusion equation with variable coefficients in a rectangle. We now present the numerical discretization.

2.3.1 Numerical approximation of (2.44)

To solve this convection-diffusion problem, we use time stepping and operator splitting. In each small time interval, the problem is split into 2 parts: *the transport problem* and *the diffusion problem*, see Section 2.1.1. More in detail, let $\tau = T/n$ be a time step and $C_i \approx C((u,v),t_i)$ for $i = 1,\ldots,n$. Given C_{i-1}, the relation

$$C_i = D^i(\tau)T^i(\tau)C_{i-1}, \quad \tau = t_i - t_{i-1},$$

determines C_i. The transport $T^i(\tau)$ corresponds to the solution ϕ_T of the transport equation

$$\partial_t\phi - G\partial_v\phi = 0, \tag{2.51}$$

with the inflow condition $\phi_T((u,v^{(2)}),t) = C_0(t)$ and the initial condition

$$\phi_T((u,v),t_{i-1}) = C_{i-1}.$$

The diffusion $D^i(\tau)$ is obtained by solving the diffusion equation

$$\partial_t\phi = g\left\{\partial_u(a\partial_u\phi) + \partial_v(b\partial_v\phi)\right\}, \tag{2.52}$$

with initial condition $\phi((u,v),t_{i-1}) = C_i^{1/2}$, where

$$C_i^{1/2} := T^i(\tau)C_{i-1} \equiv \phi_T((u,v),t_i).$$

Then, we set

$$C_i = D^i(\tau)C_i^{1/2} = D^i(\tau)T^i(\tau)C_{i-1} \equiv \phi((u,v),t_i).$$

The convergence of this approximation scheme is based on convergence results for operator splitting, see [16], [31], and our own result in Section 2.5.

We still must mention the boundary conditions, in particular how the original boundary conditions must be applied to the split problem. This is important to get the correct physical solution. On $\Gamma_2 \cup \Gamma_3 \cup \Gamma_4$ there is the homogeneous Neumann condition (2.46), (2.47) that must be split. During transport, there is no flux on $\Gamma_2 \cup \Gamma_3$, so no boundary condition is needed. Γ_4 is an outflow boundary, so a boundary condition is impossible. During the diffusion step, the boundary condition on $\Gamma_2 \cup \Gamma_3 \cup \Gamma_4$ is still the homogeneous Neumann condition.

It remains to determine how to split the boundary condition on Γ_1. If the boundary condition that needs to be satisfied is (2.45), we propose to take this Dirichlet condition as boundary condition for the transport part, as well as for the diffusion part. If the boundary condition is (2.50), we propose to split this condition. During transport we consider the Dirichlet condition $C(u,v^{(2)}) = C_0(t)$. In this manner all mass flux of contaminant into the domain during a time step has been realized during the transport step. Therefore, during the diffusion step, we consider the homogeneous Neumann condition $\partial_v C(u,v^{(2)}) = 0$. This approach guarantees the correct mass balance.

The space discretization is based on the cell-centered finite volume concept, with one unknown per cell, see Section 2.1.3. Let $\{u_i\}_{i=0}^N$ and $\{v_j\}_{j=0}^M$ be the nodal points for a (not necessarily equidistant) partitioning in u and v, respectively. We will construct a cell around these (u_i,v_j) points and the cell value will be the concentration value in this point. We generally take a non-equidistant v partitioning following from an equidistant x-partitioning along the x-axis between the two wells. In the points $\{u_i,v_j\}$ for $j=0$, the Dirichlet conditions for inflow concentration C are prescribed. We have $v_0 = v^{(2)}$, $v_M = v^{(1)}$, so $v_0 > v_M$, and $0 < u_0 < u_N < \pi$, see Fig. 2.9. Let $\{u_i,v_j\}$ be an inner point in $\tilde{\Omega}$. We define $\Delta u_+ = u_{i+1} - u_i$, $\Delta u_- = u_i - u_{i-1}$, $u_{i+1/2} = u_i + \Delta u_+/2$, $u_{i-1/2} = u_i - \Delta u_-/2$, $\Delta u = u_{i+1/2} - u_{i-1/2}$. We proceed analogously for v, where, $\Delta v_+ = v_{j-1} - v_j$, etc. Thus we obtain u_i-strips defined by $(u_{i-1/2},u_{i+1/2}) \times (v^{(1)},v^{(2)})$, and in these strips the finite volume

$V_{ij} = (u_{i-1/2}, u_{i+1/2}) \times (v_{j+1/2}, v_{j-1/2})$ corresponding to (u_i, v_j). For the edges of $\widetilde{\Omega}$ we set $u_{-1/2} \equiv 0$, $u_{N+1/2} \equiv \pi$, $v_{-1/2} \equiv v_0$ and $v_{M+1/2} \equiv v_M$.

2.3.2 Solution of the general transport problem

The solution of the transport problem (2.51) will be based on a piecewise constant initial profile $\phi_0(v)$, i.e., the solution of a multiple Riemann problem, which is obtained in analytical form. We shall solve (2.51) in the strip $(u_{i_0-1/2}, u_{i_0+1/2}) \times (v^{(1)}, v^{(2)})$, with shocks on the edges of the finite volume $v_0^e \equiv v^{(2)}$, $v_1^e \equiv v_{3/2}$, $v_2^e \equiv v_{5/2}$, $\ldots$, $v_{M-1}^e \equiv v_{M-1/2}$, $v_M^e \equiv v^{(1)}$, see Fig. 2.9. Denote $\phi_0(v) = U^j$ for $v \in (v_j^e, v_{j-1}^e)$. We transform (2.51), with $G = K\lambda^2/h_{\text{eff}}$,

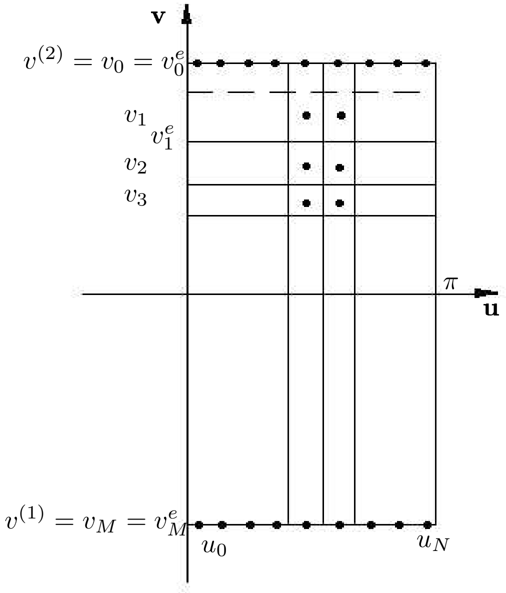


Figure 2.9: The (u, v) domain divided in strips and cells.

where K is a constant, using the new variable $y = y(v)$ where

$$y = \tilde{G}(v) = \int_{v^{(1)}}^{v} \frac{h_{\text{eff}} dv}{K\lambda^2}, \quad \lambda^2 = (\cosh(v) - p_{i_0})^2, \quad p_{i_0} = \cos u_{i_0}. \tag{2.53}$$

Then $\overline{\phi}(y,t) = \phi(v,t)$ satisfies

$$\partial_t \overline{\phi} - \partial_y \overline{\phi} = 0, \quad \overline{\phi}(y,0) = \phi_0(v).$$

Since λ is positive, the transformation is one to one. The solution can be written in the form

$$\overline{\phi}(y,t) = \overline{\phi}(y+t,0), \quad \text{or} \quad \overline{\phi}(y,0) = \overline{\phi}(y-t,t),$$

and, consequently, using the inverse $\tilde{G}^{-1} : y \mapsto v$, we obtain $\phi(v,t)$ from $\phi_0(v)$. Notice that we need not compute the inverse in all points y, since it follows that $\overline{\phi}(y,t)$ is piecewise constant with the values $\{U_j\}_{j=1}^{M}$. It is sufficient to compute

$$y_k = \int_{v^{(1)}}^{v_k^e} \frac{h_{\text{eff}} dv}{K\lambda^2}, \qquad \text{for } k = 0,\dots,M,$$

and then to shift it over the time step τ, and compute the inverse $\tilde{G}^{-1}(y_k - \tau) := \delta_k$ for $k = 0,\dots,M$. The solution $\phi(v,\tau)$ attains the constant value $U^{j_k} \in \{U_j\}_{j=1}^{N}$ in the interval (δ_k, δ_{k-1}).

The initial condition $\overline{\phi}(y,0) = \phi_0(v)$ has to be appoximated by a piecewise constant function $\phi_0^c(v)$, so that the Riemann solution also applies at the inflow boundary.

The final output, which will be used as input of the diffusion part (see Section 2.3.3), is obtained by projecting $\phi(v,\tau)$ to a piecewise constant function on intervals (v_j^e, v_{j-1}^e), $j = 1,\dots,M$. This corresponds to taking averages over (v_j^e, v_{j-1}^e). For example, if $\delta_k \in (v_j^e, v_{j-1}^e)$ and $\delta_{k-1}, \delta_{k+1} \notin (v_j^e, v_{j-1}^e)$, then we have

$$C_l^{1/2}(v) = U^{j_k} \frac{v_{j-1}^e - \delta_k}{v_{j-1}^e - v_j^e} + U^{j_{k+1}} \frac{\delta_k - v_j^e}{v_{j-1}^e - v_j^e}, \quad \text{for } v \in (v_j^e, v_{j-1}^e),$$

and, similarly, in other cases. Thus, if $v_{j-1}^e < \delta_{k-1} < v_{j-2}^e$ and $v_{j+1}^e < \delta_k < v_j^e$, then $C_l^{1/2}(v) = U^{j_k}$ for $v \in (v_j^e, v_{j-1}^e)$. Then, $C_l^{1/2} = PC^{1/2} = PT(\tau)$ is piecewise constant and we can switch to the diffusion.

Recall that in the confined setting $(h_2 > h_1 \geq H)$, we can express $\tilde{G}(v)$ in an analytical form, since h_{eff} is then the constant H, i.e. $\tilde{G}(v) = \tilde{G}(v;p) = \left(\overline{G}(v;p) - \overline{G}(v^{(1)};p)\right) H/K$ and

$$\overline{G}(v,p) := \frac{2pz-2}{(1-p^2)(z^2-2pz+1)} + \frac{2p}{(1-p^2)^{3/2}} \arctan \frac{z-p}{\sqrt{1-p^2}} \tag{2.54}$$

where $z = e^v$, $p = \cos u_{i_0}$ (when we are in the strip $u_{i_0-1/2} < u < u_{i_0+1/2}$). In the unconfined setting, $\tilde{G}(v)$ must be determined numerically. We use a Newton iteration to determine $\delta_k = \tilde{G}^{-1}(y_k - \tau)$ for $y_k - \tau \in (y_j, y_{j-1})$ so that $\delta_k \in (v_j^e, v_{j-1}^e)$: we look for the zero point of $\psi(v) \equiv \tilde{G}(v) - (y_k - \tau)$ starting from v_j where $\psi(v_j) = y_j - y_k + \tau$. Note that $\psi(v)$ and the derivative $\psi'(v) = \frac{h_{\text{eff}}}{K\lambda^2}$ can be easily computed for every v.

If we neglect D_0, α_L, α_T (i.e. we don't consider diffusion), the response of contaminant injection at $t = 0$ is expected exactly at time T_r, the reponse time, with

$$y_0 - T_r = y_N, \quad \text{i.e. } T_r = \int_{v^{(1)}}^{v^{(2)}} \frac{h_{\text{eff}} dv}{K\lambda^2}, \tag{2.55}$$

for $u = \pi$ ($p = -1$) in λ, which corresponds to the line connecting the centers of the wells. This time of response should correspond to the beginning of a withdrawal curve corresponding to the pulse type injection of contaminant.

2.3.3 Solution of the diffusion part

As grid points we use the ordered pairs $\{u_i, v_j\}$, $i = 0, \dots, N$; $j = 0, \dots, M$. The diffusion part of (2.44) is

$$\partial_t \phi = g \left\{ \partial_u (a(u,v) \partial_u C) + \partial_v (b(u,v) \partial_v C) \right\}. \tag{2.56}$$

As mentioned before, the boundary condition is a Dirichlet boundary or a homogeneous Neumann condition at the inflow, and homogeneous Neuman conditions elsewhere. We integrate (2.56) over (t_{k-1}, t_k) and V_{ij}. We assume further that the values C_{ij} and $g_{ij} = g(u_i, v_j)$ are dominant over V_{ij}. Let us denote by $C^E = C^E_{ij} = C_{i+1,j}$, $C^W = C^W_{ij} = C_{i-1,j}$, $C^N = C^N_{ij} = C_{i,j-1}$, $C^S = C^S_{ij} = C_{i,j+1}$ and $a^E = a^E_{ij} = a_{i+1/2,j}$ and similarly a^W, a^N, a^S, b^E, b^W, b^N, b^S. Then, applying the finite volume method, we consider

$$V_{ij} \int_{\Delta t} \partial_t C_{ij} \, dt = g_{ij} \int_{\Delta t} \int_{V_{ij}} \left\{ \partial_u (a(u,v) \partial_u C) + \partial_v (b(u,v) \partial_v C) \right\} du \, dv \, dt.$$

The approximation can be done using integration by parts and approximating $\partial_u C$ on the edge $(u_{i+1/2}, v)$, resp. $(u_{i-1/2}, v)$, for $v \in (v_{j+1/2}, v_{j-1/2})$ by $(C^E - C)/(\Delta u_+)$ resp. $(C - C^W)/(\Delta u_-)$, and similarly $\partial_v C$ on the edges $(u, v_{j+1/2})$, $(u, v_{i-1/2})$. Combined with an implicit time step, we obtain the approximation scheme

$$\begin{aligned}\left[\omega+\left(a^E\frac{\Delta v}{\Delta u_+}+a^W\frac{\Delta v}{\Delta u_-}+b^N\frac{\Delta u}{\Delta v_+}+b^S\frac{\Delta u}{\Delta v_-}\right)\tau\right]C_{i,j}&=\\ \left[\tau\frac{\Delta v}{\Delta u_-}a^W\right]C_{i-1,j}&+\left[\tau\frac{\Delta v}{\Delta u_+}a^E\right]C_{i+1,j}\\ +\left[\tau\frac{\Delta u}{\Delta v_+}b^N\right]C_{i,j-1}&+\left[\tau\frac{\Delta u}{\Delta v_-}b^S\right]C_{i,j+1}+\omega C_{i,j}^{k-1},\end{aligned}\tag{2.57}$$

where $\omega=\omega_{ij}=\frac{|V_{ij}|}{g_{ij}}$. Taking into account the boundary conditions, we have to put $a^W\equiv 0$ for the points $\{u_0,v_j\}$ and $a^E\equiv 0$ for the points $\{u_N,v_j\}$, $j=1,\ldots,N$. Moreover, for $\{u_i,v_M\}$, $i=0,\ldots,N$, we take $b^S\equiv 0$ in (2.57). The inflow boundary condition still needs to be considered. If we have to satisfy a Neumann condition, we just need to put $b^N\equiv 0$ for $\{u_i,v_0\}$, $i=0,\ldots,N$. In the case a Dirichlet condition must be applied, we know the value of the concentration. In this case we join the finite volume around this point with the next finite volume so that V_{i1} is $(u_{i-1/2},u_{i+1/2})\times(v_0,v_{3/2})$, and we set the derivative equal to $(C_{i,0}-C)/(v_0-v_1)$, where $C_{i,0}$ is the prescribed Dirichlet condition at time $i\tau$.

The above scheme corresponds to the matrix system

$$\mathbf{A}\mathbf{C}^k=\mathbf{d}.\tag{2.58}$$

Here, the matrix $\mathbf{A}$ is diagonal dominant and positive definite. The vector $\mathbf{d}$ can be constructed from the previous timestep and the Dirichlet boundary condition (if present), i.e, $d_{i,j}=\omega C_{i,j}^{k-1}+\delta_{j1}b^N\frac{\Delta u}{\Delta v_+}\tau C_{i,0}$, where δ_{lk} is the Kronecker symbol.

For the solution of this matrix equation we use the pre-conditioned conjugate gradient method as implemented in the package Meschach. If $\alpha_T=0$ and $D_0=0$, the diffusion is reduced to only the v-direction ($a(u,v)=0$). Then a simple TDMA (tridiagonal matrix algorithm, see [62]) can be used to solve in each strip the one-dimensional diffusion, which we present in the next section.

2.3.4 Benchmark solution: the case $\alpha_T=0=D_0$

It is important to validate a model. Benchmark solutions can perform part of this task. It is a solution of a simplified case, which can be solved more accurately. By comparing the solution of the general model when the data converge to the data of the benchmark solution, one can estimate the accuracy of the general model.

If we have only longitudinal dispersion, (i.e., $\alpha_T = 0 = D_0$), a different approach than the one outlined in the previous section is possible. In this discretization, numerical diffusion arises due to the projection of the solution after transport to piecewise constants, which is performed to get an initial condition for the diffusion part. We can skip this projection step when only longitudinal dispersion is taken into account, as the problem reduces to solving several one-dimensional problems. We obtain a precise numerical approximation that can be used as a benchmark solution.

We now apply in every u_i strip the transformation (2.53) to the general form (2.44). This gives for each strip $u \in (u_{i-1/2}, u_{i+1/2})$ the 1D convection-diffusion problem

$$\partial_t \tilde{C}_i - \partial_y \tilde{C}_i = \partial_y \left(\frac{2\alpha_L h_{\text{eff}}}{\gamma K \ \lambda \left(\tilde{G}^{-1}(y), u_i \right)} \partial_y \tilde{C}_i \right), \ y \in (y_0^{(i)}, y_M^{(i)}), \tag{2.59}$$

where we have $\tilde{\lambda}(y, u_i) = \lambda \left(\tilde{G}^{-1}(y), u_i \right)$.

We avoid numerical dispersion due to the projection onto step functions corresponding to the fixed grid $\{v_j\}_{j=1}^M$, by doing all computations in the y coordinate frame. For this we use a fixed (coarse) discretization $\{y_j\}_{j=1}^N$ which is uniform on $(y_0^{(i)}, y_M^{(i)})$ with stepsize $\Delta y^{(i)}$. Along with (2.59) we have boundary and initial conditions

$$\tilde{C}(y_0^{(i)}, t) = C_0(t) \text{ or } \partial_y \tilde{C}(y_0, t) = 0, \text{ and } \partial_y \tilde{C}(y_N, t) = 0, \ \tilde{C}(y, 0) = 0.$$

Here, $C_0(t)$ is again a step or a pulse input. To resolve the shocks that are present in the input profile, we add a few grid points to do front tracking. This means that the extra grid points are placed close to the front (fine grid), and will move with the transport velocity -1. We denote the set of both fixed and moving grid points of $(y_0^{(i)}, y_M^{(i)})$ by $\{y_j\}_{j=1}^{N+m}$

The method of approximation for (2.59) is again operator splitting of the transport part and diffusion part. For the simple one-dimensional hyperbolic transport part, we take a time step $\Delta t = q \Delta y^{(i)}$, $q \in \mathbb{N}_0$, which relates the coarse grid stepsize with the time step. This is substantial as this implies that the concentration in the fixed grid points is exactly known, and no interpolation is necessary. We have exactly $C(y_i, t + \Delta t) = C(y_{i-q}, t)$. Note that Δt changes from strip to strip. The moving grid points keep their concentration value, but have to be shifted over a distance $-\Delta t$.

The diffusion part can be handled as in Section 2.3.3, where we set $a(u, v) = 0$. Special care needs to be taken with the moving grid points, because this changes the grid after every time step. By means of the moving grid points we can approximate the front of the wave much more precisely with a relatively small number of fixed grid points and with preservation of the mass balance (locally and globally).

To retrieve a result at a specific time t, (2.59) must be solved in all strips up to a time $t_e^{(i)} \geq t$, (with $t_e^{(i)} = r\Delta y^{(i)}$, $r \in \mathbb{N}_0$), and simple linear interpolation in time is used to obtain the concentration value at t for each strip separately.

2.3.5 General remark on operator splitting

In all our numerical approximations of (2.44), we apply operator splitting. This means that we first consider transport during a time Δt, and then let the system diffuse during the same time. Because in (2.56) and (2.59) the diffusion is place-dependent, an error due to the operator splitting is introduced. This error will be small if the time step of transport is sufficiently small, so that there is only a small change in diffusion coefficients between the initial position and the new position. The diffusion coefficients of the final position can then be taken as a good approximations of the coefficients over the entire transport length. Another approach would be to take as diffusion coefficients the average values of the initial diffusion and final diffusion coefficients.

2.4 Solution of advection dominated diffusion with equilibrium adsorption

Now we will start adding nonlinearity to the problem. The solution obtained in Section 2.3 is used to evaluate the accuracy of the discretization: when the nonlineary disappears, we should recover the linear solution.

For contaminant transport, the main nonlinearity comes from adsorption processes. When the contaminant flows through the subsurface a portion of it sticks to the surface of the grains, which therefore behave as sinks. At the same time, the contaminant can detach of the grain, which gives rice to a source of contaminant.

Contaminant transport with adsorption is a very dynamical and difficult research area. Precise mathematical models are available and a significant effort has been done to develop efficient numerical methods for the solution. However, the solution of strongly nonlinear convection-diffusion problems with dominant convection and nonlinear adsorption is still an open problem. The main reason is that the solution can be localized with the sharp fronts and is very dynamical. This is a very difficult task for precise numerical approximation. Various types of regularizations (e.g., up winding) must be applied to stabilize numerical oscilations and instabilities. This, in turn, leads to numerical dispersion which shadows the influence of, and sensitivity on, the model data. In some special cases desirable results have been obtained. Also in our model setting, the dual-well, a contribution towards a precise numerical solution is obtained by us.

Generally, adsorption gives rises to a source/sink term I, see (2.37). The form of this term is deduced in [7, 69], and explained later in Section B.7. One has

$$h_{\text{eff}}\partial_t C = \nabla \cdot (D h_{\text{eff}} \nabla C) - \nabla(h_{\text{eff}} \boldsymbol{v} C) - \frac{h_{\text{eff}}}{\theta_0} \varrho \, \partial_t S, \tag{2.60}$$

which for equilibrium adsorption is written as

$$h_{\text{eff}}\partial_t F(C) = \nabla \cdot (D h_{\text{eff}} \nabla C) - \nabla(h_{\text{eff}} \boldsymbol{v} C), \tag{2.61}$$

where $F(C) = C + \Psi(C)$, with $\Psi(C)$ the adsorption isotherm (hiding the ϱ/θ_0 terms in its definition). This is the same equation, apart from the time derivative, as (2.37). The transformations done previously on the space domain can be repeated to obtain an equation over a 2-D rectangle in the case of the dual-well. Then (2.44) is replaced by

$$\partial_t F(C) = g\{\partial_u(a\partial_u C) + \partial_v(b\partial_v C)\} + G\partial_v C, \quad \text{in } \Omega, \tag{2.62}$$

with the same definition for g, a, b, G and Ω, and also with the same initial condition (2.48) and boundary conditions (2.46)-(2.47) and (2.45) or (2.50). For brevity we will only work with BC (2.50), as this is the most physical one.

The results of this Section appeared in [39].

2.4.1 Numerical approximation of (2.61)

To solve the convection diffusion problem above we use time stepping and operator splitting in which, for each small time interval, the problem is splitted into 2 parts: *the transport problem* and *the diffusion problem.* More in detail, let $\tau = T/L$, $(L \in \mathbf{N})$, be a time step and let $C_n \approx C((u, v), t_n)$ for $n = 1, \dots, L$.

If C^{n-1} is known, then the relation

$$C^n = D^n(\tau)T^n(\tau)C^{n-1}, \quad \tau = t_n - t_{n-1},$$

determines C^n. The transport $T^n(\tau)$ corresponds to the solution ϕ of the transport equation

$$\partial_t F(\phi) - G(u,v)\partial_v \phi = 0, \tag{2.63}$$

with the inflow condition

$$\phi((u, v^{(2)}), t) = C_0(t)$$

and the initial condition

$$\phi((u,v), t_{n-1}) = C^{n-1}.$$

Then, we put

$$C^{n,1/2} := T^n(\tau)C_{n-1} \equiv \phi((u,v), t_n).$$

The diffusion $D^n(\tau)$ is obtained by solving the diffusion equation

$$\partial_t F(\phi) = g\left\{\partial_u(a\partial_u \phi) + \partial_v(b\partial_v \phi)\right\}, \tag{2.64}$$

along with the initial condition

$$\phi((u,v), t_{n-1}) = C^{n,1/2},$$

and the boundary condition

$$\partial_\nu \phi = 0 \quad \text{on } \partial\Omega.$$

The arguments for the specific splitting of the boundary condition are the same as in the linear setting. Next, we set

$$C^n = D^n(\tau)C^{n,1/2} = D^n(\tau)T^n(\tau)C^{n-1} \equiv \phi((u,v), t_n).$$

We will prove convergence for this approximation in Section 2.5, based on convergence results for operator splitting approximation, [16] and [31].

The space discretization for the nonlinear problem is the same as before, see Section 2.3.1.

2.4.2 Solution of the nonlinear transport problem

We consider (2.63) in the strip $(u_{i_0-1/2}, u_{i_0+1/2}) \times (v^{(1)}, v^{(2)})$ with shocks on the edges $v_{j+1/2}$ of the finite volumes. The resulting 1D-problem can be solved by a semi-analytical Riemann method without a time step limitation for the case of Langmuir or Freundlich type isotherms, see [37, 45]. The solution of the transport problem (2.63) will again be based on a piecewise constant initial profile $\phi_0(v)$, i.e., the solution of the multiple Riemann problem, see Section 2.1.2.

In the general case of isotherms we transform (2.63) by using the new variable $y = y(v)$ where

$$y = G_i(v) \equiv G(u_i, v) = \int_{v^{(1)}}^{v} \frac{h_{\text{eff}} dv}{K\lambda^2}, \tag{2.65}$$

$$\lambda^2 = (\cosh(v) - p_i)^2, \quad p_i = \cos u_i, \quad \overline{\varphi}(u_i, y) = \varphi(u_i, v) \quad i = 1, ..., N.$$

We obtain (index i is omitted)

$$\partial_t F(\overline{\phi}) - \partial_y \overline{\phi} = 0, \quad \overline{\phi}(y, 0) = \phi_0(v). \tag{2.66}$$

If the initial profile $\phi_0(v)$ in (2.66) is piecewise constant, then the solution consists of the values of $\phi_0(v)$ and rarefactions in the intervals given by the positions of the original shocks (i.e. original intervals) after time evolution τ. If the original shock at $y = y_j$ was acceptable, then its position will be given by the Rankin-Hugoniot speed movement

$$\dot{s}(t) = \frac{\bar{\phi}(y_j^+, 0) - \bar{\phi}(y_j^-, 0)}{F[\bar{\phi}(y_j^+, 0)] - F[\bar{\phi}(y_j^-, 0)]},$$

where y^+ denotes the upstream limit, and y^- the downstream limit. If the original shock at $y = y_j$ was not acceptable, then it develops into the rarefaction along the y-interval

$$(y_j - \frac{1}{F'[\bar{\phi}(y_j^-)]}\tau, y_j - \frac{1}{F'[\bar{\phi}(y_j^+)]}\tau).$$

If F is convex and $\phi(y_j^-, 0) \leq \phi(y_j^+, 0)$, then the shock is acceptable. If $\phi(y_j^-, 0) \geq \phi(y_j^+, 0)$, then the shock is unacceptable. If F is concave (e.g. for Langmuir adsorption or Freundlich adsorption with $p < 1$), then the role of acceptable and unacceptable shocks is interchanged.

In the case of Freundlich isotherm, i.e. $F(s) = s + K_0 s^p$, the general form of rarefaction is (see [37])

$$\bar{\phi}(y,t) = \left(\frac{t + y - y_j}{pK_0(y_j - y)}\right)^{\frac{1}{p-1}}, \quad \text{for } 0 < y_j - y < t.$$

In the case of Langmuir sorption isotherm, i.e. $F(s) = s + K_1 \frac{s}{1+K_2 s}$, the rarefaction is of the form (see [37])

$$\bar{\phi}(y,t) = \frac{1}{K_2}\left(-1 + \sqrt{K_1}\frac{2(y_j - y)}{\sqrt{t^2 - (2(y_j - y) - t)^2}}\right), \quad \text{for } 0 < y_j - y < t.$$

To construct the global entropy solution of the multiple Riemann problem, we find the position of grid points (of the original shocks) after time length τ and put together the constant values and the local solutions of the rarefaction waves. This holds for small time step τ during which no collision arise between the neighbouring shocks or rarefactions. The collisions of neighbouring shocks can be treated as follows. The constant value between shocks disappears and we are left with only one shock with the jump equal to the summation of the original jumps. The collision of the rarefaction which meets the shocks can also be described in an analytical form (see [37]), but we shall limit our time step up to the first collision of rarefaction and shock, and project onto piecewise constants. Then, we continue the transport. Since we consider only a very special initial profile, it is simple to calculate this time limitation. From the solution of (2.66) we obtain the desired solution of (2.63) for each strip $i = 1, ..., N$ using the backward transformation. In case that $h_2 > h_1 \geq H$, (confined aquifer), we can express (2.65) as

$$G_i(v) = [\bar{G}(u_i, v) - \bar{G}(u_i, v^{(1)})]\frac{\gamma^2\theta_0 H}{4A},$$

where

$$\bar{G}(u_i, v) = \frac{2pz - 2}{(1-p^2)(z^2 - 2pz + 1)} + \frac{2p}{(1-p^2)^{3/2}} \arctan\frac{z - p}{\sqrt{1-p^2}}$$

with

$$z = e^v, p = \cos u_i.$$

2.4.3 Solution of the projection problem

After transport we obtain the profile $\bar{\phi}(y)$ which consists of piecewise constant parts and rarefactions waves. Before starting the diffusion, this profile must be projected onto a piecewise constant profile $\bar{\phi}(y_j)$ that will be used for diffusion.

From (2.66) is follows that $F(\bar{\phi})$ is the conserved quantity, which is also clear from a mass balance consideration. Thus, we need to find the value $\bar{\phi}(y_j)$ for which we have

$$F(\bar{\phi}(y_j))\Delta y = F_j \Delta y = \int_{y_{j-1/2}}^{y_{j+1/2}} F(\bar{\phi}(y))dy. \tag{2.67}$$

The right hand side of (2.67) can be readily obtained in the case of a piecewise constant profile. When rarefaction waves are encountered, exact projection has been worked out in [37], and is given by

$$\int_{\alpha}^{\beta} F(\bar{\phi}(y)) = J(\beta) - J(\alpha), \quad \text{with } J(\beta) = \beta\left[F(\bar{\phi}(\beta)) - \bar{\phi}(\beta)F'(\bar{\phi}(\beta))\right].$$

Having determined F_j, we can calculate $\bar{\phi}(y_j)$ by a hybrid Newton-Raphson/bisection algorithm, see [70], in the case of Freundlich adsorption. For Langmuir adsorption the inverse is a known function.

2.4.4 Solution of the nonlinear diffusion problem

The same approximation scheme that has led to (2.57), now leads to

$$\begin{aligned}\omega F(C_{i,j}) + \left(a^E \frac{\Delta v}{\Delta u_+} + a^W \frac{\Delta v}{\Delta u_-} + b^N \frac{\Delta u}{\Delta v_+} + b^S \frac{\Delta u}{\Delta v_-}\right)\tau C_{i,j} = \\ \left[\tau \frac{\Delta v}{\Delta u_-} a^W\right] C_{i-1,j} + \left[\tau \frac{\Delta v}{\Delta u_+} a^E\right] C_{i+1,j} \\ + \left[\tau \frac{\Delta u}{\Delta v_+} b^N\right] C_{i,j+1} + \left[\tau \frac{\Delta u}{\Delta v_-} b^S\right] C_{i,j-1} + \omega F(C_{i,j}^{n-1}),\end{aligned} \tag{2.68}$$

where $\omega = \omega_{ij} = \frac{|V_{ij}|}{g_{ij}}$, and $F(C_{i,j}^{n-1})$ is the value F_{ij} obtained in the projection step.

Taking into account the boundary conditions we have to put $a^W \equiv 0$ for the points $\{u_1, v_j\}$ and $a^E \equiv 0$ for the points $\{u_N, v_j\}$, $j = 1, \ldots, M$. Moreover, for $\{u_i, v_1\}$, $i = 1, \ldots, N$, we take $b^S \equiv 0$ in (2.57), and $b^N \equiv 0$ for the points $\{u_i, v_M\}$.

We suggest two possible solution methods for the nonlinear system of algebraic equations (2.68). First, we can use Newton type iterations, starting with $C \equiv C^{n-1}$. This implies solving a matrix equation in every iteration step. If $\alpha_T = 0$ and $D_0 = 0$, the diffusion is reduced to only the v-direction $(a(u,v) = 0)$ and in each step a simple TDMA (tridiagonal matrix algorithm) can be used to solve in each strip the one-dimensional diffusion problem. Note that only if convergence of the Newton method is reached, mass balance will be kept. Thus, timestep and gridsize should be carefully chosen to obtain this convergence after a small number of iteration steps.

Secondly, we also implement a relaxation method as in [30] and [35] as follows

$$\omega\lambda_{i,j}^{(l-1)}(C_{i,j}^{(l)} - C_{i,j}^{n-1}) + \left(a^E\frac{\Delta v}{\Delta u_+} + a^W\frac{\Delta v}{\Delta u_-} + b^N\frac{\Delta u}{\Delta v_+} + b^S\frac{\Delta u}{\Delta v_-}\right)\tau C_{i,j} =$$
$$\left[\tau\frac{\Delta v}{\Delta u_-}a^W\right]C_{i-1,j} + \left[\tau\frac{\Delta v}{\Delta u_+}a^E\right]C_{i+1,j}$$
$$+\left[\tau\frac{\Delta u}{\Delta v_+}b^N\right]C_{i,j+1} + \left[\tau\frac{\Delta u}{\Delta v_-}b^S\right]C_{i,j-1}, \quad (2.69)$$

where l is an iteration parameter and

$$\lambda_{i,j}^{(l)} := \frac{F(C_{i,j}^{(l)}) - F(C_{i,j}^{n-1})}{C_{i,j}^{(l)} - C_{i,j}^{n-1}}, \quad \lambda_{i,j}^{(0)} := F'(C_{i,j}^{n-1})$$

is a relaxation function.

We stop the iterations and define $C_{i,j} := C_{i,j}^{(l_0)}$ as soon as

$$\text{(a) } |\lambda_{i,j}^{(l_0)} - \lambda_{i,j}^{(l_0-1)}| < \tau \quad \text{and} \quad \text{(b) } \sum_{i,j}\left|C_{i,j}^{(l_0)} - C_{i,j}^{(l_0-1)}\right| < \epsilon, \quad (2.70)$$

ϵ being a small tolerance. This is the *stopping criterium.*

Remark 2.4.1. *The relaxation method is more robust than the Newton method, allowing to solve also the higly nonlinear cases with large time steps. Therefore, it is the prefered choice. However, close to the exact solution the Newton method converges much faster. Furthermore, the stopping criterium (2.70) is not optimal: around the moving interface, where the concentration is very small, λ can oscilate, preventing (2.70-b) to be fulfilled. To overcome this difficulty we suggest several optimizations. We will illustrate the consequence by comparing the typical number of iterations in the case of Freundlich adsorption with* $p = 0.25$.

With the standard stopping criterium there are typically 220 relaxation iterations per timestep at the beginning of the injection.

- *A first optimization neglects what happens to the small concentrations, and controls the mass balance instead. We stop the iterations and define* $C_{i,j} := C_{i,j}^{(l_0)}$ *as soon as*

$$\text{(a) for all} \quad C^{(l_0)} > \epsilon \text{ or } C^{(l_0-1)} > \epsilon : \qquad |\lambda_{i,j}^{(l_0)} - \lambda_{i,j}^{(l_0-1)}| < \tau, \quad (2.71)$$

$$\text{(b)} \sum_{i,j} \Delta u_i \Delta v_j \lambda_{i,j}^{(l_0)} (C_{i,j}^{(l_0)} - C_{i,j}^{n-1}) \approx 0. \quad (2.72)$$

  Condition (2.72) follows from the conservation of mass argument. During the diffusion step we have homogeneous Neumann BCs. Consequently,

$$\int_\Omega \partial_t F(C) = 0 \approx \sum_{i,j} \int_{\Omega_{i,j}} \left(F(C_{i,j}^{(l)}) - F(C_{i,j}^{n-1}) \right) / \tau,$$

  from which (2.72) follows. This procedure guarantees that the iterations are continued in those cases where the solutions varies only a little when passing from one iteration to the next, but the correct solution is not yet obtained. We now have typically 110 relaxation iterations per timestep, and the problem of oscilating λ *is completely avoided.*

- *One of the reasons of slow convergence is that during the transport process the interface was sharpened. During diffusion the interface then smooths out again. However, the initial value of the relaxation parameter* λ *to the right is very large up to infinity (as the starting concentration is nearly zero or zero), and needs to convergence to a value* $F(C)/C$. *The slow convergence can be improved by taking for the values to the right of the interface the same values for* $\lambda^{(0)}$ *as those found in the previous run of the diffusion step. After a run, we keep in memory the values to the right of the interface which are smaller than* D *(*$D >> 1$*) and, at the beginning of the next diffusion step, we take to the right of the interface these values as the values for* $\lambda^{(0)}$ *in stead of* $F'(C_{i,j}^{n-1})$*.This typically reduces the number of iterations to 65.*

- *Finally, we use the fast convergence of the Newton method close to the solution. After every iteration step, we check wether the mass is conserved up to a precision of* 1%*, i.e. we relax condition (2.72). If this is the case,* *we stop the relaxation iterations and continue with Newton iterations using the last solution found during the relaxation method. We do a minimim of 5 relaxation iterations before starting the Newton iterations. This typically reduces the number of iterations to 6 relaxation iterations and 15 Newton iterations.*

We conclude that we have developed a numerical approximation scheme for the nonlinear dual-well problem, based upon operator splitting, the Riemann problem, and a nonlinear finite volume scheme solved with a relaxation scheme.

2.5 Convergence of the numerical method in 2D

We now prove convergence of the operator splitting method applied to a boundary value problem of the form

$$\begin{gathered}
\tfrac{1}{g(x,y)}\partial_t F(v) - h(x,y)\partial_y v - (\partial_x(a(x,y)\partial_x v + \partial_y(b(x,y)\partial_y v) = 0 \\
v(x,y,0) = v_0(x,y), \\
b(x,y)\partial_y v + h(x,y)v = h(x,y)v^0(x,y,t) \quad \text{on } \Gamma_1 \text{ (inflow)}, \\
\partial_\nu v = 0 \quad \text{on } \Gamma_2, \\
\partial_\nu v = 0 \quad \text{on } \Gamma_3 \text{ (outflow)},
\end{gathered} \tag{2.73}$$

where $(x, y, t) \in \Omega \times I = [x^{(1)}, x^{(2)}] \times [y^{(1)}, y^{(2)}] \times [0, T]$, Γ_1 is the inflow boundary ($hn_y < 0$) given by the line $y = y^{(2)}$ ($h(x, t) \geq 0$), Γ_3 is the outflow boundary ($hn_y > 0$) given by the line $y = y^{(1)}$, and Γ_2 is such that $\partial\Omega = \overline{\Gamma}_1 \cup \overline{\Gamma}_2 \cup \overline{\Gamma}_3$. Here, we are interested in functions F of the type $F(v) = v + c\psi_e(v)$, c a constant and ψ_e a so-called equilibrium sorption isotherm. As always, ∂_ν denotes the outward normal derivative.

We will need the following assumptions: F^{-1} is Lipschitz continuous, monotone increasing with $F(0) = 0$ and $F(s) < C_L$ if $s < L$. Furthermore a and b are smooth and positive and g and h are smooth, positive and bounded.

Remark 2.5.1.

- *We have taken F^{-1} Lipschitz continuous as then the case of Freundlich adsorption with power less than 1 is included (without need of special treatment). This is the interesting case.*
- *The dual well fully fits into the scheme (2.73), except for one minor point: $g(u, v)$ as defined in (2.44) is zero in the point $(0, 0)$. This point can however be disregarded as it is the point at infinity where the contaminant*

concentration is zero. We may regularize g by $g_\epsilon > \epsilon$, a positive function, see further Remark 2.5.11.

The proof is based on the ideas presented in [16, 27, 25, 31, 46].

Let us first define the weak solution to problem (2.73). We obtain the variational weak formulation by multiplying (2.73) by a test function $\phi(x, y, t) \in C^\infty(\Omega \times I)$, by integrating over $\Omega \times I$ and performing once integration by parts. This gives

$$\begin{aligned}
&\int_{\Omega\times I} (\partial_t \phi)\frac{F(v)}{g} - \int_\Omega \phi\frac{F(v)}{g}\Big|_{t=T} + \int_\Omega \phi\frac{F(v)}{g}\Big|_{t=0} \\
&\quad - \int_{\Omega\times I} (\partial_y h\phi)v + \int_0^T \int_{x^{(1)}}^{x^{(2)}} \left[(h\phi v)\big|_{y=y^{(2)}} - (h\phi v)\big|_{y=y^{(1)}}\right] dx\, dt \\
&\qquad - \int_{\Omega\times I} [(\partial_x\phi)(a\partial_x(v)) + (\partial_y\phi)(b\partial_y(v))] \\
&\qquad + \int_0^T \int_{y^{(1)}}^{y^{(2)}} \left[\phi a\partial_x(v)\big|_{x=x^{(2)}} - \phi a\partial_x(v)\big|_{x=x^{(1)}}\right] \\
&\qquad + \int_0^T \int_{x^{(1)}}^{x^{(2)}} \left[\phi b\partial_y(v)\big|_{y=y^{(2)}} - \phi b\partial_y(v)\big|_{y=y^{(1)}}\right] = 0. \qquad (2.74)
\end{aligned}$$

Next, we impose the boundary conditions of (2.73), with special care for the inflow boundary. We consider test functions ϕ with $\phi = 0$ on the outflow $(y = y^{(1)})$ and with $\phi(u, v, T) = 0$. We are led to the following definition.

Definition 2.5.1. *A weak solution u of (2.73) satisfies*

$$\begin{aligned}
&\int_{\Omega\times I} (\partial_t \phi)\frac{F(u)}{g} + \int_\Omega \frac{F(v_0(x,y))}{g}\phi(x, y, 0) \\
&\qquad - \int_{\Omega\times I} [(\partial_x\phi)(a\partial_x(u)) + (\partial_y\phi)(b\partial_y(u))] \\
&\qquad - \int_{\Omega\times I} (\partial_y h\phi)u + \int_0^T \int_{x^{(1)}}^{x^{(2)}} hv^0(t)\phi\, dx\, dt\Big|_{y=y^{(2)}} = 0, \qquad (2.75)
\end{aligned}$$

for all $\phi(x, y, t) \in C^\infty(\Omega \times I)$, with $\phi = 0$ at $t = T$, a.e. in Ω, and with $\phi = 0$ on the outflow boundary Γ_3 (i.e. $\phi(x, y^{(1)}, t) = 0$) for a.e. $t > 0$..

We can also introduce a *very* weak solution as follows

Definition 2.5.2. *A very weak solution u of (2.73) satisfies*

$$\int_{\Omega\times I}(\partial_t\phi)\frac{F(u)}{g}+\int_{\Omega}\frac{F(v_0(x,y))}{g}\phi(x,y,0)$$
$$+\int_{\Omega\times I}u\left[\partial_x(a\partial_x(\phi))+\partial_y(b\partial_y(\phi))\right]$$
$$-\int_{\Omega\times I}(\partial_y h\phi)u+\int_0^T\int_{x^{(1)}}^{x^{(2)}}hv^0(t)\phi\,dx\,dt|_{y=y^{(2)}}=0, \quad (2.76)$$

for all $\phi(x,y,t)\in C^\infty(\Omega\times I)$ which have compact support near Γ_3, and which moreover fullfill $\phi=0$ at $t=T$, a.e. in Ω, and also obey $\partial_y\phi=0$ on Γ_1 and $\partial_x\phi=0$ on Γ_2 for a.e. $t>0$.

Remark 2.5.2. *Note that since $\partial_y\phi=0$ on Γ_1 for a very weak solution, the term $-\int_0^T\int_{x^{(1)}}^{x^{(2)}}bv\partial_y\phi|_{y=y^{(2)}}$ does no longer arise, which is desirable. The requirement ϕ compact in the entire domain in stead of only near Γ_3 would also set this term equal zero, but then the important inflow boundary condition on Γ_1, i.e. the last term of (2.76), would not be present in the very weak solution. The same goes for Γ_2.*

We first consider the *semi-discrete method* in which operator splitting is used but the separate subproblems are solved exactly, and next the *fully discrete method* in which the subproblems are solved numerically.

2.5.1 Semi-discrete method

As mentioned before, there are several ways to split the equation, see e.g., [16] and [31]. Here we split with respect to the physical properties. Therefore, (2.73) splits into a hyperbolic part

$$\partial_t F(v)-G(x,y)\partial_y v=0, \quad (2.77)$$

where $G=gh$, and with inflow condition

$$v(x,y^{(2)},t)=v^0(x,y^{(2)},t) \quad (2.78)$$

and initial condition

$$v(x,y,0)=v_0(x,y),$$

and a parabolic part

$$\partial_t F(w) = g(x,y)\left\{\partial_x(a(x,y)\partial_x w) + \partial_y(b(x,y)\partial_y w)\right\}, \tag{2.79}$$

with initial condition

$$w(x,y,0) = w_0(x,y)$$

and boundary condition

$$\partial_\nu w = 0 \quad \text{on } \partial\Omega.$$

In what follows we choose a time step Δt and an integer n such that $n\Delta t = T$. We denote further $t_n = n\Delta t$.

The corresponding (semi-discrete) splitting method then reads

$$C_{\Delta t}(t) = \left[\mathcal{D}_{\Delta t} \circ \mathcal{T}_{\Delta t}\right]^n C_0, \quad \text{for } t \in (t_{n-1}, t_n], \quad n = 1, \ldots, N,$$

where $\mathcal{T}_t$ and $\mathcal{D}_t$ denote the solution operators of (2.77) and (2.79), respectively. Note that due to the specific form of (2.77), where the characteristics are along one of the axes, we have that this solution will be identical to the solution in the case where the hyperbolic part is further split into one-dimensional equations:

$$C_{\Delta t}(t) = \left[\mathcal{D}_{\Delta t} \circ I \circ \mathcal{T}^2_{\Delta t}\right]^n C_0, \quad \text{for } t \in (t_{n-1}, t_n], \quad n = 1, \ldots, N, \tag{2.80}$$

where $\mathcal{T}_t^2$ is the exact solution operator associated with variable y, and the hyperbolic operator for the x variable is the identity, or $\mathcal{T}_t^2 = I$.

2.5.2 Aim

Our goal is to prove convergence to the solution of (2.73) of the solution obtained by the used operator splitting. Uniqueness of this solution follows from the original porous media equation (2.37), which is in standard form, and is non-degenerate.

A first possible approach is to put (2.37) into the form used in [25], and use the operator splitting method as developed therein. This method would clearly converge. In this section, we want to show that analogue techniques can be used on the transformed equation (2.73), obtaining a proof of convergence for our operator splitting method which starts from these transformed equations.

Let us first highlight the differences between (2.73) and [25]. First of all, it is needed to set $F(v) = u$, for which not necessarily a known inverse function is known. Furthermore, the term h_{eff}, when discarded in the unsaturated case, leads to a non-divergence free velocity field when handled as in [25], which is not considered there. Finally, we consider the full IVBP problem. Therefore other complications arise, such as the correct splitting of the boundary conditions.

As there is no degeneracy of the diffusion term, the solution cannot contain shocks: thus entropy solutions need not be considered for (2.73). However, as we use operator splitting, the entropy condition plays an important role in the hyperbolic part of the split equation. The solution of (2.73) satisfies entropy conditions, even if they are not used to select the physical solution.

2.5.3 Fully discrete method

Practically, the exact solution operators need to be replaced by suitable numerical methods. For the convergence proof, we use a relaxation type FVM to approximate $\mathcal{D}_t$, and a large step front tracking method to approximate $\mathcal{T}_t$. In this front tracking method we will work in a grid where the front tracking is done strip-wise in every x-strip.

Let $\mathcal{D}_{\Delta xy}(t)$ stand for the FV solution operator associated with (2.79) at time t and let $\mathcal{T}_{\delta,\Delta xy}(t)$ stand for the front tracking solution operator associated with (2.77). As mentioned before, we have $\mathcal{T}_{\delta,\Delta xy}(t) = I \circ \mathcal{T}_{\delta,\Delta x}(t)$.

After solving the hyperbolic step, we must project the solution onto a Cartesian grid. We consider a non-equidistant grid x_i, y_j, with $i = 1, \ldots, N_1$ and $j = 1, \ldots, N_2$, and set $\Delta x_{ij} = \Delta x_i \Delta y_j$ where $\Delta x_i = x_{i+1} - x_i$, and analogously for Δy_j. The projection operator is constructed in such a way as to maintain mass balance, $(\int F(v) \mathrm{d}\Omega)$. It is given by

$$\pi v(x,y) = F^{-1}\left(\frac{1}{|\Omega_{ij}|}\int_{\Omega_{ij}} F(v(x,y))\, d\Omega\right) = F^{-1}\left(\tilde{\pi} F(v)\right), \quad \text{for } (x,y) \in \Omega_{ij}, \tag{2.81}$$

where $\Omega_{ij} = [x_i, x_{i+1}) \times [y_j, y_{j+1})$, with $i = 1, \ldots, N_1 - 1$ and $i = j, \ldots, N_2 - 1$.

If we assume that over the entire timestep the front tracking algorithm provides a solution, our fully discrete splitting method reads

$$C_\eta(t) = \left[\mathcal{D}_{\Delta xy,\Delta t} \circ \pi \circ \mathcal{T}_{\delta,\Delta y,\Delta t}\right]^n C_0, \quad \text{for } t \in (t_{n-1}, t_n], \quad n = 1, \ldots, N, \tag{2.82}$$

where we used $\mathcal{D}_{\Delta xy,\Delta t} = \mathcal{D}_{\Delta xy}(\Delta t)$, the same for $\mathcal{T}_{\delta,\Delta y,\Delta t}$, and where $\eta = (\delta, \Delta xy, \Delta t)$ represents the discretization parameters.

Note that the front tracking algorithm that we use is based on exact solutions. This is in contrast with more general algorithms, based on piecewise linear interpollations of the flux and velocity functions. These methods require a stability result which allows to prove convergence of the solution of the interpolants to the real solution. This will not be needed when it is possible to work with the exact flux and velocity function. The disadvantage however is that the exact solution is no longer known when a shock reaches the beginning of a rarefaction wave. This can be overcome by performing an extra projection during the hyperbolic step. Our fully discrete splitting method then reads

$$C_\eta(t) = \left[\mathcal{D}_{\Delta xy,\Delta t} \circ [\pi \circ \mathcal{T}_{\delta,\Delta y,\Delta t_k}]^l\right]^n C_0, \quad \text{where } \sum_{k=1}^{l} \Delta t_k = \Delta t, \tag{2.83}$$

and $t \in (t_{n-1}, t_n]$, $n = 1, \ldots, N$.

It is necessary to show that $l < \infty$ when $\Delta t, \Delta x \to 0$. We have the following Lemma.

Lemma 2.5.1. *For the hyperbolic problem (2.66) solved by a Riemann solver (2.83), with $F(\phi) = u$, $\phi = f(u)$, $0 \leq f' < L$, the number of projections l needed during transport in (2.83) is for a given timestep Δt, $\Delta t/\Delta x = C$, bounded above by $l_{\max}$, with $l_{\max}$ independent of the timestep.*

Proof. Set t=0. The initial condition is a piecewise constant function over a grid $\{x_i\}$. First, let us suppose that we project to piecewise constants everytime a front passes a gridpoint. With front we mean the beginning of a rarefaction wave, the end of a rarefaction wave, or a shock. As the flux function $f = F^{-1}$ satisfies $f' < L$, the speed of a front is less than L. The minimum time after which a projection can happen is $\Delta x/L$. and the maximum number of projections is therefore $l_{\max} = \Delta t/(\Delta x/L) = CL$, independent of the timestep.

We must still consider the fact that projection might happen before a front passes a gridpoint. We only need to consider the case where a shock (as in Fig. 2.1, u_l to u_r, speed $s = (f(u_l) - f(u_r))/(u_l - u_r)$) meets the beginning of a rarefaction wave (u_l, speed $f'(u_l)$) as only then projection is necessary (the case where the end of a rarefaction wave meets a shock can be solved exactly). As we started with a piecewise continuous initial condition the shock must be at $t = 0$ at a gridpoint x_i, and the beginning of the rarefaction wave at x_{i+1}. Thus, in order for the projection to happen before a front passes a gridpoint, we must have $s > 0 > f'(u_l)$, which is not possible in our setup as $f' \geq 0$. □

Remark 2.5.3. *Lemma 2.5.1 plays a role for the exact Riemann solver which is similar to the role of the Lemma showing that there are a finite number of shock collisions for fixed δ for the front tracking method, where the flux f is approximated by a piecewise constant profile f^δ, see [27, 51]. In our setup shock* *collisions for a given initial condition will be finite for the same reasons as in [27, 51]. However, every time a projection has to be done, the method must be restarted. Therefore Lemma 2.5.1 is important.*

Remark 2.5.4. *The need for projection might seriously endanger the efficiency of our solution method. In the expermints this will not be a problem as we consider single injection pulses. In real experiments, recirculation of contaminant might be applied, and the problem of shocks meeting rarefaction waves will arise. In a worst case scenario, it might be best to use the front tracking method instead of an exact Riemann solver to avoid spurious projections. Other possible methods are projection on a larger resolution than the grid size, or only projection around the problematic collision and not over the entire domain.*

While using a numerical method, we obtain an approximate solution C^k in every point t_k of the time discretization. In order to obtain convergence results, we need functions that are defined on the whole interval $[0, T]$. This has to be done carefull, as in one timestep Δt transport and diffusion happen simultaneously, whereas in our discretization this has been split in two parts. Therefore, we need to compress the two separate steps into one timestep. To this end, all the transport is considered in the first half of the timestep, and the diffusion in the other half, with projection at $\frac{\Delta t}{2}$. Let us therefore define the following sequences:

$$v_\nu(x,y,t) = \begin{cases} \mathcal{T}_{\delta,\Delta y}(2(t-t_k))v^k(x,y) & t \in [t_k, t_{k+1/2}) \\ \mathcal{D}_{\Delta xy,\Delta t}(2(t-t_{k+1/2}))v^{k+1/2}(x,y) & t \in [t_{k+1/2}, t_{k+1}) \end{cases}$$

where v^k is the solution obtained at time t_k. The parameter ν corresponds to time and space discretization. The definition of the sequences formally corresponds to the operator splitting procedure.

The convergence proof consists of several basic steps:

- We prove that v_ν is uniformly bounded
- We show that v_ν has bounded total variation in the space variable
- We prove that v_ν is L_1-Hölder continuous in time with coefficient $\frac{1}{2}$
- Applying Riesz-Frechet-Kolmogorov's compactness criterion we prove the existence of convergent subsequences of v_ν converging for $\nu \to 0$ in L_1-sense to some function $v(x,y,t)$
- We show that the limit function $v(x,y,t)$ satisfies the variational formulation (2.75)

For the fully discrete method (2.83) the result is summarized in the following main theorem

Theorem 2.5.1 (Fully-discrete convergence). *Let the retardation function $F(v)$ be such that it is nondecreasing and that F^{-1} is Lipschitz continuous. In addition, let the functions $g(x,y)$, $h(x,y)$, $a(x,y)$ and $b(x,y)$ be smooth and let $g(x,y)$, $a(x,y)$, $b(x,y)$ be positive. If $v_0(x,y)$ and $v^0(x,y,.)$ are nonnegative, bounded and of bounded total variation, then the numerical approximation $v^n(x,y)$, obtained by the operator splitting scheme (2.83), converges to the very weak solution of the convection-diffusion-adsorption problem (2.73) for $n \to \infty$.*

For the semi-discrete method, a similar result can be obtained. In what follows we concentrate on the fully discrete method.

2.5.4 The hyperbolic step and the projection

The transport step and projection step are obviously uniformly bounded and have bounded total variation. However, these obvious results are obtained in different coordinates/variables than the ones used in the subsequent diffusion step. Therefore, we will write down these parts in detail. A further complication is the dimensional splitting: total variation in one space dimension might be bounded, but this has to be connected to the result obtained in the other space dimension.

1D Hyperbolic step

The main construction in the numerical algorithm is the transformation of the one dimensional hyperbolic equation into

$$\partial_t V - \partial_{\widetilde{y}} f(V) = 0, \tag{2.84}$$

which is solved by front tracking with IC $V(\widetilde{y},0) = V_0(\widetilde{y})$ and the BC $V(\widetilde{y}^{(2)},t) = V^0(t)$ at the inflow $\widetilde{y} = \widetilde{y}^{(2)}$. Here, the variable $\widetilde{y}$ is obtained with a 1-1 transformation from y, and we have $F(v) = V$. As F is strictly monotone, also $f = F^{-1}$ is strictly monotone. After the front tracking, projection to a Cartesian grid in y is done. This grid is transformed 1-1 to a grid $\{\widetilde{y}_j\}$. This gives us values V_j which can be transformed back to the value v_{ij} of the original variable v, where we use $V(j)|_{\widetilde{y}=\widetilde{y}^j} = F(v_{ij})$. To this end a numerical procedure with a Newton type of iteration can be used.

In general, the initial data v_0, (or V_0), is approximated by a step function $v_{0,\Delta x} = \pi v_0$, (or $V_{0,\Delta x}$), before the hyperbolic step is executed. Note that the result of our FVM will be a step function. Therefore, the stepwise approximation must only be performed at $t = 0$. Note that in the dual-well problem we have $v_0(x) \equiv 0$.

We follow the reasoning of [25]. By construction, V is not increasing in the L_∞-norm and has bounded variation. This is clear for the Cauchy problem, from [51] where interpolated fluxes and velocities are used. For the IBVP the boundary condition has to be taken into account. The outflow boundary has no influence on the solution. The inflow boundary $V^0(t)$ is approximated by a step function $V^0_{\Delta t}$. The wave propagation from the inflow boundary will, by construction, be decreasing in the L_∞-norm. It adds not more to the variation than the incoming wave has itself, see also [32]. Furthermore, all waves have finite speed of propagation, also the ones at the inflow boundary, so the solution is Lipschitz-continuous in time with respect to the L_1-norm. Each solution satisfies the entropy condition for the perturbed equation ($V_{0,\Delta x}, V^0_{\Delta t}$ the piecewise constant approximations of the initial and inflow condition respectively) and thus is an entropy solution.

In [32] also a stability result is given. This results allows us to consider $V_{0,\Delta x}$ and $V^0_{\Delta t}$ for the IC and BC instead of the exact conditions. We do not interpolate the flux and velocity field, so the stability result needs not to be as general as in [25].

We summerize these considerations in the following lemma

Lemma 2.5.2. *Let $V(\widetilde{y}, t)$ be a solution of (2.84) obtained by the front tracking method. Then V satisfies the following estimates*

$$\|V(.,t)\|_\infty \leq \max\left(\|V_0\|_\infty, \left\|V^0(.)\right\|_\infty\right),$$

$$TV_{\widetilde{y}}(V(.,t)) \leq TV_{\widetilde{y}}(V_0) + TV_t(V^0(.)) \leq TV_{\widetilde{y}}(V_0) + Ct,$$

$$\|V(.,t) - V_0\|_{1,\widetilde{y}} \leq Ct,$$

where C is a constant depending on the data. The solution can be constructed by front tracking in a finite number of steps for any $t > 0$. It is stable with respect to initial and boundary data: let V^1 and V^2 be two solutions corresponding to ICs V_0^1 and V_0^2, respectively, and corresponding to BCs $V^{0,1}$ and $V^{0,2}$, respecively. Then, we have

$$\begin{aligned} \left\|V^1(.,t) - V^2(.,t)\right\|_{1,\widetilde{y}} &\leq \left\|V_0^1 - V_0^2\right\|_{1,\widetilde{y}} + t\,\|f\|_{Li} \left\|V^{0,1} - V^{0,2}\right\|_\infty \\ &\leq \left\|V_0^1 - V_0^2\right\|_{1,\widetilde{y}} + Ct, \end{aligned}$$

where the constant C depends on the data (flux, velocity and initial condition).

This lemma can be easily rewritten in terms of variable v and coordinate y as F is a monotone increasing function.

Lemma 2.5.3. *Let $v(y,t)$ be a solution of (2.77) obtained by the front tracking method as described before. Then v satisfies the following estimates*

$$\|F(v(.,t))\|_\infty \leq \max\left(\|F(v_0)\|_\infty, \left\|F(v^0(.))\right\|_\infty\right),$$

$$\|v(.,t)\|_\infty \leq \max\left(\|v_0\|_\infty, \left\|v^0(.)\right\|_\infty\right),$$

$$TV_y(F(v(.,t))) \leq TV_y(F(v_0)) + TV_t(F(v^0(.))) \leq TV_y(F(v_0)) + Ct,$$

$$\left\|\frac{F(v(.,t)) - F(v_0)}{G}\right\|_1 \leq Ct,$$

where C is a constant depending on the data. The solution can be constructed by front tracking in a finite number of steps for any $t > 0$. It is stable with respect to initial and boundary data: let v^1 and v^2 be two solutions corresponding to ICs v_0^1 and v_0^2, respectively, and corresponding to BCs $v^{0,1}$ and $v^{0,2}$, respectively;, as the value of u cannot change left or right of the shock (it is the constant u_l and u_r respectively) then we have

$$\left\|\frac{F(v^1) - F(v^2)}{G}\right\|_1 \leq \left\|\frac{F(v_0^1) - F(v_0^2)}{G}\right\|_1 + t\left\|F^{-1}\right\|_{Li}\left\|F(v^{0,1}) - F(v^{0,2})\right\|_\infty,$$

where the constant C depends on the data (flux, velocity and initial condition).

Remark 2.5.5. *The stability result can be extended to different flux and different velocity fields with a Kružkov analysis, see [25, 51] and (2.4). Qualitatively, the result is the same:*

$$TV_{\widetilde{y}}(V(.,t)) \leq TV_{\widetilde{y}}(V_0) + Ct$$

and

$$\|V^1 - V^2\|_{1,\widetilde{y}} \leq \|V_0^1 - V_0^2\|_{1,\widetilde{y}} + Ct,$$

where the constant C depends on the data (flux, velocity and initial condition). More importantly, as different velocity fields are considered, the transformation y to $\widetilde{y}$ can be seen as a velocity field, allowing to write ([51])

$$\|F(v^1) - F(v^2)\|_1 \leq \|F(v_0^1) - F(v_0^2)\|_1 + Ct, \tag{2.85}$$

which is a stronger stability result than the one of Lemma 2.5.3. We stress again that to obtain this result a separate Kružkov analysis must be done taking into account velocity fields (like G).

Remark 2.5.6. *Lemma 2.5.3 can be further refined by invoking the Lipschitz continuity of* F^{-1}*:* $|v_1 - v_2| = \frac{dF^{-1}}{ds}(\epsilon)|V_1 - V_2| \leq C|V_1 - V_2| = C|F(v_1) - F(v_2)|$*, and by taking into account that there are two constants* C_1, C_2 *such that* $0 < C_1 < G < C_2$*.*

1D-Projection step

After the transport step, a projection step is done. If we consider a timestep Δt, then starting from $v(x, t_n) = v^n$, we arrive after one transport step at $\mathcal{T}_{\delta,\Delta x,\Delta t} v^n = v(x, t_{n+1}) = \widetilde{v}^{n+\frac{1}{2}}$. With projection to the fixed grid, we then obtain

$$\pi \widetilde{v}^{n+\frac{1}{2}} = v^{n+\frac{1}{2}}.$$

The following lemma is straigthforward ([26]).

Lemma 2.5.4. *Let* $h(x,t) \in BV(\mathbb{R})$ *be a function consisting from piecewise constant parts and continuous monotone rarefaction waves, and let* π *and* $\widetilde{\pi}$ *be the projection operators from (2.81). We have that*

$$TV(F(h)) = TV(F(g)) \geq TV(\widetilde{\pi}F(h)) = TV(F(\pi h)),$$

where $g(x,t)$ *is the function consisting of* h*, where the rarefaction waves are replaced by piecewise constant functions interpolating the rarefaction waves.*

Proof. Let $F(h_c)$ be a continuous approximation to $F(h)$ defined as follows. In a small neighbourhoud of each jump, we let $F(h_c)$ be a linear interpolation between the two constant values $F(v_l)$ and $F(v_r)$. Then $TV(F(h)) = TV(F(h_c)) \geq TV(\widetilde{\pi}F(h))$, since $\widetilde{\pi}F(h)$ is a particular partition of $F(h_c)$.
Now let g be as defined. As the rarefaction waves are monotone, and F is an ascending function, we will have for every rarefaction wave s that $TV(F(s)) = TV(F(g|_{\mathrm{dom}(s)}))$. Therefore, by extension $TV(F(h)) = TV(F(g))$. □

We can also derive a result for the variation in time. We may do this multidimensional as this does not add complexity.

Lemma 2.5.5. *If* $C = \Delta x/\tau$*, the projection operator satisfies*

$$\|F(v^{n+\frac{1}{2}}) - F(\widetilde{v}^{n+\frac{1}{2}})\|_1 = \int_\Omega |\widetilde{\pi}F(v) - F(v)| \mathrm{d}x\mathrm{d}y \leq C\tau TV(F(\widetilde{v}^{n+\frac{1}{2}})).$$

Proof. We have that $F(v^{n+\frac{1}{2}}) = F_{ij}$, a constant in the volume $\Omega_{ij} = (x_i, x_{i+1}) \times (y_i, y_{i+1})$. Due to the properties of the projection operator (conservation of $F(v)$), we have $\underline{F}_{ij} \leq F_{ij} \leq \overline{F}_{ij}$, where $\underline{F}_{ij} = \min(F(\widetilde{v}^{n+\frac{1}{2}}))$ in the volume Ω_{ij}, and analogously, $\overline{F}_{ij}$ is the maximum.

If F is a smooth function, there is a point ξ_{ij} where $F(\widetilde{v}^{n+\frac{1}{2}}(\xi_{ij})) = F_{ij}$. Applying the mean value theorem to $F(\widetilde{v}^{n+\frac{1}{2}})$, and with the point ξ_{ij} less than a distances Δx and Δy separated from every point of Ω_{ij}, we obtain

$$\|F(v^{n+\frac{1}{2}}) - F(\widetilde{v}^{n+\frac{1}{2}})\|_1 \leq \int_\Omega |\partial_x F(\widetilde{v}^{n+\frac{1}{2}}))| \Delta x \mathrm{d}x \mathrm{d}y + \int_\Omega |\partial_y F(\widetilde{v}^{n+\frac{1}{2}}))| \Delta y \mathrm{d}x \mathrm{d}y$$
$$= C\tau TV(F(\widetilde{v}^{n+\frac{1}{2}})).$$

In the case that F consists of shocks, we have

$$\begin{aligned} \|F(v^{n+\frac{1}{2}}) - F(\widetilde{v}^{n+\frac{1}{2}})\|_1 &= \sum_{ij} \int_{\Omega_{ij}} |F_{ij} - F(\widetilde{v}^{n+\frac{1}{2}})| \, dxdy \\ &\leq \sum_{ij} (\overline{F}_{ij} - \underline{F}_{ij}) \Delta x \Delta y \leq C\tau TV(F(\widetilde{v}^{n+\frac{1}{2}})). \end{aligned}$$

□

The above lemma will be usefull to relate all errors made to the total variation (TV) of the initial condition. Note that, independently, we have the following result:

Lemma 2.5.6. *If $\|F(\widetilde{v}^{n+\frac{1}{2}})\|_\infty \leq C$, the projection operator satisfies*

$$\|F(v^{n+\frac{1}{2}}) - F(\widetilde{v}^{n+\frac{1}{2}})\|_1 \to 0, \quad \textit{as} \quad \Delta x \to 0.$$

Proof. Due to the properties of the projection operator (conservation of $F(v)$), we have $\underline{F}_{ij} \leq F_{ij} \leq \overline{F}_{ij}$, where $\underline{F}_{ij} = \min(F(\widetilde{v}^{n+\frac{1}{2}}))$ in the volume Ω_{ij}, and analogously, $\overline{F}_{ij}$ is the maximum. We have

$$\|F(v^{n+\frac{1}{2}}) - F(\widetilde{v}^{n+\frac{1}{2}})\|_1 \leq \sum_{ij} |\Omega_{ij}| (\overline{F}_{ij} - \underline{F}_{ij}) \to 0, \quad \text{as} \quad \Delta x \to 0,$$

which follows from the fact that two Riemann sums of the same integral $0 < \int F(\widetilde{v}^{n+\frac{1}{2}}) < C$ are considered. □

2D hyperbolic step and projection

We now consider the full 2 dimensional problem. The boundedness is evident. The TV result however is less clear.

Recall that

$$[\pi \circ \mathcal{T}_{\delta,\Delta x,\Delta t} \circ \pi \circ \mathcal{T}_{\delta,\Delta y,\Delta t}]\, v^n = [\pi \circ I \circ \mathcal{T}_{\delta,\Delta y,\Delta t}]\, v^n = \pi \widetilde{v}^{n+\frac{1}{2}} = v^{n+\frac{1}{2}}, \tag{2.86}$$

where the transport operator is in terms of $u = F(v)$, instead of v. For the combination of transport and projection it is best to perform both calculations in the same coordinates. We do both in terms of u. We have, with $F(v) = u$,

$$[\widetilde{\pi} \circ \mathcal{T}_{\delta,\Delta x}(\Delta t) \circ \widetilde{\pi} \circ \mathcal{T}_{\delta,\Delta y}(\Delta t)]\, u^n = [\widetilde{\pi} \circ I \circ \mathcal{T}_{\delta,\Delta y}(\Delta t)]\, u^n = \widetilde{\pi} \widetilde{u}^{n+\frac{1}{2}} = u^{n+\frac{1}{2}}, \tag{2.87}$$

where $\mathcal{T}_{\delta,\Delta y}(\Delta t)$ is the transport operator acting on v. We prove the following lemma.

Lemma 2.5.7. *For (2.86) we have that*

$$TV_{x,y}(F(v^{n+\frac{1}{2}})) \leq TV_{x,y}(F(v^n)) + C\tau,$$

or equivalently

$$TV_{x,y}(u^{n+\frac{1}{2}}) \leq TV_{x,y}(u^n) + C\tau,$$

Proof. The proof goes along the lines of [26]. The operator $\mathcal{T}_{\delta,\Delta x}(\Delta t)$ is the identity in our case, so it remains to show the assertion for $\widetilde{\pi} \circ \mathcal{T}_{\delta,\Delta y}(\Delta t)$, the transport in the y direction. Two ingredients will be used in this proof: first the properties of the solution for the 1D problem in y and, secondly, the stability of the solution for the 1D problem in y. These two parts correspond to the two parts of the 2 dimensional total variation, given by

$$TV_{xy}h(x,y) = \int_{y^{(1)}}^{y^{(2)}} TV_x(h(x,y))dy + \int_{x^{(1)}}^{x^{(2)}} TV_y(h(x,y))dx.$$

In the hyperbolic step the initial condition is a piecewise constant profile u_{ij}, $i = 1, \ldots, N_1$ and $j = 1, \ldots, N_2$. If, for each fixed x, $u(x,y)$ is a piecewise constant function in y, we write

$$u_j(x) = u|_{j\Delta y < y < (j+1)\Delta y}(x,y),$$

and similarly

$$u_i(y) = u|_{i\Delta x < x < (i+1)\Delta x}(x,y).$$

We also denote

$$u_i(y,t) = \mathcal{T}_{\delta,\Delta y}(t)u_i(y),$$

which contains shocks and rarefaction waves. Due to the absence of transport in the x direction,

$$u_j(x,t) = \mathcal{T}_{\delta,\Delta x}(t)u_j(x) = u_j(x).$$

According to Lemma 2.5.2, the solution of the 1D problem satisfies the inequality

$$TV_y(u_i(y,\Delta t)) \leq TV_y(u_i(y)) + C\Delta t,$$

where y may be used instead of $\widetilde{y}$ as this is the one dimensional variation. Furthermore, Lemma 2.5.4 yields:

$$TV_y(u_{i,j}^{n+\frac{1}{2}}) \leq TV_y(u_i(y,\Delta t)).$$

Combining these results we have

$$TV_y(u^{n+\frac{1}{2}})\Delta x \leq TV_y(u^n)\Delta x + C\tau\Delta x.$$

For $TV_x(u^{n+\frac{1}{2}})$ we will use the stability result (2.85). If we consider two adjacent x-strips i and $i+1$ with solution $u_i(y,t)$ and $u_{i+1}(y,t)$, then these solutions are obtained in a y strip starting from different initial conditions $u_i(y)$ and $u_{i+1}(y)$ and slightly different velocity functions $G(x,y)$. The stability result says that

$$\begin{aligned}\int_{y^{(1)}}^{y^{(2)}} |u_{i+1}(t+\Delta t) - u_i(t+\Delta t)|dy &\leq \int_{y^{(1)}}^{y^{(2)}} |u_{i+1}(t) - u_i(t)|dy + C(\eta)\Delta t \\ &\leq \sum_{j=1}^{N_2} |u_{i+1,j} - u_{i,j}|\Delta y + C(\eta)\Delta t,\end{aligned}$$

where $C(\eta)$ indicates that C depends on the data: flux function, velocity field and initial data, which are all different from one strip to the other. However, due to refinement of the grid, the two strips are less separated (i.e. at least one is in a different position), and so $C(\eta)$ changes as the data on which it depends changes. To be more exact, from the Kružkov analysis (Lemma 2.5.2, [51]) we know the form of $C(\eta)$, and we can extract the dependance on Δx, which gives

$C(\eta) = C\Delta x + O(\Delta x^2)$, where C is a constant depending on the flux, velocity and initial data of strip i only. Therefore, we obtain

$$\int_{y^{(1)}}^{y^{(2)}} |u_{i+1}(t+\Delta t) - u_i(t+\Delta t)|dy \leq \sum_{j=1}^{N_2} |u_{i+1,j} - u_{i,j}|\Delta y + C\Delta t\Delta x. \quad (2.88)$$

By definition of the projection operator we have

$$\begin{aligned} &\int_{y^{(1)}}^{y^{(2)}} |u_{i+1}(y,t+\Delta t) - u_i(y,t+\Delta t)|dy \\ = &\sum_{j=1}^{N_2} \int_{y_j}^{y_{j+1}} |u_{i+1}(y,t+\Delta t) - u_i(y,t+\Delta t)|dy \\ \geq &\sum_{j=1}^{N_2} \Big| \int_{y_j}^{y_{j+1}} u_{i+1}(y,t+\Delta t) - u_i(y,t+\Delta t)dy\Big| = \sum_{j=1}^{N_2} |u_{i+1,j}^{n+\frac{1}{2}} - u_{i,j}^{n+\frac{1}{2}}|\Delta y \end{aligned}$$

Combining with (2.88), and summing over $i = 1\ldots, N_1 - 1$ yields

$$\sum_{ij} |u_{i+1,j}^{n+\frac{1}{2}} - u_{i,j}^{n+\frac{1}{2}}|\Delta y \leq \sum_{ij} |u_{i+1,j} - u_{i,j}|\Delta y + C\Delta t$$

where $\sum_i \Delta x = C$.

Then

$$TV_{xy}u^{n+1/2} \leq TV_{xy}u^n + C\tau.$$

□

2.5.5 The parabolic step

The boundedness of the diffusion and total variation decreasing property might be expected in advance. For the exact solution of the parabolic step it is straightforward to show this property using the kernel, see [27]. However, for a FVW type of discretization the proof is not as simple, and has not been done to our knowledge for the 2D case. However, it turns out that the property is not needed to prove convergence. It suffices to show that the TV is bounded over the combined 3 steps: transport, projection and diffusion. We will proceed in this way. First, we give the approximation scheme used, then we prove boundedness and a stability result as obtained for the other steps. Finally, we prove a TV property, that will allow us in the next section to combine the diffusion step with the other steps, and get an upper bound on the total variation.

Possible approximation schemes

The approximation scheme (2.68) is used to prove convergence. We rewrite it as

$$\begin{aligned}\omega F(w_{i,j}) + \left(a_{i+1,j}\tfrac{\Delta y}{\Delta x^+} + a_{i,j}\tfrac{\Delta y}{\Delta x^-} + b_{i,j+1}\tfrac{\Delta x}{\Delta y^+} + b_{i,j}\tfrac{\Delta x}{\Delta y^-}\right)\tau w_{i,j} = \\ \left[\tau\tfrac{\Delta y}{\Delta x^+}a_{i+1,j}\right] w_{i+1,j} + \left[\tau\tfrac{\Delta y}{\Delta x^-}a_{i,j}\right] w_{i-1,j} \\ + \left[\tau\tfrac{\Delta x}{\Delta y^+}b_{i,j+1}\right] w_{i,j+1} + \left[\tau\tfrac{\Delta x}{\Delta y^-}b_{i,j}\right] w_{i,j-1} + \omega F(w^{n-1}_{i,j}),\end{aligned} \qquad (2.89)$$

where $\omega = \omega_{ij} = \frac{|V_{ij}|}{g_{ij}}$, $F(w^{n-1}_{i,j})$ is the value of F_{ij} obtained in the projection step, and where we allowed for a nonequidistant grid, so $\Delta x^+ = x^1_{i+1} - x^1_i$, etc.

Taking into account the boundary conditions, we put $a_{ij} \equiv 0$ for the points $\{x_1, y_j\}$ and $a_{i+1,j} \equiv 0$ for the points $\{x_{N_1}, y_j\}$, $j = 1, \ldots, N_2$. Moreover, for $\{x_i, y_1\}$, $i = 1, \ldots, N_1$, we take $b_{ij} \equiv 0$ in (2.89), and $b_{i,j+1} \equiv 0$ for the points $\{x_i, y_{N_2}\}$.

We will use (2.89), assuming $w_{i,j}$ to be its exact solution. In practice, several different approximation schemes are considered. We revise them for future reference. The nonlinear system of algebraic equations (2.89) will be solved numerically by Newton type iterations, starting with $C \equiv C^{n-1}$. This implies solving a matrix equation in every iteration step. This can be done up to a desired order of accuracy.

We also consider the relaxation method, (2.69), which we write here as

$$\begin{aligned}\omega\lambda^{(l-1)}_{i,j}(w^{(l)}_{i,j} - w^{n-1}_{i,j}) + \left(a_{i+1,j}\tfrac{\Delta y}{\Delta x^+} + a_{i,j}\tfrac{\Delta y}{\Delta x^-} + b_{i,j+1}\tfrac{\Delta x}{\Delta y^+} + b_{i,j}\tfrac{\Delta x}{\Delta y^-}\right)\tau w^{(l)}_{i,j} = \\ \left[\tau\tfrac{\Delta y}{\Delta x^+}a_{i+1,j}\right] w^{(l)}_{i+1,j} + \left[\tau\tfrac{\Delta y}{\Delta x^-}a_{i,j}\right] w^{(l)}_{i-1,j} \\ + \left[\tau\tfrac{\Delta x}{\Delta y^+}b_{i,j+1}\right] w^{(l)}_{i,j+1} + \left[\tau\tfrac{\Delta x}{\Delta y^-}b_{i,j}\right] w^{(l)}_{i,j-1},\end{aligned} \qquad (2.90)$$

where l is a iteration parameter and

$$\lambda^{(l)}_{i,j} := \frac{F(w^{(l)}_{i,j}) - F(w^{n-1}_{i,j})}{w^{(l)}_{i,j} - w^{n-1}_{i,j}}, \quad \lambda^{(0)}_{i,j} := F'(w^{n-1}_{i,j})$$

is a relaxation function. If the convergence conditions

$$|\lambda^{(l_0)}_{i,j} - \lambda^{(l_0-1)}_{i,j}| < \tau, \quad \sum_{ij}\left|w^{(l_0)}_{i,j} - w^{(l_0-1)}_{i,j}\right| < \epsilon, \qquad (2.91)$$

ϵ being a small tolerance, are met, we stop the iterations and define $w_{i,j} := w_{i,j}^{(l_0)}$.

For convenience, we rewrite (2.89) in two equivalent forms. For simplicity, set $\Delta x_+ = \Delta x_- = \Delta y_+ = \Delta y_- = \Delta x = \Delta y$, and rewrite (2.89) as

$$\frac{F(w_{i,j}^n) - F(w_{i,j}^{n-1})}{g_{ij}} = \frac{\tau}{\Delta x \Delta y}\left(a_{i+1,j}\widetilde{Z}_{i+1,j}^i - a_{ij}\widetilde{Z}_{ij}^i + b_{i,j+1}\widetilde{Z}_{i,j+1}^j - b_{ij}\widetilde{Z}_{ij}^j\right), \tag{2.92}$$

where $\widetilde{Z}_{ij}^i = w_{ij} - w_{i-1,j}$ and $\widetilde{Z}_{ij}^j = w_{ij} - w_{i,j-1}$. This equation is transformed by setting

$$\frac{F(w_{ij})}{g_{ij}} = u_{ij}, \quad w_{ij} = F^{-1}(g_{ij}u_{ij}) = A(x_i, y_j, u_{ij}). \tag{2.93}$$

The function A is monotone increasing in u: F is strictly increasing in w, so F^{-1} is strictly increasing in u, as g is fixed at a gridpoint. The scheme is therefore equivalent to

$$\begin{aligned} u_{i,j} - u_{i,j}^{n-1} = \frac{\tau}{\Delta x^2}\,&(a_{i+1,j}(A(u_{i+1,j}) - A(u_{i,j})) - a_{ij}(A(u_{i,j}) - A(u_{i-1,j})) + \\ &b_{i,j+1}(A(u_{i,j+1}) - A(u_{i,j})) - b_{ij}(A(u_{i,j}) - A(u_{i,j-1}))) . \end{aligned} \tag{2.94}$$

This is a more complicated version of the implicit scheme used in [25], as here A also depends on the space coordinates because of g. We will apply the mean value theorem as follows:

$$\begin{aligned} &A(x_{i+1}, y_j, u_{i+1,j}) - A(x_i, y_j, u_{i,j}) \\ &\quad = A(x_{i+1}, y_j, u_{i+1,j}) - A(x_{i+1}, y_j, u_{i,j}) + A(x_{i+1}, y_j, u_{i,j}) - A(x_i, y_j, u_{i,j}) \\ &\qquad = \frac{\partial A}{\partial u}(x_{i+1}, y_j, \xi_{i+1,j})(u_{i+1,j} - u_{ij}) + \left(\frac{\partial A}{\partial x}\right)_{i+1,j}\Delta x, \end{aligned}$$

for a suitable $\xi_{i+1,j}$ and where $(\partial A/\partial x)_{i+1,j} = \partial A/\partial x(\eta_{i+1}, v_j, u_{ij})$ for a suitable η_{i+1}. Then, we rewrite (2.94) as

$$\begin{aligned} u_{i,j} - u_{i,j}^{n-1} \quad = \quad & \widetilde{\alpha}_{i+1,j}Z_{i+1,j}^i - \widetilde{\alpha}_{i,j}Z_{ij}^i + \widetilde{\beta}_{i,j+1}Z_{i,j+1}^j - \widetilde{\beta}_{i,j}Z_{i,j}^j \\ & + \tfrac{\tau}{\Delta x^2}\,(a_{i+1,j}(\partial A/\partial x)_{i+1,j}\Delta x - a_{ij}(\partial A/\partial x)_{i,j}\Delta x + \\ & + b_{i,j+1}(\partial A/\partial y)_{i,j+1}\Delta y - b_{ij}(\partial A/\partial y)_{i,j}\Delta y) \end{aligned} \tag{2.95}$$

where $Z_{i,j}^i = u_{i,j} - u_{i-1,j}$, $Z_{i,j}^j = u_{i,j} - u_{i,j-1}$ and

$$\widetilde{\alpha}_{i+1,j} = a_{i+1,j}\tfrac{\tau}{\Delta x^2}A'(\xi_{i+1,j}),$$

and identically

$$\widetilde{\beta}_{i,j+1} = b_{i,j+1}\tfrac{\tau}{\Delta x^2}A'(\zeta_{i,j+1}),$$

for suitable $\xi_{i+1,j}$, $\zeta_{i,j+1}$. Note that a, b are bounded and strictly positive. The degeneracy only comes from the A' term. We can rewrite the above as,

$$\begin{aligned} u_{i,j} - u_{i,j}^{n-1} \quad = \quad & \widetilde{\alpha}_{i+1,j}Z^i_{i+1,j} - \widetilde{\alpha}_{i,j}Z^i_{ij} + \widetilde{\beta}_{i,j+1}Z^j_{i,j+1} - \widetilde{\beta}_{i,j}Z^j_{i,j} \\ & +\tfrac{\tau}{\Delta x}\left(\gamma_{i+1,j} - \gamma_{i,j} + \delta_{i,j+1} - \delta_{i,j}\right), \end{aligned} \tag{2.96}$$

where

$$\gamma_{i+1,j} = a_{i+1,j}(\partial A/\partial x)_{i+1,j}, \qquad \delta_{i,j+1} = b_{i,j+1}(\partial A/\partial y)_{i,j+1}.$$

An approximation of (2.96) is given by

$$u_{i,j} - u_{i,j}^{n-1} \quad = \quad \widetilde{\alpha}_{i+1,j}Z^i_{i+1,j} - \widetilde{\alpha}_{i,j}Z^i_{ij} + \widetilde{\beta}_{i,j+1}Z^j_{i,j+1} - \widetilde{\beta}_{i,j}Z^j_{i,j}, \tag{2.97}$$

which is precisely the scheme used in [27].

Remark 2.5.7. *In the degenerate point* $(0,0)$ *we have that* $A(u) \equiv 0$ *and* $A' = 0$. *In the region where no concentration is present,* $u = 0$*, this is also the case. At the moving interface, we have* $\lim_{v\to 0+} F'(v) = +\infty$*, so* $\lim_{u\to 0+} A'(u) = 0$.

We summerize the 4 schemes that we have deduced:
Scheme (1): relaxation method. $W^{n,l}$ is the solution of

$$\begin{aligned} & \frac{\lambda_{ij}}{g_{ij}}\left(W^{n,l}_{i,j} - W^{n-1,l}_{i,j}\right) + \left(a_{i+1,j} + a_{ij} + b_{i,j+2} + b_{ij}\right)\frac{\tau}{\Delta x^2}W^{n,l}_{i,j} - \\ & \qquad \frac{\tau}{\Delta x^2}\left[a_{ij}W^{n,l}_{i-1,j} + a_{i+1,j}W^{n,l}_{i+1,j} + b_{i,j+1}W^{n,l}_{i,j+1} + b_{ij}W^{n,l}_{i,j-1}\right] = 0, \end{aligned} \tag{2.98}$$

where λ_{ij} is the last calculated relaxation parameter.
Scheme (2): FV scheme. W^n is the solution of

$$\begin{aligned} & \frac{F(W_{i,j}) - F(W^{n-1}_{i,j})}{g_{ij}} + \left(a_{i+1,j} + a_{ij} + b_{i,j+1} + b_{ij}\right)\frac{\tau}{\Delta x^2}W_{i,j} - \\ & \qquad \frac{\tau}{\Delta x^2}\left[a_{ij}W_{i-1,j} + a_{i+1,j}W_{i+1,j} + b_{i,j+1}W_{i,j+1} + b_{ij}W_{i,j-1}\right] = 0. \end{aligned} \tag{2.99}$$

Scheme (3): degenerate variable coefficient diffusion method. w is the solution of

$$\begin{aligned} w_{i,j} - w^{n-1}_{i,j} = {} & \widetilde{\alpha}_{i+1,j}Z^i_{i+1,j} - \widetilde{\alpha}_{i,j}Z^i_{ij} + \widetilde{\beta}_{i,j+1}Z^j_{i,j+1} - \widetilde{\beta}_{i,j}Z^j_{i,j} \\ & + \frac{\tau}{\Delta x}\left(\gamma_{i+1,j} - \gamma_{i,j} + \delta_{i,j+1} - \delta_{i,j}\right), \end{aligned} \tag{2.100}$$

Scheme (4): degenerate diffusion method. u is the solution of

$$u_{i,j} - u_{i,j}^{n-1} = \widetilde{\alpha}_{i+1,j} Z_{i+1,j}^{i} - \widetilde{\alpha}_{i,j} Z_{ij}^{i} + \widetilde{\beta}_{i,j+1} Z_{i,j+1}^{j} - \widetilde{\beta}_{i,j} Z_{i,j}^{j} \tag{2.101}$$

Scheme (1) and (2) will be used in our practical computations. Convergence of scheme (1) to the solution (2.79) has been proven in [35] in the L_2-norm. We will prove convergence of the operator splitting method by using W^n. As an extra point we can show afterwards that $W^{n,l}$ converges to W^n. We have added scheme (4), as results for this scheme exist in the litarature. A connection of scheme (4) with our scheme (2) can be obtained through scheme (3).

Boundedness and stability

We only consider scheme (2). We prove the boundedness and stability in an analogue form to the one obtained for the transport and projection.

Lemma 2.5.8. *Let W^n and V^n be approximate solutions of (2.79) generated by the scheme (2.99). Then, one has*

$$\|F(W^n)\|_\infty \leq \|F(W^0)\|_\infty, \quad \left\|\frac{F(W^n) - F(V^n)}{g}\right\|_1 \leq \left\|\frac{F(W^0) - F(V^0)}{g}\right\|_1 .$$

Proof. Choose in (2.99) $i = l$, $j = k$, such that $W_{l,k}^n = \max_{ij} W_{ij}$. Due to the properties of F, $F(W_{l,k}^n) = \max_{ij} F(W_{ij})$. We directly obtain $\max F(W_{i,j}^n) \leq \max F(W_{i,j}^{n-1})$, and therefore also $\max W_{i,j}^n \leq \max W_{i,j}^{n-1}$. This can be repeated for $\min_{ij} W_{ij}$, which proves the first assertion by induction on n. The second assertion follows by subtracting (2.99) for W from the equation for V, with $d_{ij} = W_{ij}^n - V_{ij}^n$. This gives

$$\left[\frac{F'(\xi_{ij})}{g_{ij}} + (a_{i+1,j} + a_{ij} + b_{i,j+1} + b_{ij}) \frac{\tau}{\Delta x^2}\right] d_{ij}$$
$$= \frac{\tau}{\Delta x^2} \left[a_{ij} d_{i-1,j} + a_{i+1,j} d_{i+1,j} + b_{i,j+1} d_{i,j+1} + b_{ij} d_{i,j-1}\right] + \frac{F(W_{i,j}^{n-1}) - F(V_{i,j}^{n-1})}{g_{ij}}.$$

Taking absolute values and summation over i and j, leads to

$$\left\|\frac{F(W^n) - F(V^n)}{g}\right\|_1 \leq \left\|\frac{F(W^{n-1}) - F(V^{n-1})}{g}\right\|_1 ,$$

which proves the lemma. □

Total variation result of [25]

In [25] the boundedness of the total variation of scheme (4) is suggested for 2 or more dimensions without proof. The techniques given are however not extendable to higer dimensions. For clarity, we recall the results. We begin with two esimates that can be readily extended to two or more dimensions.

Lemma 2.5.9. *Let u^n and v^n be two solutions of (2.101). Then*

$$\|u^n\|_\infty \leq \|u^0\|_\infty, \quad \|u^n - v^n\|_1 \leq \|u^0 - v^0\|_1.$$

Proof. Eq. (2.101) is the implicit scheme used in [25] and the above result is Lemma 2.2 from [25]. The L_∞ estimate follows directly from (2.101) by the same reasoning as in Lemma 2.5.8. The L_1-stability is not proven explicitely in [25, 27], but can be easily obtained by the arguments given. We show this stability in 2 dimensions.

Rewrite (2.101) for u^n in its original form (2.94), and substract it from the equation for v^n. By the mean value theorem we have $A(u^n_{i,j}) - A(v^n_{i,j}) = A'(\zeta^n_{ij})(u^n_{i,j} - v^n_{i,j})$ for some appropriate ζ^n_{ij}. Writing $U^n_{ij} = u^n_{i,j} - v^n_{i,j}$, we obtain

$$\Big(1 + \frac{\tau}{\Delta x} A'(\zeta^n_{i,j}) \left(a^E + a^W + b^N + b^S\right)\Big) U^n_{ij} = U^{n-1}_{ij} + \frac{\tau}{\Delta x} \big(a^E A'(\zeta^n_{i+1,j}) U^n_{i+1,j} \\ + a^W A'(\zeta^n_{i-1,j}) U^n_{i-1,j} + b^N A'(\zeta^n_{i,j+1}) U^n_{i,j+1} + b^S A'(\zeta^n_{i,j-1}) U^n_{i,j-1}\big) .$$

The first factor in the left hand side is positive as A is an increasing function in u. Taking absolute values, and summing over i and j, we obtain

$$\|u^n - v^n\|_1 \leq \|u^{n-1} - v^{n-1}\|_1,$$

from which the required result follows by induction on n. □

The following result cannot be extended to higher dimensions:

Lemma 2.5.10. *Let u^n be an approximate solution generated by the 1D version of (2.101). Then*

$$TV(u^n) \leq TV(u^0)$$

Proof. We rely upon [25]. In 1D, the equation is

$$u_i = u^{n-1}_i + \alpha_{i+1} Z_{i+1} - \alpha_i Z_i \tag{2.102}$$

We have that $TV(u) = \sum_i |Z_i|$. Proceeding similar as in the previous lemmas we deduce from (2.102) that

$$Z_i = Z_i^n + \alpha_{i+1} Z_{i+1} + \alpha_{i-1} Z_{i-1} - 2\alpha_i Z_i.$$

We bring the third term of the rhs to the left, take absolute values and sum over i to obtain the result. □

Another approach is needed to obtain TV diminishing properties of scheme (2.99).

A sufficient total variation result

We begin by rewriting (2.99) as

$$\begin{aligned} \tfrac{F(W_{i,j}) - F(W_{i,j}^{n-1})}{g_{ij}} &= \tfrac{\tau}{\Delta x} \left[a_{i+1,j} D_{i+1,j}^{i} W - a_{ij} D_{i,j}^{i} W \right] \\ &\quad + \tfrac{\tau}{\Delta y} \left[b_{i,j+1} D_{i,j+1}^{j} W - b_{ij} D_{i,j+1}^{j} W \right], \end{aligned} \tag{2.103}$$

where

$$D_{ij}^{i} W = \tfrac{W_{i,j} - W_{i-1,j}}{\Delta x}, \quad D_{ij}^{j} W = \tfrac{W_{i,j} - W_{i,j-1}}{\Delta y}.$$

We have the following lemma:

Lemma 2.5.11. *Let W^n be the solution of (2.99) with $a,\, b \geq \delta > 0$, then*

$$\tau TV(W^n) \leq C_1 \alpha \tau - \frac{C_2}{\alpha} \sum_{ij} \frac{1}{g_{ij}} \left[F(W_{ij}^n) - F(W_{ij}^{n-1}) \right] W^n \Delta x \Delta y, \tag{2.104}$$

where C_1 and $C_2 > 0$ and where $\alpha > 0$ is arbitrary but fixed.

Proof. We multiply both sides of (2.103) with W_{ij} and $\Delta x \Delta y$. We sum over i and j, and apply Abels' summation (A.2) in the rhs (integration by parts in discrete form),. Using the boundary conditions ($D_{ij} W = 0$), we arrive at

$$\begin{aligned} &\sum_{ij} \frac{1}{g_{ij}} \left[F(W_{ij}^n) - F(W_{ij}^{n-1}) \right] W_{ij}^n \Delta x \Delta y \\ &\qquad + \tau \sum_{ij} \left[a (D_{ij}^{i} W^n)^2 + b (D_{ij}^{j} W^n)^2 \right] \Delta x \Delta y = 0. \end{aligned} \tag{2.105}$$

For arbitrary $\alpha > 0$ (fixed), the Cauchy inequality can be used, i.e. $|f| \leq \frac{\alpha}{2} + \frac{1}{2\alpha} f^2$. Combination with (2.105) leads to

$$\begin{aligned}
&\sum_{ij} \left[\left| D^i_{ij} W^n \right| + \left| D^j_{ij} W^n \right| \right] \Delta x \Delta y \qquad (2.106)\\
\leq \;& \frac{1}{2}\alpha|\Omega| + \frac{1}{2}\alpha|\Omega| + \frac{1}{2\alpha\delta} \sum_{ij} \left[a(D^i_{ij} W^n)^2 + b(D^j_{ij} W^n)^2 \right] \Delta x \Delta y \\
\leq \;& \alpha|\Omega| - \frac{1}{2\alpha\delta} \frac{1}{\tau} \sum_{ij} \frac{1}{g_{ij}} \left[F(W^n_{ij}) - F(W^{n-1}_{ij}) \right] W^n_{ij} \Delta x \Delta y,
\end{aligned}$$

where we used the assumption $a, b \geq \delta > 0$. □

Remark 2.5.8. *Note that for a pure diffusion problem, the scheme (2.99), does provide an easy way to prove boundedness of total variation, see Lemma 2.5.12 below. From this lemma, it's clear that all difficulties arise from our hyperbolic step performed between* 2 *diffusion steps, for which a result as in (2.107) is not available.*

Lemma 2.5.12. *Let W^n be the solution of a pure diffusion problem obtained by (2.99), where $a,\ b \geq \delta > 0$. Then*

$$TV(W^n) \leq C_1 \alpha + \frac{C_2}{\alpha} \sum_{ij} \left[a(D^i_{ij} W^0)^2 + b(D^j_{ij} W^0)^2 \right] \Delta x \Delta y,$$

where C_1 and $C_2 > 0$ and where $\alpha > 0$ is arbitrary but fixed.

Proof. Denote $\alpha_{ij} = a_{ij} D^i_{ij} W$ and $\beta_{ij} = b_{ij} D^j_{ij} W$. Multiplying both sides of (2.103) with $W^n_{ij} - W^{n-1}_{ij}$ and $\Delta x \Delta y$, and summing on $i = 1, \ldots, N_1$ and $j = 1, \ldots, N_2$ gives

$$\begin{aligned}
\sum_{ij} \frac{F'(\xi_{ij})}{g_{ij}} (W_{ij} - W^{n-1}_{ij})^2 = \frac{\tau}{\Delta x} \sum_{ij} &\left\{ (\alpha_{i+1,j} - \alpha_{ij}) \left(W_{ij} - W^{n-1}_{ij} \right) \right. \\
&\left. + (\beta_{i,j+1} - \beta_{ij}) \left(W_{ij} - W^{n-1}_{ij} \right) \right\}.
\end{aligned}$$

We apply Abels' summation in the rhs (taking in (A.2) once $a_i = \alpha_{i+1,j}$ and once $a_j = \beta_{i,j+1}$). As we consider a homogeneous Neumann BC, we have $\alpha_{N_1+1,j} = \alpha_{1,j} = 0 = \beta_{i,1} = \beta_{i,N_2+1}$ (see the beginning of this section). Recalling that

$F' > 0$, we obtain

$$
\begin{aligned}
0 \le -\frac{\tau}{\Delta x}\sum_{ij} a_{i,j} D^i_{ij} W^n \left(D^i_{ij} W^n \Delta x - D^i_{ij} W^{n-1}\Delta x\right) \\
- \frac{\tau}{\Delta x}\sum_{ij} b_{i,j} D^j_{ij} W^n \left(D^j_{ij} W^n \Delta x - D^j_{ij} W^{n-1}\Delta x\right).
\end{aligned}
$$

We now use the identity

$$s(s-r) = \tfrac{s^2}{2} - \tfrac{r^2}{2} + \tfrac{(s-r)^2}{2},$$

from which it follows that

$$
\begin{aligned}
\sum_{ij}\left(a_{i,j}\left(D^i_{i,j}W^n\right)^2 + b_{i,j}\left(D^j_{i,j}W^n\right)^2\right) \\
\le \sum_{ij}\left(a_{i,j}\left(D^i_{i,j}W^{n-1}\right)^2 + b_{i,j}\left(D^j_{i,j}W^{n-1}\right)^2\right). \qquad (2.107)
\end{aligned}
$$

Combined with (2.106) this gives the required result. □

Note that if we do not consider a homogeneous Neumann BC, the BC will also influence the above total variation result.

2.5.6 The three steps combined

We use the following notation:

$$
\begin{aligned}
\left[\mathcal{D}_{\Delta x,\Delta y,\Delta t} \circ \pi \circ \mathcal{T}_{\delta,\Delta x,\Delta t} \circ \pi \circ \mathcal{T}_{\delta,\Delta y,\Delta t}\right] v^n &= \left[\mathcal{D}_{\Delta x,\Delta y,\Delta t} \circ \pi \circ \mathcal{T}_{\delta,\Delta y,\Delta t}\right] v^n \\
&= \mathcal{D}_{\Delta x,\Delta y,\Delta t} \circ \pi \widetilde{v}^{n+\frac{1}{2}} \\
&= \mathcal{D}_{\Delta x,\Delta y,\Delta t} \circ v^{n+\frac{1}{2}} \\
&= v^{n+1}, \qquad (2.108)
\end{aligned}
$$

Recall that F^{-1} is Lipschitz continuous, so that $(F^{-1})'(s) \le L_{iF}$, $\forall |s| < L$, or, equivalently, $F'(s) \ge 1/L_{iF}$. Before being able to prove the important TV boundedness, we need to define an auxiliar function. We introduce

$$B(s) := sF(s) - \int_0^s F(z)\,dz.$$

We have that

$$[F(u) - F(v)]\, u \geq B(u) - B(v). \tag{2.109}$$

Indeed, note that $[F(u) - F(v)]\, u = F(u)u - F(v)v - (u - v)F(v) \geq F(u)u - F(v)v - \int_v^u F(z)\, dz \equiv B(u) - B(v)$, since $F(u)$ is monotone increasing. We also have $B(s) > 0$ if $s > 0$, and $B(0) = 0$, as well as $B'(s) = sF'(s)$. Consequently, for Langmuir and Freundlich adsorption we have that $B'(s) \leq L_B$, $\forall |s| < L$. We shall not use this last estimate, as the boundedness of $B(s)$, $\forall |s| < L$, following from the boundedness of F, $\forall |s| < L$, will be sufficient for our purposes. We have the following lemma.

Lemma 2.5.13. *The approximation scheme given by (2.108) satisfies*

$$\|v^n\|_\infty \leq C, \quad \textit{and} \quad \int_0^t TV_{xy}(v_{\Delta t}(t))\, \mathrm{d}t \leq C,$$

where the constant C is independent of the space and time discretization and only depends on the domain and on a, b, $\|v^0\|_\infty$ and $\|v_0\|_\infty$. Moreover, $v_{\Delta t}(x, y, t) = v^n(x, y)$ for $t \in (t_{n-1}, t_n)$ is a piecewise constant function in (x, y).

Proof. The first inequality follows directly from Lemma 2.5.3 and Lemma 2.5.8, using the Lipschitz continuity of F^{-1}, i.e. $\|v^n\|_\infty < C\|F(v^n)\|_\infty$.

For the second inequality, we use Lemma 2.5.11, where $W^{n-1} \equiv v^{n+\frac{1}{2}}$, and we apply (2.109) to get

$$\begin{aligned}
&\tau TV(v^{n+1}) \\
&\leq C_1\alpha\tau - \tfrac{C_2}{\alpha} \textstyle\sum_{ij} \tfrac{1}{g_{ij}} \left[F(v_{ij}^{n+1}) - F(v_{ij}^{n+\frac{1}{2}})\right] v_{ij}^{n+1} \Delta x \Delta y \\
&\leq C_1\alpha\tau - \tfrac{C_2}{\alpha} \textstyle\sum_{ij} \tfrac{1}{g_{ij}} \left[B(v_{ij}^{n+1}) - B(v_{ij}^{n+\frac{1}{2}})\right] \Delta x \Delta y \\
&= C_1\alpha\tau - \tfrac{C_2}{\alpha} \textstyle\sum_{ij} \tfrac{1}{g_{ij}} \displaystyle\int_{\Omega_{ij}} \left[B(v_{ij}^{n+1}) - B(v_{ij}^{n+\frac{1}{2}})\right] \mathrm{d}x\mathrm{d}y \\
&= C_1\alpha\tau - \tfrac{C_2}{\alpha} \textstyle\sum_{ij} \tfrac{1}{g_{ij}} \displaystyle\int_{\Omega_{ij}} \left[B(v_{ij}^{n+1}) - B(v_{ij}^{n})\right] \mathrm{d}x\mathrm{d}y \\
&\quad + \tfrac{C_2}{\alpha} \textstyle\sum_{ij} \tfrac{1}{g_{ij}} \displaystyle\int_{\Omega_{ij}} \left[B(\widetilde{v}^{n+\frac{1}{2}}(x, y)) - B(v_{ij}^{n})\right] \mathrm{d}x\mathrm{d}y \\
&\quad + \tfrac{C_2}{\alpha} \textstyle\sum_{ij} \tfrac{1}{g_{ij}} \displaystyle\int_{\Omega_{ij}} \left[B(v_{ij}^{n+\frac{1}{2}}) - B(\widetilde{v}^{n+\frac{1}{2}}(x, y))\right] \mathrm{d}x\mathrm{d}y
\end{aligned}$$

$$
\begin{aligned}
&\leq C_1\alpha\tau - \tfrac{C_2}{\alpha}\textstyle\sum_{ij}\frac{1}{g_{ij}}\displaystyle\int_{\Omega_{ij}}\left[B(v_{ij}^{n+1}) - B(v_{ij}^{n})\right]\mathrm{d}x\mathrm{d}y\\
&\quad+\tfrac{C_2}{\alpha}\textstyle\sum_{ij}\frac{1}{g_{ij}}\displaystyle\int_{\Omega_{ij}}\left[F(\widetilde{v}^{n+\frac{1}{2}}(x,y)) - F(v_{ij}^{n})\right][\widetilde{v}^{n+\frac{1}{2}}(x,y)]\mathrm{d}x\mathrm{d}y\\
&\quad+\tfrac{C_2}{\alpha}\textstyle\sum_{ij}\frac{v_{ij}^{n+\frac{1}{2}}}{g_{ij}}\displaystyle\int_{\Omega_{ij}}\left[F(v_{ij}^{n+\frac{1}{2}}) - F(\widetilde{v}^{n+\frac{1}{2}}(x,y))\right]\mathrm{d}x\mathrm{d}y.
\end{aligned}
$$

The last term at the rhs is zero, as this is exactly the projection (2.81). The third term can be estimated by using $F(\|v^n\|_\infty) = \|F(v^n)\|_\infty \leq \|F(v^0)\|_\infty = F(\|v^0\|_\infty)$ so that $\|v^n\|_\infty \leq \|v^0\|_\infty = C\|F(v^0)\|_\infty$. Moreover, invoking Lemma 2.5.3 (Lipschitz continuity in time of the transport) and $g > \epsilon$ we find

$$
\begin{aligned}
\tau TV(v^{n+1}) \leq \left(C_1\alpha + \tfrac{C(\|F(v^0)\|_\infty)}{\alpha}\right)\tau&\\
- \tfrac{C_2}{\alpha}\textstyle\sum_{ij}\displaystyle\int_{\Omega_{ij}}\frac{1}{g}\left[B(v_{ij}^{n+1}) - B(v_{ij}^{n})\right]\mathrm{d}x\mathrm{d}y.&
\end{aligned}
$$

Next, by summing on n –fixing α arbitrarely– we may arrive at

$$
\begin{aligned}
\sum_{i=1}^{n}\tau TV(v^i) &\leq C(\|F(v^0)\|_\infty)T - C_2\textstyle\sum_{ij}\frac{1}{g_{ij}}\left[B(v_{ij}^{n}) - B(v_{ij}^{0})\right]\Delta x\Delta y\\
&\leq C(\|F(v^0)\|_\infty)T + C_2\textstyle\sum_{ij}\frac{1}{g_{ij}}B(v_{ij}^{0})\Delta x\Delta y\\
&\leq C(\|F(v^0)\|_\infty)T + C_2\tfrac{\|F(v^0)\|_\infty\|v^0\|_\infty}{\epsilon}\|\Omega\| \leq C,
\end{aligned}
$$

where we still used the fact that $0 \leq \frac{1}{g}B(v) \leq \frac{1}{g}\|F(v^0)\|_\infty\|v^0\|_\infty$. This is the required result. □

The TV boundedness property of Lemma 2.5.13 can be rephrased as follows.

Lemma 2.5.14. *The approximation scheme given by (2.108) satisfies*

$$
\int_0^T\int_\Omega |v_{\Delta t}(x + k\Delta x, y + l\Delta y, t) - v_{\Delta t}(x,y,t)|\,\mathrm{d}x\mathrm{d}y\mathrm{d}t \leq C(l\Delta x + k\Delta y), \tag{2.110}
$$

where C only depends on the domain, on a and b, and on $\|v^0\|_\infty$ and $\|v_0\|_\infty$. Moreover, $v_{\Delta t}(x,y,t) = v^n(x,y)$ for $t \in (t_{n-1}, t_n)$ is a piecewise constant function in (x,y).

We estimate the total variation in t. We take $t \in (t_n, t_{n+1})$. For the hyperbolic and the projection step this result has already been obtained. It remains to consider the parabolic part.

Lemma 2.5.15. *The approximation scheme given by (2.108) satisfies*

$$\int_0^T \int_\Omega |v_{\Delta t}(x,y,t+k\tau) - v_{\Delta t}(x,y,t)| \,\mathrm{d}x\mathrm{d}y\mathrm{d}t \leq C\sqrt{k\tau}, \tag{2.111}$$

where C only depends on the domain, on a and b and on $\|v^0\|_\infty$ and $\|v_0\|_\infty$. Moreover, $v_{\Delta t}(x,y,t) = v^n(x,y)$ for $t \in (t_{n-1}, t_n)$ is a piecewise constant function in (x,y).

Proof. We consider a smooth function ϕ, and its piecewise constant approximation $\phi_h \equiv \phi_{\Delta x, \Delta t} = \phi_{ij}$, for $(x,y) \in \Omega_{ij}$. We multiply both sides of (2.103) by ϕ_{ij} and sum up over i and j. Using the notation (2.108) and applying (A.2), we get

$$\begin{aligned}
&\sum_{ij} \int_{\Omega_{ij}} \frac{1}{g_{ij}} \left[F(v_{ij}^{n+1}) - F(v_{ij}^{n+\frac{1}{2}})\right] \phi_h \mathrm{d}x\mathrm{d}y \\
&\quad \leq \tau \left| \sum_{ij} \left[a_{ij}[D_{ij}^i v^{n+1}] D_{ij}^i \phi_h + b_{ij}[D_{ij}^j v^{n+1}] D_{ij}^j \phi_h\right] \Delta x \Delta y \right| \\
&\quad \leq \tau C \max\left(\|D^x \phi_h\|_\infty, \|D^y \phi_h\|_\infty\right) TV(v^{n+1}) \\
&\quad \leq C\tau \|\nabla \phi_h\|_\infty TV(v^{n+1}) \\
&\quad \leq C\tau \|\nabla \phi\|_\infty TV(v^{n+1}).
\end{aligned} \tag{2.112}$$

Here, we used the estimate $\|F(v^{n+1})\|_\infty \leq \|F(v^0)\|_\infty \leq C$, the properties of a and b, and the fact that $\phi_h \to \phi$ for $\tau \to 0$, dropping higher order terms in τ.

We establish weak Lipschitz continuity in time similarly as in [27] (with a technique due to Kružkov). Due to the strong Lipschitz continuity of the hyperbolic operator and the properties of the projection, (2.81), we have that

$$\sum_{ij} \int_{\Omega_{ij}} \frac{1}{g_{ij}} \left[F(v_{ij}^{n+\frac{1}{2}}) - F(\widetilde{v}^{n+\frac{1}{2}}(x,y))\right] \phi_h \mathrm{d}x\mathrm{d}y = 0, \tag{2.113}$$

and

$$\sum_{ij} \int_{\Omega_{ij}} \frac{1}{g_{ij}} \left[F(\widetilde{v}^{n+\frac{1}{2}}(x,y)) - F(v_{ij}^n)\right] \phi_h \mathrm{d}x\mathrm{d}y \leq C\|\phi\|_\infty \tau. \tag{2.114}$$

Combining the three steps (2.112), (2.113), (2.114), and repeating the argument for time steps $n, n+1, \dots, n+k$, we find,

$$\sum_{ij} \int_{\Omega_{ij}} \frac{1}{g_{ij}} \left[F(v_{ij}^{n+k}) - F(v_{ij}^{n}) \right] \phi_h \mathrm{d}x\mathrm{d}y \leq C\|\phi\|_\infty k\tau + C\|\nabla\phi\|_\infty \tau \sum_{i=1}^{k} TV(v^{n+i}).$$

Multiplying by τ, summing on n, we get

$$\int_0^T \int_\Omega \frac{1}{g} \left[F(v_{\Delta t}(x,y,t+k\tau)) - F(v_{\Delta t}(x,y,t)) \right] \phi \mathrm{d}x\mathrm{d}y\mathrm{d}t \leq C(\|\phi\|_\infty + \|\nabla\phi\|_\infty) k\tau,$$

by Lemma 2.5.13. Changing the summation by integrals, errors are introduced due to numerical differentiation (e.g. g_{ij} replaced by $g(x,y)$). However, these go to zero as $\Delta x \to 0$. Now, let ω_h be a smooth mollifier with support in $[-h,h]^3$, define

$$\psi = \mathrm{sgn}(v_{\Delta t}(x,y,t+k\tau) - v_{\Delta t}(x,y,t)),$$

and set $\phi = \omega_h(x,y,t) * \psi$. Standard arguments based upon the properties of ω_h allow us to write, see [33, 46]

$$\int_0^T \int_\Omega \frac{1}{g} \left| F(v_{\Delta t}(x,y,t+k\tau)) - F(v_{\Delta t}(x,y,t)) \right| \mathrm{d}x\mathrm{d}y\mathrm{d}t \leq C(h + \frac{1}{h}) k\tau \leq C\sqrt{k\tau}, \tag{2.115}$$

where $h = \sqrt{k\tau}$. Recall that $v_{\Delta t}$ is bounded. Use the fact that $TV(v) \leq C\,TV(F(v))$, $(F'(v) \geq \frac{1}{C})$, and notice that $g < C$. We conclude that

$$\int_0^T \int_\Omega \left| v_{\Delta t}(x,y,t+k\tau) - v_{\Delta t}(x,y,t) \right| \mathrm{d}x\mathrm{d}y\mathrm{d}t \leq C\sqrt{k\tau}. \tag{2.116}$$

□

2.5.7 Existence

The convergence of the sequence $(v_{\Delta t}(x,y,t))$ for $\Delta t \to 0$ follows from the Riesz-Fréchet-Kolmogorov theorem, see Appendix A, Theorem A.2.2. With this Theorem we can state:

Lemma 2.5.16. *If $\Delta t \to 0$, then there exists a subsequence $v_{\nu_j}(x,y,t)$ of the sequence $v_{\Delta t}(x,y,t)$ such that $v_{\nu_j} \to v$ for $j \to \infty$ in $L_{1,\mathrm{loc}}(\Omega \times I)$, $\Omega \times I = (x^{(1)}, x^{(2)}) \times (y^{(1)}, y^{(2)}) \times (0,T)$.*

Proof. Lemma 2.5.13 implies that $v_{\Delta t}(x,y,t)$ is uniformly bounded. From Lemma 2.5.14 and 2.5.15 it follows that

$$\int_0^T \int_\Omega |v_{\Delta t}(x+k\Delta x, y+l\Delta y, t+m\Delta t) - v_{\Delta t}(x,y,t)|\,\mathrm{d}x\mathrm{d}y\mathrm{d}t \\ \leq C(k\Delta x + l\Delta y + \sqrt{m\Delta t}),$$

Thus the condition of the compactness criterion in the Riesz-Fréchet-Kolmogorov theorem is satisfied. Consequently, there exists a subsequence $v_{\nu_j}(x,y,t)$ that converges to some $v(x,y,t)$ in $L_{1,\mathrm{loc}}(\Omega \times I)$. □

Remark 2.5.9. *We emphasize the fact that the convergence is in $L_{1,\mathrm{loc}}(\Omega \times I)$, so nothing can be said on the value of v on the boundary. The weak formulation (2.75) does not need the value of v on the boundary, as only v_0 and v^0, given functions, appear in the boundary terms of the weak formulation. Therefore, this weak formulation is consistent with our approach for proving convergence.*

Before continuing, we need a more elaborated function than $v_{\Delta t}$, which is piecewise continuous in time. We define $v_\nu(x,y,t)$ as

$$v_\nu(x,y,t) = \begin{cases} \mathcal{T}_{\delta,\Delta y}(2(t-t_k))v^k(x,y), & t \in [t_k, t_{k+1/2}) \\ \mathcal{D}_{\Delta xy,\Delta t}(2(t-t_{k+1/2}))v^{k+1/2}(x,y), & t \in [t_{k+1/2}, t_{k+1}) \end{cases}$$

where v^k is the solution obtained at time t_k and $v^{k+1/2} = \pi\mathcal{T}_{\delta,\Delta y}(2(t_{k+1/2} - t_k))v^k(x,y) = \pi\widetilde{v}^{k+\frac{1}{2}}(x,y)$, with $k = 0,\ldots,N-1$. Furthermore, $t_{k+1/2} = (t_{k+1} - t_k)/2$. We write $\tau_k = t_{k+1} - t_k$.

All the results obtained for $v_{\Delta t}$ are also valid for v_ν. We now prove Theorem 2.5.1, the convergence of the solution obtained by our operator splitting method to the very weak solution (2.76) of (2.73).

Proof of Theorem 2.5.1. Lemma 2.5.16 claims that the sequence $\{v_\nu\}_{\Delta_t > 0}$ converges to some $v(x,y,t)$. To complete the convergence proof for the splitting procedure, it is now sufficient to show that this limit is the very weak solution (2.76) of problem (2.73).

Let us consider test functions $\phi(x,y,t) \in C^\infty(\Omega \times I)$, with compact support near the outflow boundary. At the inflow boundary, $y = y^{(2)}$, we impose $\partial_y \phi\big|_{y=y^{(2)}} = 0$ and at the no-flow boundary we impose $\partial_x \phi\big|_{x=x^{(1)}} = 0 = \partial_x \phi\big|_{x=x^{(2)}}$. Furthermore, we require $\phi(x,y,T) = 0$. The variational formulation is then given by (2.76). We need to show that the limit function $v(x,y,t)$ satisfies (2.76). We use the ideas from [16, 27].

We begin with the transport part for $t \in (t_k, t_{k+\frac{1}{2}})$, and consider the new variable $z = 2(t - t_k)$, and the accompanying transformation of the test function $\overline{\phi}(x, y, z) = \phi(x, y, \frac{z}{2} + t_k)$. Write formally $v_\nu(u, v, t) = v^k_{\mathcal{T}}(2(t - t_k))$, where $v^k_{\mathcal{T}}(t) = \mathcal{T}_{\delta,\Delta y}(t) v^k(x, y)$. In the considered time interval, v_ν is the exact solution of the transport problem (2.77) with initial condition the piecewise constant function v^k, and with inflow condition $v(x, y^{(2)}, t) = v^0(x, y^{(2)}, t)$. Therefore, we can write

$$
\begin{aligned}
&\int_\Omega \int_{t_k}^{t_{k+\frac{1}{2}}} \left(\tfrac{1}{2} \frac{F(v_\nu)}{g} \partial_t \phi - v_\nu \partial_y \left(h\phi \right) \right) d\Omega\, dt \\
&\quad = \int_\Omega \int_0^{\tau_k} \left(\frac{F\left(v^k_{\mathcal{T}}(z) \right)}{g} \partial_z \overline{\phi} - v^k_{\mathcal{T}}(z) \partial_y \left(h\overline{\phi} \right) \right) d\Omega\, \tfrac{1}{2} dz \\
&\quad = \tfrac{1}{2} \int_\Omega \frac{F\left(v^k_{\mathcal{T}}(z) \right)}{g} \overline{\phi} \Big|_{z=0}^{z=\tau_k} d\Omega - \tfrac{1}{2} \int_0^{\tau_k} \int_{x^{(1)}}^{x^{(2)}} h v^k_{\mathcal{T}}(z) \overline{\phi} \Big|_{y=y^{(1)}}^{y=y^{(2)}} dz \\
&\quad = \tfrac{1}{2} \int_\Omega \frac{F\left(\widetilde{v}^{k+\frac{1}{2}} \right)}{g} \phi(t_{k+\frac{1}{2}})\, d\Omega - \tfrac{1}{2} \int_\Omega \frac{F\left(v_\nu(t_k) \right)}{g} \phi(t_k)\, d\Omega \\
&\quad -\tfrac{1}{2} \int_{t_k}^{t_{k+1}} \int_{x^{(1)}}^{x^{(2)}} h v^0(x, y^{(2)}, \tilde{t}) \phi(\frac{\widetilde{t} - t_k}{2} + t_k)\, dx\, d\widetilde{t}, \\
&\quad = \tfrac{1}{2} \int_\Omega \frac{F\left(\widetilde{v}^{k+\frac{1}{2}} \right)}{g} \phi(t_{k+\frac{1}{2}})\, d\Omega - \tfrac{1}{2} \int_\Omega \frac{F\left(v_\nu(t_k) \right)}{g} \phi(t_k)\, d\Omega \\
&\quad -\tfrac{1}{2} \int_{t_k}^{t_{k+1}} \int_{x^{(1)}}^{x^{(2)}} h v^0(x, y^{(2)}, \tilde{t}) \phi(\tilde{t})\, dx\, d\widetilde{t} + \mathcal{O}(\Delta t^2). \qquad (2.117)
\end{aligned}
$$

For the last equality, we used $\widetilde{t} = 2(t - t_k) + t_k$, and $\phi \in C^1(\Omega \times I)$ so $\phi(\frac{\tilde{t} - t_k}{2} + t_k) = \phi(\tilde{t}) + \mathcal{O}(\Delta t)$ for $\tilde{t} \in (t_k, t_{k+1})$. The error goes to zero, even after summation over k (i.e. $\sum_k (\Delta t)^2 \to 0$), so we can drop the error term.

We now turn our attention to the diffusion part over the time interval $(t_{k+\frac{1}{2}}, t_k)$ with initial condition $v^{k+1/2}$. This corresponds to scheme (2.103). For simplicity set $\Delta x^+ = \Delta x^- = \Delta x$ and $\Delta y^+ = \Delta y^- = \Delta y$. Multiply both sides of (2.103) with $\phi_{ij} = \phi(x_i, y_j, t_k)$, and sum over i and j. Using the standard notation (2.108) and $\phi_{ij}^{k+1} = \phi(x_i, y_j, t_{k+1})$, we get that

$$
\begin{aligned}
I := \sum_{i=0}^{N_1}\sum_{j=0}^{N_2} \Bigg[& \frac{\Delta x \Delta y}{g_{ij}} \frac{F(v_{ij}) - F(v_{ij}^{k+\frac{1}{2}})}{\tau_k}\phi_{ij} \\
& - \phi_{ij}\left(a_{i+1,j}\frac{v_{i+1,j} - v_{ij}}{\Delta x}\Delta y - a_{i,j}\frac{v_{i,j} - v_{i-1,j}}{\Delta x}\Delta y \right) \\
& \quad - \phi_{ij}\left(b_{i,j+1}\frac{v_{i,j+1} - v_{ij}}{\Delta x}\Delta x - b_{i,j}\frac{v_{i,j} - v_{i,j-1}}{\Delta y}\Delta x \right) \Bigg] = 0.
\end{aligned}
$$

We rearrange the first term, and apply Abel's summation (A.2) on the last two terms. This allows to write

$$
\begin{aligned}
I = \sum_{i=0}^{N_1}\sum_{j=0}^{N_2} & \frac{\Delta x \Delta y}{g_{ij}} \left(-\frac{\phi_{ij}^{k+1} - \phi_{ij}}{\tau_k} F(v_{ij}) + \frac{F(v_{ij})\phi_{ij}^{k+1}}{\tau_k} - \frac{F(v_{ij}^{k+\frac{1}{2}})\phi_{ij}}{\tau_k} \right) \\
& + \sum_{i=1}^{N_1}\sum_{j=0}^{N_2} \frac{\phi_{ij} - \phi_{i-1,j}}{\Delta x} a_{i,j} \frac{v_{i,j} - v_{i-1,j}}{\Delta x} \Delta x \Delta y \\
& \quad + \sum_{i=0}^{N_1}\sum_{j=1}^{N_2} \frac{\phi_{ij} - \phi_{i,j-1}}{\Delta y} b_{i,j} \frac{v_{i,j} - v_{i,j-1}}{\Delta y} \Delta x \Delta y = 0, \qquad (2.118)
\end{aligned}
$$

where we used $a_{0,j} = a_{N_1+1,j} = 0 = b_{i,0} = b_{i,N_2+1}$ because of the homogeneous Neumann boundary condition. We again apply Abel's summation on the second and third double sum of (2.118), to obtain

$$
\begin{aligned}
I = \sum_{i=0}^{N_1}\sum_{j=0}^{N_2} \frac{\Delta x \Delta y}{g_{ij}} &\left(-\frac{\phi_{ij}^{k+1}-\phi_{ij}}{\tau_k} F(v_{ij}) + \frac{F(v_{ij})\phi_{ij}^{k+1}}{\tau_k} - \frac{F(v_{ij}^{k+\frac{1}{2}})\phi_{ij}}{\tau_k} \right) \\
&- \sum_{i=2}^{N_1}\sum_{j=0}^{N_2} \left(a_{i,j}\frac{\phi_{ij}-\phi_{i-1,j}}{\Delta x} - a_{i-1,j}\frac{\phi_{i-1,j}-\phi_{i-2,j}}{\Delta x} \right) \frac{v_{i-1,j}}{\Delta x}\Delta x \Delta y \\
&- \sum_{i=0}^{N_1}\sum_{j=2}^{N_2} \left(b_{i,j}\frac{\phi_{ij}-\phi_{i,j-1}}{\Delta y} - b_{i,j-1}\frac{\phi_{i,j-1}-\phi_{i,j-2}}{\Delta y} \right) \frac{v_{i,j-1}}{\Delta y}\Delta x \Delta y \\
&+ \sum_{j=0}^{N_2} a_{N_1,j}\frac{\phi_{N_1,j}-\phi_{N_1-1,j}}{\Delta x}\frac{v_{N_1,j}}{\Delta x} - \sum_{j=0}^{N_2} a_{1,j}\frac{\phi_{1,j}-\phi_{0,j}}{\Delta x}\frac{v_{0,j}}{\Delta x} \\
&+ \sum_{i=0}^{N_1} b_{i,N_2}\frac{\phi_{i,N_2}-\phi_{i,N_2-1}}{\Delta y}\frac{v_{i,N_2}}{\Delta y} - \sum_{i=0}^{N_1} b_{i,1}\frac{\phi_{i,1}-\phi_{i,0}}{\Delta y}\frac{v_{i,0}}{\Delta y} = 0. \qquad (2.119)
\end{aligned}
$$

The four single sums contain values of our solution v on the boundary. These terms are all zero for sufficiently small Δx, Δy. The b_{i,N_2} term vanishes because of $\partial_y \phi = 0$ on the outflow boundary because of the compact support there; the other 3 terms vanish because of the boundary conditions imposed on ϕ at the other boundaries. We therefore have

$$
\begin{aligned}
I = \sum_{i=0}^{N_1}\sum_{j=0}^{N_2} \frac{\Delta x \Delta y}{g_{ij}} &\left(-\frac{\phi_{ij}^{k+1}-\phi_{ij}}{\tau_k} F(v_{ij}) + \frac{F(v_{ij})\phi_{ij}^{k+1}}{\tau_k} - \frac{F(v_{ij}^{k+\frac{1}{2}})\phi_{ij}}{\tau_k} \right) \\
&- \sum_{i=1}^{N_1-1}\sum_{j=0}^{N_2} \left(a_{i+1,j}\frac{\phi_{i+1,j}-\phi_{i,j}}{\Delta x} - a_{i,j}\frac{\phi_{i,j}-\phi_{i-1,j}}{\Delta x} \right) \frac{v_{i,j}}{\Delta x}\Delta x \Delta y \\
- \sum_{i=0}^{N_1}\sum_{j=1}^{N_2-1} &\left(b_{i,j+1}\frac{\phi_{i,j+1}-\phi_{i,j}}{\Delta y} - b_{i,j}\frac{\phi_{i,j}-\phi_{i,j-1}}{\Delta y} \right) \frac{v_{i,j}}{\Delta y}\Delta x \Delta y = 0. \qquad (2.120)
\end{aligned}
$$

By reordening terms, multiplying by τ_k, writing formally $v_\nu(u,v,t) = v_{\mathcal{D}}^k(2(t-t_k))$, where $v_{\mathcal{D}}^k(t) = \mathcal{D}_{\delta,\Delta x,\Delta y}(t)v^{k+\frac{1}{2}}(x,y)$, (2.120) can be seen as an approximation of the equality

$$\int_\Omega \int_0^{\tau_k} \left[\frac{F(v_{\mathcal{D}}^k(z))}{g} \partial_z \overline{\phi}(z) + v_{\mathcal{D}}^k(z) \left(\partial_x a \partial_x \overline{\phi}(z) - \partial_y b \partial_y \overline{\phi}(z) \right) \right] d\Omega dz$$
$$= \int_\Omega \frac{F(v_{\mathcal{D}}^k(\tau_k))}{g} \overline{\phi}(\tau_k)\, d\Omega - \int_\Omega \frac{F(v_{\mathcal{D}}^k(0))}{g} \overline{\phi}(0)\, d\Omega, \quad (2.121)$$

By passing from (2.120) to (2.121), errors appear that are due to numerical differentiation and integration in time and space. However, these errors go to zero as $\Delta t \to 0$, also after summation on k. Therefore, we need not consider them further. As a last step we rewrite (2.121) as

$$\int_\Omega \int_{t_{k+\frac{1}{2}}}^{t_{k+1}} \left[\tfrac{1}{2} \frac{F(v_\nu(t))}{g} \partial_t \phi(t) + v_\nu(t) \left(\partial_x a \partial_x \phi(t) + \partial_y b \partial_y \phi(t) \right) \right] d\Omega dt$$
$$= \tfrac{1}{2} \int_\Omega \frac{F(v_\nu(t_{k+1}))}{g} \phi(t_{k+1})\, d\Omega - \tfrac{1}{2} \int_\Omega \frac{F(v^{k+\frac{1}{2}}))}{g} \phi(t_{k+\frac{1}{2}})\, d\Omega. \quad (2.122)$$

Combining the two results by adding (2.117) and (2.122) for $k = 0, \ldots, N-1$, we arrive at

$$\int_0^T \int_\Omega \left(\tfrac{1}{2} \frac{F(v_\nu)}{g} \partial_t \phi - \chi_{\mathcal{T}}(t) v_\nu \partial_y (h\phi) \right.$$
$$\left. + \chi_{\mathcal{D}}(t) v_\nu(t) \left[\partial_x a \partial_x \phi(t) + \partial_y b \partial_y \phi(t) \right] \right) d\Omega\, dt$$
$$= \tfrac{1}{2} \int_\Omega \frac{F(v_\nu(T))}{g} \phi(T)\, d\Omega - \tfrac{1}{2} \int_\Omega \frac{F(v_\nu(0))}{g} \phi(0)\, d\Omega$$
$$+ \tfrac{1}{2} \sum_{k=0}^{n-1} \int_\Omega \left[F(\widetilde{v}^{k+\frac{1}{2}}) - F(v^{k+\frac{1}{2}}) \right] \frac{\phi(t_{k+\frac{1}{2}})}{g}\, d\Omega$$
$$- \tfrac{1}{2} \int_0^T \int_{x^{(1)}}^{x^{(2)}} h v^0(x, y^{(2)}, t) \phi(t)\, dx\, dt. \quad (2.123)$$

Here, $\chi_{\mathcal{T}}(t)$ and $\chi_{\mathcal{D}}(t)$ are characteristic functions defined as

$$\chi_{\mathcal{T}}(t) = \begin{cases} 1 & \text{for } t \in \cup_k [t_k, t_{k+1/2}) \\ 0 & \text{otherwise} \end{cases}, \quad \chi_{\mathcal{D}}(t) = \begin{cases} 1 & \text{for } t \in \cup_k [t_{k+1/2}, t_{k+1}) \\ 0 & \text{otherwise.} \end{cases},$$

We have ([16, 27])

$$\chi_{\mathcal{T}}(t) \text{ and } \chi_{\mathcal{D}}(t) \rightharpoonup \frac{1}{2} \text{ in } L_2(0,T) \quad \text{for} \quad \Delta t \to 0$$

Recall further that the test function ϕ was chosen so as to satisfy $\phi(T) = 0$. Moreover, for $\Delta t \to 0$ ($n \to \infty$), the projection error represented by the third term on the rhs of (2.123), tends to zero. This property follows from

$$\begin{aligned}
&\tfrac{1}{2}\sum_{k=0}^{N-1} \int_\Omega \left[F(\widetilde{v}^{k+\frac{1}{2}}) - F(v^{k+\frac{1}{2}})\right] \frac{\phi(t_{k+\frac{1}{2}})}{g}\, d\Omega \\
= \;& \tfrac{1}{2}\sum_{k=0}^{N-1}\sum_{ij} \int_{\Omega_{ij}} \left[F(\widetilde{v}^{k+\frac{1}{2}}) - F(v_{ij}^{k+\frac{1}{2}})\right] \frac{\phi_{ij}(t_{k+\frac{1}{2}})}{g_{ij}}\, d\Omega \\
&+\tfrac{1}{2}\sum_{k=0}^{N-1}\sum_{ij} \int_{\Omega_{ij}} \left[F(\widetilde{v}^{k+\frac{1}{2}}) - F(v_{ij}^{k+\frac{1}{2}})\right] \left[\frac{\phi(t_{k+\frac{1}{2}})}{g} - \frac{\phi_{ij}(t_{k+\frac{1}{2}})}{g_{ij}}\right] d\Omega \\
= \;& I_1 + I_2.
\end{aligned}$$

By the definition of the projection, (2.81), we have that $I_1 \equiv 0$, $\forall k$. For I_2 we can use the smoothness of ϕ and g, $g > \epsilon$, Lemma 2.5.5 and Lemma 2.5.13, to obtain

$$\begin{aligned}
I_2 \;&\leq\; \tfrac{1}{2}\sum_{k=0}^{N-1} \int_\Omega \left|F(\widetilde{v}^{k+\frac{1}{2}}) - F(v_{ij}^{k+\frac{1}{2}})\right| \left[\frac{\|\nabla\phi\|_\infty}{\epsilon} + \frac{\|\phi\|_\infty \|\nabla g\|_\infty}{\epsilon^2}\right] \Delta x\, d\Omega \\
&\leq\; \tfrac{1}{2}\sum_{k=0}^{N-1} C\Delta t\, TV(F(\widetilde{v}^{k+\frac{1}{2}}))\Delta x \leq C\Delta x.
\end{aligned}$$

We now pass to the limit $\Delta t \to 0$ in (2.123), with $\Delta t = C\Delta x$. Taking into account the convergence of v_ν to v in $L_{1,\mathrm{loc}}(\Omega \times I)$, we finally obtain that the limit function v satisfies (2.76). □

Remark 2.5.10 (Uniqueness). *Having proved convergence of our approximation scheme to a very weak solution of (2.73), we may address the question of uniqueness. We can refer to other works to partly answer this question (like [27, 61]). For our problem, uniqueness of a one-dimensional version of (2.73), however with other BCs, has been proved in [44]. Some work has also been done in [42]. Under a transformation $w = F(v)$, we can also refer to [34], at least when $\Omega = \mathbb{R}^d$.*

Remark 2.5.11 (Lipschitz continuity and cut-off). *At the beginning of this Section, we noted that we need to regularize g by a function g_ϵ, everywhere positive. So the obtained very weak solution is in fact the very weak solution of this regularized problem, (v_ϵ). According to [38] there exists a subsequence $\{\epsilon_j\}_{j>0}$ such that (v_{ϵ_j}) is weakly convergent in appropriate function spaces and the limit is the very weak solution of the original problem (2.73). We further required the Lipschitz continuity of F^{-1}. If this condition is not fulfilled we may pass to a regularization (eg. the case of Freundlich adsorption $p > 1$, where $F'(0) = 0$). Again, a convergent subsequence can be found, converging to the very weak solution of the original problem, [38].*

2.5.8 Boundary conditions and operator splitting

It is important to comment on the relation between boundary conditions and the operator splitting method. As far as we know, boundary conditions are rarely considered in previous works for convection-diffusion equations and operator splitting: the authors limit themselves to the Cauchy problem. The reason for this restriction is the local convergence obtained with Riesz-Fréchet-Kolmogorov compactness theorem. This does not allow to consider the limit function on the boundary. Hence, Dirichlet and Neuman boundary conditions, which are commonly considered in convection-diffusion problems, are difficult to be handled.

Most problem in the literature are purely hyperbolic or are degenerate parabolic equations, requiring the introduction of entropy solutions. This complicates the inclusion of boundary conditions. For an overview of the difficulties with Dirichlet BCs and hyperbolic problems, see [12]. The difficulties concern the fact that the characteristics may intersect $\partial\Omega$ from the interior, such that the boundary condition does not hold pointwise for all times. This argument has led us to consider different types of boundary conditions on the inflow boundary and on the outflow boundary.

Before continuing, we mention that splitting methods are also used in combination with reduction to ODEs, allowing the use of solution methods for ODEs (Method of Lines). Much work has been done to incorporate boundary conditions in this setting with specific boundary correction techniques, see e.g. [29].

In our case, (2.73) is a parabolic equation, and the weak formulation can be readily obtained. We have a known, fixed velocity field, allowing the identification of no-flux boundaries, inflow and outflow boundaries. In a general 2-D setting this is no longer the case. If one considers for instance the equation $\partial_t u + \nabla f(u) - \Delta u = 0$, imposing a no-flow boundary condition is more complicated. Therefore, an operator splitting method is not optimal for general types of boundary value problems. In our case, the velocity field and flux function is such that flow boundary conditions (inflow/outflow/no-flow) are satisfied independent of the value of u.

As seen, some special boundary conditions can be considered when a parabolic equation like (2.73) is solved by operator splitting of the transport and diffusion part. The first task is the inclusion of the initial condition in the weak formulation. Moreover, we have proved that an advective influx boundary condition can also be considered. We can state the following:

Proposition 2.5.1. *A convection-diffusion problem solved by operator splitting of the transport and diffusion part, with a Diriclet BC over the inflow boundary during transport, and a homogeneous Neumann BC during the diffusion, corresponds to a convection-diffusion problem with an advective influx BC (mixed type) over the inflow boundary.*

This follows directly from the proof of Theorem 2.5.1.

Remark 2.5.12. *Operator splitting is mainly used to overcome the difficulties due to the dominant convection. This dominant convection over the inflow boundary implies two things. First, the difference between a Dirichlet BC and the advective influx BC will be small as the diffusion part is much smaller than the convective part. Secondly, dominant convection over the inflow boundary will make it physically very difficult to implement a Dirichlet BC in a real experiment. Some feedback mechanism is needed, such as equilibrium desorption to garantee the Dirichlet BC. For flow problems, an advective influx BC appears to be most appropriate.*

Remark 2.5.13. *The above consideration implies that, although the diffusion step has homogeneous Neumann BC and it is the last step of the operator splitting, the resulting limit function will not obey a homogeneous Neumann BC, but is the weak solution with advective influx BC. This is consistent with the fact that we have only convergence in $L_{1,loc}$-sense.*

Remark 2.5.14. *The advective influx BC corresponds in a pure hyperbolic problem to a Dirichlet BC ($a = 0 = b$). Thus, one is led to introducing boundary entropy conditions. This can be avoided in our approach. However, the solution of the transport part does satisfy not only (2.77) but also entropy conditions, see [12, 32].*

The second type of BC that does not give problems is a Neumann boundary condition over a no-flow boundary.

Proposition 2.5.2. *A Neumann boundary condition over a no-flow boundary implies no flux during the transport part, and the Neumann BC during the diffusion part.*

The homogeneous Neumann boundary condition for a hyperbolic PDE in a general setting has been considered in detail in [12, 32]

It remains to consider the outflow boundary. During transport no boundary condition is considered, while during diffusion a homogeneous Neumann BC is taken. From Remark 2.5.13 we deduce that the limiting function will not satisfy a homogeneous Neumann BC, as is required by the original problem (2.73). We are not concerned with this complication since we consider dominant convection and it seems difficult to physically impose a homogeneous Neumann BC at the outflow. There is no better alternative for the outflow BC in (2.73). A free outflow BC, as is implemented by the operator splitting method, is reasonable, and has the advantage that the outflow is not affecting the solution, which would be the case with a real "no-flow" boundary condition. In fact, the original BC has limited sense as it conflicts with the physical meaning of the extraction well. However, the BC can be interpreted as follows: roughly speaking $b(x,y)\partial_y v + h(x,y)v = h(x,y)v_{\text{out}}$, and the convection is dominant, $h >> b$. Therefore, $v \approx v_{\text{out}}$, hence $\partial_y v \approx 0$.

Dirichlet BC and non-homogeneous Neumann BC

It is an open problem wether or not a Dirichlet BC or non-homogeneous Neumann BC for a convection-diffusion problem can be implemented with an operator splitting method. The common way to implement a Dirichlet BC would be to require a Dirichlet BC during transport and during diffusion. This will give a convective and a diffusive flux over the inflow. As seen, a non-homogeneous Neumann BC over a part of the boundary where there is no velocity field (no flux for the transport part) can be implemented in the operator splitting by taking this BC only for the diffusion part. During transport no BC is needed over a no-flux boundary.

A Dirichlet BC over a no-flux boundary, or a Neumann BC over a flux-boundary, or a Robin type BC which is not of the advective flux type, would all imply difficulties for an operator splitting approach, just as the handling of these types of BC still give rise to many open problems for hyperbolic problems. One can argue however that these type of BCs are not the most physically reasonable.

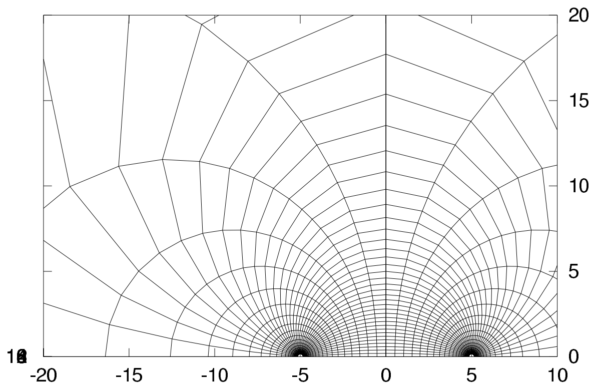


Figure 2.10: Computational grid in Cartesian coordinates for equally spaced, 40×40, (u, v) nodal points.

2.6 Numerical experiments

2.6.1 Without adsorption

For the direct problems, we shall consider a "standard" example with the following defining data: the wells each have a radius of $r_1 = r_2 = 15$cm and their centers are placed 10m from each other ($d = 5$m, $c = 0$). The height of the aquifer is $H = 10$m, the porosity of the soil is $\theta_0 = 0.2$ and the hydraulic conductivity $k = 10^{-5}$m/s $= 0.864$m/day. The longitudinal dispersivity α_L, and its transversal counterpart α_T, as well as the prescribed head value at the extraction well and at the injection well, will be varied in several experiments. We inject the tracer with constant concentration $C^{(2)}(t) = C_0$. For ease of presentation we set $C_0 = 1$. The tracer does not decay ($\mu = 0$) and molecular diffusion is neglected. In this Section we only present data obtained with the flux boundary condition (2.50).

The computational grid, for 40×40 equally spaced nodal points in the transformed (u, v) variables, is plotted (transformed back to Cartesian coordinates) in Figure 2.10.

Experiment 1: convergence

We use the benchmark scheme to investigate the convergence of the standard scheme. For this purpose we calculate the root-mean-square error (RMS) for different grid sizes and time steps. As a reference we use the break through curve (BTC) of the benchmark solution with 80 strips, 800 fixed divisions on the y-axis and 10 moving gridpoints around the shock. The standard example is used with $\alpha_T = 0$, $\alpha_L = 0.1$, $h_1 = 10$ and $h_2 = 20$. The BTC is determined over a period of 18 days after the injection of a step input.

In Table 2.1 the results are given for the benchmark method, and in Table 2.2 for the general method. For the benchmark solution we have better behaviour, and a low error. Taking a larger operator splitting time step increases the error more than decreasing the number of grid points does. Apparently, the operator splitting error is such that adding moving grid points makes it more prominent when the timestep is above a certain treshold, here $\Delta t = 0.1$ days. Adding moving grid points does reduce the error when the splitting time step is less than 0.1 days. It is then possible to use a fixed grid of 200 divisions with low error. Therefore, we use the benchmark method with 80 strips, 200 fixed y-axis divisions, 10 moving gridpoints around the inflow shocks, and a time step of 0.05 days. To make the influence of the moving grid points clearly visible in a picture, we show in Fig. 2.11 a cut-out of the BTC obtained for this example, but with a pulse input of C_0 during 1 day. Moving grid points are used for the beginning and for the end of the shock. The BTC for a 800 grid, with no moving gridpoints or with 20 moving grid points per shock, is hardly distinguishable from the 200 grid with 10 moving gridpoints per shock. On the other hand, the 200 grid with no moving gridpoints, doesn't provide an acceptable approximation.

The standard scheme gives overall larger errors, as can be seen in Table 2.2. This is mainly due to the projection error. For further experiments we shall use the 80×400 grid or 80×200 grid, with time step 0.05 days. Similar results as in Tables 2.1, 2.2 are obtained for a pulse input.

In Table 2.3, we give the RMS difference of the BTC obtained with increasing longitudinal dispersivity in the benchmark or standard scheme, compared to the BTC obtained with the benchmark scheme for $\alpha_L = 0$m. This RMS difference should converge to 0 with decreasing α_L. This is the case for the benchmark method, but for the general method there is a threshold below which we cannot reduce the RMS difference. Again, this is due to the numerical dispersion caused by the projection operator. We conclude that the standard scheme in this example can only be used for a value of the longitudinal dispersivity larger than

strips	$\#y_{div}$	$\#y_{mov}$	Δt (days)	RMS
80	800	0	0.05	0.0001409
80	400	10	0.05	0.0001146
80	400	0	0.05	0.0003805
80	200	10	0.05	0.0003021
80	200	0	0.05	0.0008496
80	100	10	0.05	0.0005780
80	100	0	0.05	0.0017165
80	800	10	0.1	0.0001696
80	800	0	0.1	0.0002189
80	800	10	0.2	0.0004247
80	800	0	0.2	0.0004194
80	800	10	0.4	0.0009222
80	800	0	0.4	0.0009181

Table 2.1: RMS error of the benchmark solution, depending on the number of strips, the number of fixed y-divisions, the number of moving grid points per shock, and the operator splitting time step.

0.001, or $\alpha_L/D > 0.0001$.

Experiment 2: BTC of step input

For the benchmark solution, the transversal dispersivity is neglected ($\alpha_T = 0$). We apply the method in 80 strips, with $y_{div} = 400$. The injected front is tracked with 10 moving grid points, and the time step of the operator splitting is 0.05 days. Furthermore, the head value is $h_1 = 10$m at the extraction well and $h_2 = 15$m at the injection well. The resulting BTCs of the confined flow for 7 different values for α_L/D are plotted in Fig. 2.12. The result is scaled to the same values as used in [76] to allow comparison. A careful analysis of the figures shows that for low values of α_L/D our scheme produces similar BTCs as in [76], but that for higher dispersivities clearly different results are obtained. It is important to point out that the approximation method in [76] is only valid when $\alpha_L/D < 0.1$. Therefore, the last curve that should be compared is the one corresponding to $\alpha_L/D = 0.05$; it is still in excellent agreement. Thus, our results are very reliable.

strips	cells	Δt (days)	RMS
80	800	0.05	0.0005759
80	400	0.05	0.0006244
80	200	0.05	0.0008579
80	100	0.05	0.001490
80	800	0.1	0.001017
80	800	0.2	0.001760
80	800	0.4	0.002993

Table 2.2: RMS error of the standard scheme, depending on the number of strips, the number of cells and on the operator splitting time step.

α_L (m)	RMS BM	RMS
0	0	0.003095
10^{-6}	0.0002567	0.003095
10^{-5}	0.0007136	0.003098
10^{-4}	0.001439	0.003121
10^{-3}	0.002597	0.003935
10^{-2}	0.004281	0.004351

Table 2.3: RMS difference with increasing α_L of the benchmark scheme and of the standard scheme when compared with the benchmark scheme for $\alpha_L = 0$m.

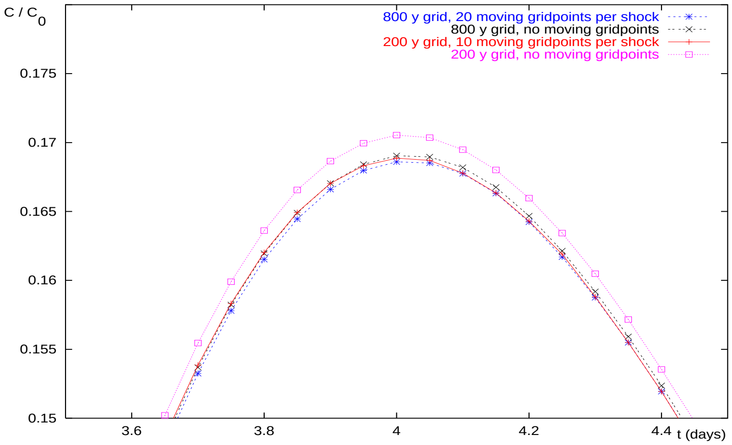


Figure 2.11: Importance of moving grid points in BTC obtained with the benchmark solution. Cut-out of BTC for $\alpha_L = 0.1$m, $\alpha_T = 0.0$m, pulse of 1 day. Top curve: 200 grid, no moving points; bottom curve: 800 grid with 20 moving gridpoints per shock.

Experiment 3: BTC pulse input

We now inject the tracer with a constant concentration $C^{(2)}(t) = 1$ during 1 day, after which we inject zero concentration, $C^{(2)}(t) = 0$. As mentioned, we use the standard method in an 80×400 grid with time step 0.05 days to get equally good results as with the benchmark method. In Fig. 2.13 we again give the scaled BTCs for several α_L/D values.

Experiment 4: influence of confined versus unconfined flow

Our numerical scheme allows for confined, unconfined or partially confined-unconfined flow as long as the Dupuit-Forchheimer approximation is valid. We can investigate the influence of this on the BTC: using the standard method in a 80×400 grid on the standard example with $\alpha_L = 0.2$m, $\alpha_T = 0 = D_0$, timestep 0.05 days and several injection and extraction head values, we obtain Fig. 2.14.

All BTCs are obtained with a head difference $\Delta h = 5$m. For all confined flow, the same BTC is retrieved, as it should. For partially confined-unconfined flow

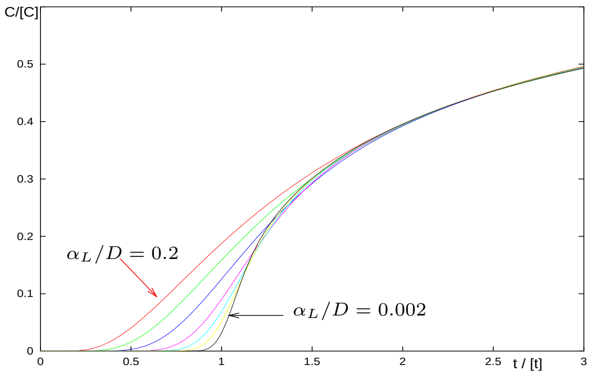


Figure 2.12: BTC with step input for $\alpha_L/D = 0.2$, 0.1, 0.05, 0.02, 0.01, 0.005 and 0.002m scaled to match [76].

($h_2 = 13$m, $h_1 = 8$m) and unconfined flow ($h_2 = 10$m, $h_1 = 5$m and $h_2 = 6$m, $h_1 = 1$m), clearly different BTCs are found. The BTC curves for unconfined flow differ from each other at the same Δh, unlike the case of confined flow, since the transport problem is nonlinear with respect to h. When Δh is constant and $h_1 \geq 5$m, the obtained BTCs do not differ significantly from the confined case. If $h_1 < 5$m, the corresponding BTC are dependent on h_2 and h_1. Also the rate of the pumping differs greatly (for the same Δh) when h_1 and h_2 are in unsaturated levels. At $\Delta h = 5$m, the pumping rate Q is 32.43m^3/day for confined flow and 11.35m^3/day for unconfined flow ($h_1 = 1$, $h_2 = 6$). The lower transmissivity for unconfined flow explains the shifting of the BTC to the right.

The differences can also be clarified by the breakthrough time (t_{BTT}), which we define as the fastest possible breakthrough of contaminant in the absence of diffusive terms. This t_{BTT} follows from (2.53) with $u = \pi$,

$$t_{\text{BTT}} = \int_{v^{(1)}}^{v^{(2)}} \frac{h_{\text{eff}} dv}{K(\cosh(v) + 1)^2},$$

which in the confined regime simplifies to

$$t^{\text{c}}_{\text{BTT}} = \frac{\kappa}{\Delta h} \int_{v^{(1)}}^{v^{(2)}} \frac{dv}{(\cosh(v) + 1)^2},$$

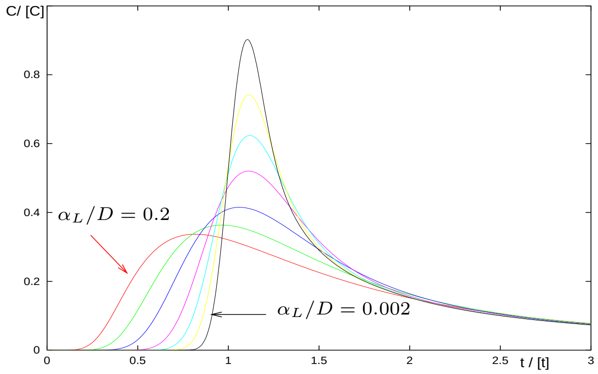


Figure 2.13: BTC with pulse input for $\alpha_L/D = 0.2$, 0.1, 0.05, 0.02, 0.01, 0.005 and 0.002m scaled to match [76].

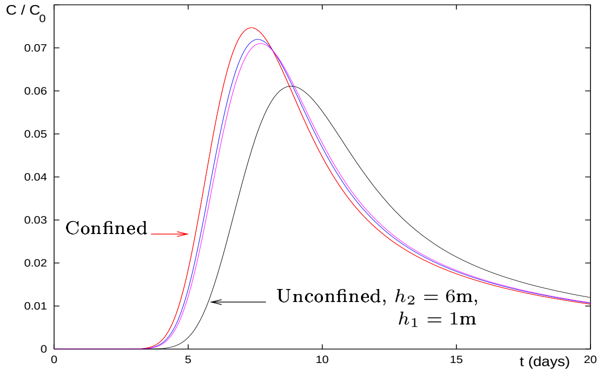


Figure 2.14: BTC for 4 different head values. The top curve corresponds to confined flow and $\Delta h = 5$. The other curves correspond to partially confined-unconfined flow with $h_2 = 13$m, $h_1 = 8$m, and unconfined flow with $h_2 = 10$m, $h_1 = 5$m and $h_2 = 6$m, $h_1 = 1$m, respectively

and in the unconfined regime to

$$t^{\mathrm{uc}}_{\mathrm{BTT}} = \frac{\kappa}{\Delta h}\int_{v^{(1)}}^{v^{(2)}} \frac{h(v)}{\frac{h_1+h2}{2}} \frac{dv}{(\cosh(v)+1)^2}.$$

Here, $\kappa = \frac{\delta^2\theta_0}{4k}(v^{(2)} - v^{(1)})$. As in the unconfined domain $h(v) = \sqrt{2\frac{A\,v+B}{k}}$, we will always have $t^{\mathrm{c}}_{\mathrm{BTT}} < t^{\mathrm{uc}}_{\mathrm{BTT}}$. Thus, for our example we have $t^{\mathrm{c}}_{\mathrm{BTT}} = 6.47$days, and for $h_1 = 1$m$, h_2 = 6$m we have $t^{\mathrm{uc}}_{\mathrm{BTT}} = 7.88$days. Note that these values correspond to the peaks of the corresponding BTCs in Fig. 2.14, which is exactly what could be expected after injection of a short pulse.

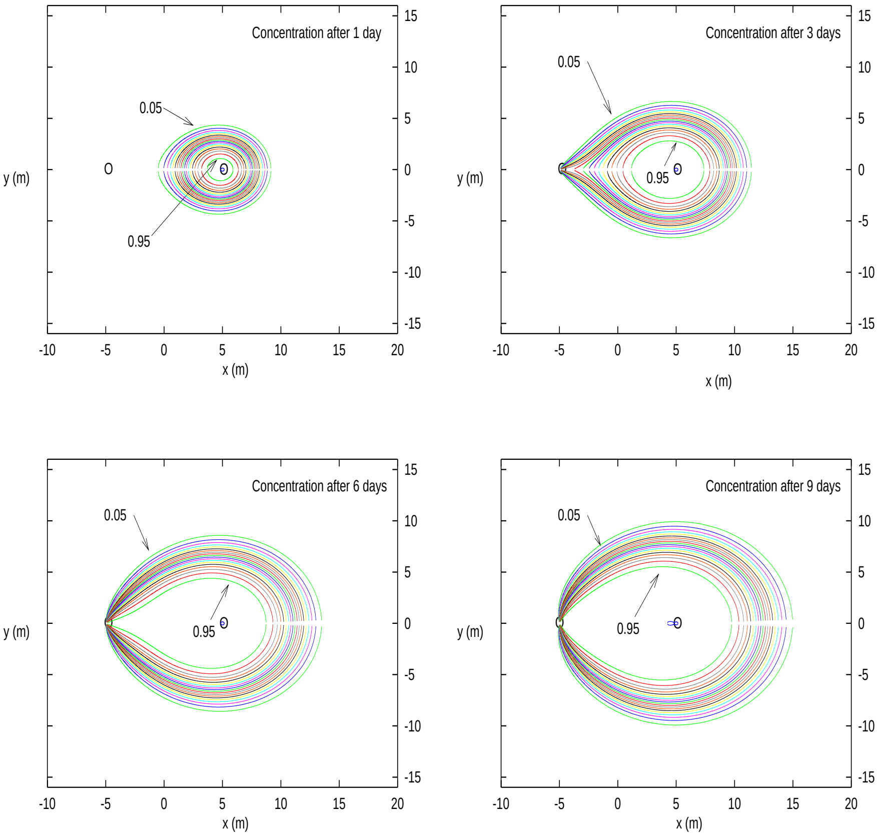


Figure 2.15: Concentration levels for relative values $0.05, 0.10, \ldots, 0.95$ for the standard example with $\alpha_L = 0.5$m and $\alpha_T = 0$m, after 1, 3, 6 and 9 days of operation.

Experiment 5: influence of dispersion coefficients

In Fig. 2.15 we plot the evolution of the tracer concentration after 1, 3, 6 and 9 days in the (x, y) plane, caused by a step input for the standard example with $\alpha_L = 0.5$m, $\alpha_T = 0$m, $h_1 = 4$m and $h_2 = 15$m. Note that the concentration levels (isoclines) are relative, i.e., the value 0.5 means that the concentration in the pore water equals half the concentration of the tracer at the injection well. As the well radius is not visible on the plots, a small circle has been drawn around the wells. Since the solution is symmetrical with respect to the x-axis, we shall only plot the positive y-axis from now on.

To illustrate the influence of the transversal dispersivity α_T on the solution, we plot in Fig. 2.16 the concentration levels for the same data and times as in Fig. 2.15, except that $\alpha_T = 0.01$m. To illustrate the influence of the longitudinal

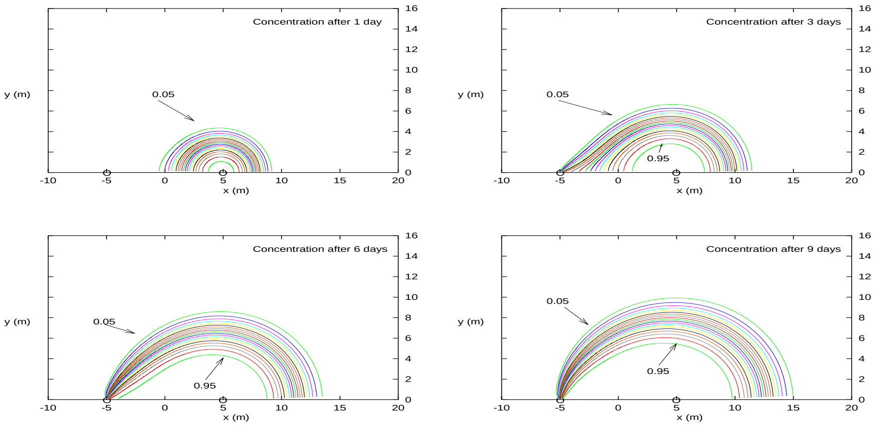


Figure 2.16: Concentration levels for relative values $0.05, 0.10, \ldots, 0.95$ for the standard example with $\alpha_L = 0.5$m and $\alpha_T = 0.01$m, after 1, 3, 6 and 9 days of operation.

dispersivity α_L on the solution, we plot in Fig. 2.17 the concentration levels for $\alpha_L = 0.05$m (keeping $\alpha_T = 0.01$m).

Finally, in Fig. 2.18, we show the result for $\alpha_L = 0 = \alpha_T (= D_0)$.

To give an overview of the effect on the BTC of different values of longitudinal and transversal dispersivity, the BTCs of the previous 4 experiments are shown

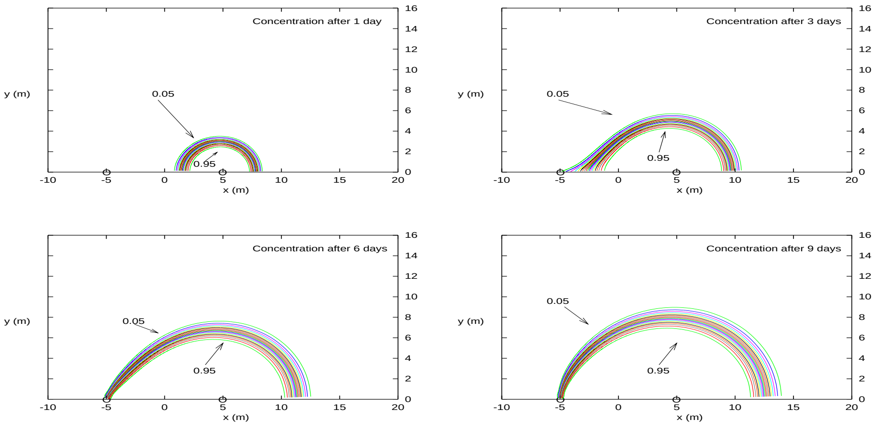


Figure 2.17: Concentration levels for relative values $0.05, 0.10, \ldots, 0.95$ for the standard example with $\alpha_L = 0.05$m and $\alpha_T = 0.01$m, after 1, 3, 6 and 9 days of operation.

in Fig. 2.19.

By comparing the results in Fig. 2.15-2.19 we want to illustrate the influence of the dispersion coefficients. When $\alpha_L = \alpha_T = 0$ ($D_0 = 0$) the concentration field should exhibit a shock (along one isocline). In Fig. 2.18 we obtain a system of isoclines creating a narrow strip which becomes broader after a long time period, especially in the regions far away from the extraction well. However, these regions are only of importance for the late time behaviour of the BTC. The smoothening of the shock is due to the numerical dispersion of our method (in the projecting process). These results are still better than those obtained by other approximation methods (e.g., upwinding). We have taken α_T in these experiments only 10% of the value of α_L, as is discussed in the literature. The influence of such a value of α_T is significantly smaller compared to that of α_L. The presence of α_T can only be recognized by the fact that the isoclines can then end behind the extraction well, which cannot appear when $\alpha_T = 0$. (Compare Fig. 2.15 and Fig. 2.16, and note at the extraction well a sharp transition in the isoclines, or not). This illustrates the ability of the transversal dispersivity to spread contaminant from one streamline to adjacent streamlines. The influence of α_T on the BTC is depicted in more detail in Fig. 2.20, see Experiment 6.

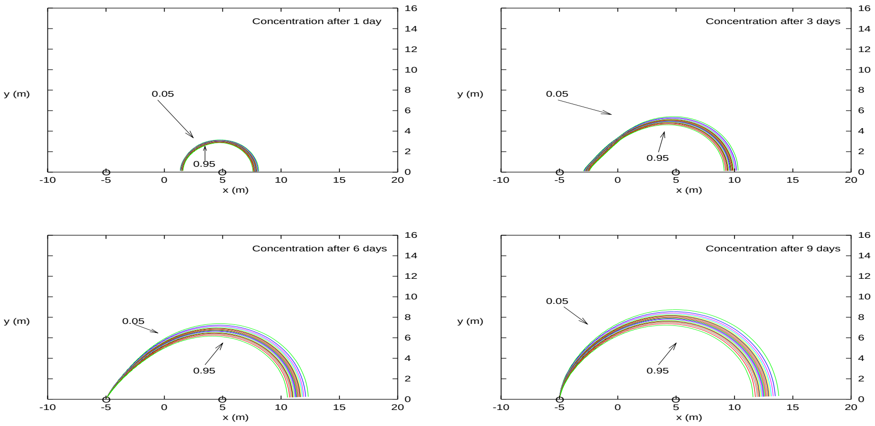


Figure 2.18: Concentration levels for relative values $0.05, 0.10, \ldots, 0.95$ for the standard scheme with $\alpha_L = 0.0$m and $\alpha_T = 0.0$m, after 1, 3, 6 and 9 days of operation.

The influence of the longitudinal dispersivity is more significant, and we can compare the time evolution of the strip of isoclines, which is broader for the higher values of α_L. Also, the time in which contaminant reaches the extraction well is shorter for larger values of α_L (compare Fig. 2.16, 2.17, 2.18.). The most convincing influence of α_L can be seen on the BTCs in Fig. 2.19. The shape of the BTCs, especially in the early phase, and also their starting points, are the most important for the determination of the parameter α_L when the inflow BC was a step input. However, if the inflow BC is of pulse type, then the time evolution of the entire BTC will be useful for the determination of α_L.

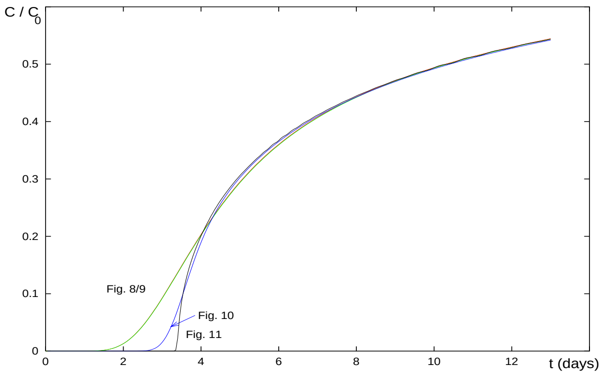


Figure 2.19: BTCs of the tests in experiment 5, Fig. 2.15-2.18.

Experiment 6: transversal dispersivity

We illustrate the influence of the transversal dispersivity on the BTC with a pulse input of 1 day in the standard example with $\alpha_L = 0.1$, $h_1 = 10$m, $h_2 = 15$m, and with α_T taking the values 0, 0.02, 0.05, 0.1 and 0.2m. The results are obtained with the standard method (since α_T is nonzero) in a 80×400 grid with an operator splitting time step of 0.05 days. In Fig. 2.20, we see that the influence of α_T is very small. The maximum value of the BTC shifts somewhat backwards and is slightly reduced with increasing α_T. We may conclude that to compute α_T from a measured BTC with an inverse method, the measurement data will have to be very precise.

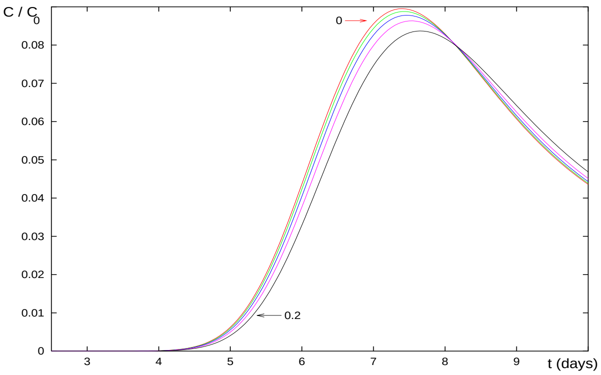


Figure 2.20: BTC with pulse input for $\alpha_L = 0.1$m, and $\alpha_T = 0$ (top curve), 0.02, 0.05, 0.1 and 0.2m (lowest curve).

Comparison with method of lines

The problem at hand was solved with an upwind method of lines in [13]. This method gave bad results. As a comparison, one of the results obtained is given for the following data: $r_1 = r_2 = 15$cm, $d = 10$m, $H = 10$m and $h_1 = 4$m, $h_2 = 15$m, porosity of the soil $\theta_0 = 0.2$, hydraulic conductivity $k = 10^{-5}$m/s $=$ 0.864m/day. Dispersivities are taken as $\alpha_L = 0.05$, $\alpha_T = 0.01$, $D_0 = 10^{-20} \approx 0$. Computations are done in a 40x40 grid with an upwind method of lines, see Fig. 2.21. These figures must be compared with Fig. 2.17. The implemented method of lines is seen to have far too much numerical dispersion to make this method suitable for parameter identification.

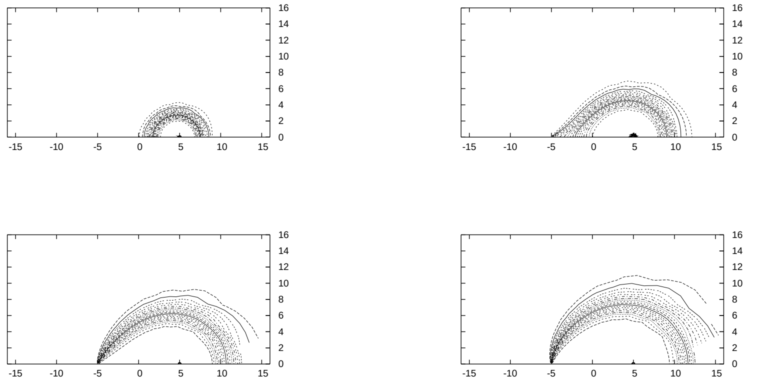


Figure 2.21: Concentration levels for relative values $0.05, 0.10, \ldots, 0.95$ with method of lines after 1, 3, 6 and 9 days of operation.

2.6.2 Mass balance and advective inflow flux

In all experiments up to now, we worked with the advective boundary condition, see (2.49), which is transformed to (2.50):

$$\left(D_0\theta_0 h_{\mathsf{eff}} + \frac{2\alpha_L\lambda}{\delta}(\partial_v\widetilde{\Phi}(v))\right)\partial_v C + (\partial_v\widetilde{\Phi}(v))C = (\partial_v\widetilde{\Phi}(v))C_0(t), \quad v = v^{(2)}.$$

We suggested to split as follows: for the transport equation take

$$C(u, v, t) = C_0(t), \quad v = v^{(2)}, \tag{2.124}$$

and for the diffusion equation

$$\partial_v C(u, v, t) = 0, \quad v = v^{(2)}. \tag{2.125}$$

We proved convergence of the solution obtained with this splitting method to the solution of the original convection-diffusion problem with an advective inflow BC. This corresponds to the physical meaning of dispersion, which cannot contribute to diffusive flux at the inflow. Note that this assumes D_0 to be neglectable, as molecular diffusion does contribute there to a diffusive flux.

For these BC's, the mass balance only consist of advective inflow and advective outflow. Generally, for the mass balance, we need the advective inflow flux $M_{\text{adv}}^{\text{in}}$,

$$\begin{aligned} M_{\text{adv}}^{\text{in}} &= \int_{\Delta t} \int_{\delta B_{r_2}(d+c,0)} \theta_0 h_{\text{eff}} C_p(x,y,t)(\boldsymbol{n}\cdot\boldsymbol{v})\, ds\, dt \\ &= (\partial_v \tilde{\Phi}(v)) \int_{\Delta t} \int_0^{\pi} C(u, v^{(2)}, t)\, dt\, du, \end{aligned}$$

the diffusive flux at the inflow,

$$\begin{aligned} M_{\text{diff}}^{\text{in}} &= -\int_{\Delta t} \int_{\delta B_{r_2}(d+c,0)} \theta_0 h_{\text{eff}} (\boldsymbol{n}\cdot D\nabla C(x,y,t))\, ds\, dt \\ &= -\int_{\Delta t} \int_0^{\pi} \left(D_0 \theta_0 h_{\text{eff}} + \frac{2\alpha_L}{\delta} \lambda (\partial_v \tilde{\Phi}(v)) \right) \partial_v C(u, v^{(2)})\, dt\, du, \end{aligned}$$

–which can be positive or negative– and the advective outflow flux,

$$\begin{aligned} M_{\text{adv}}^{\text{out}} &= \int_{\Delta t} \int_{\delta B_{r_1}(-d,0)} \theta_0 h_{\text{eff}} C_p(x,y,t)(\boldsymbol{n}\cdot\boldsymbol{v})\, ds\, dt \\ &= (\partial_v \tilde{\Phi}(v)) \int_{\Delta t} \int_0^{\pi} C(u, v^{(1)}, t)\, dt\, du, \end{aligned}$$

as well as the mass M^{pres} in the uv-space. This latter mass can be calculated in every cell by multiplying the concentration with the volume and the porosity, i.e.

$$M_{i,j}^{\text{pres}} = \theta_0 h_{\text{eff}} C_{i,j} \Delta V_{i,j}^{x,y},$$

where $C_{i,j}$ is the average concentration over the cell and where $\Delta V_{i,j}^{x,y}$ is the volume in the xy-plane that corresponds to the uv-cell:

$$\begin{aligned} \Delta V_{i,j}^{x,y} &= \int_{xy} d\sigma \\ &= \int\int \|\partial_u \boldsymbol{r} \times \partial_v \boldsymbol{r}\|\, du\, dv \\ &= \frac{\delta^2}{4} \int\int \frac{du\, dv}{\lambda^2}, \end{aligned}$$

where $\boldsymbol{r} = (x(u,v), y(u,v), 0)$. This expression cannot be calculated analytically and will be approximated numerically.

Due to mass balance we have

$$M_{\text{adv}}^{\text{in}} + M_{\text{diff}}^{\text{in}} - M_{\text{adv}}^{\text{out}} = M_{i,j}^{\text{pres}}. \tag{2.126}$$

In the case of the BC (2.45) we have the simplification $C(u, v^{(2)}, t) = C_0(t)$m and hence

$$M_{\text{adv}}^{\text{in}} = (\partial_v \tilde{\Phi}(v))\pi \int_{\Delta t} C_0(t)\, dt.$$

In the case of the advective boundary condition (2.50) we will have

$$M_{\text{adv}}^{\text{in}} + M_{\text{diff}}^{\text{in}} = (\partial_v \tilde{\Phi}(v))\pi \int_{\Delta t} C_0(t)\, dt.$$

The suggested splitting of the BC (2.45) by means of (2.124)-(2.125), will fulfill this equation.

We now present a series of experiments in which the influence of the boundary condition at the inflow is illustrated, and we present mass balance tables.

Experiment 1: step input

We repeat the step input experiment. In other words, the wells have a radius of $r_1 = r_2 = 15$cm and their centers are placed 10m from each other ($d = 5$m, $c = 0$). The height of the aquifer is $H = 10$m, the porosity of the soil is $\theta_0 = 0.2$, the hydraulic conductivity is $k = 10^{-5}$m/s $= 0.864$m/day and the transversal counterpart and the molecular diffusion are considered neglectible, $\alpha_T = 0 = D_0$. The longitudinal dispersivity α_L is varied, and we measure the BTC. We inject the tracer with constant concentration $C^{(2)}(t) = C_0$. For an easy presentation we set $C_0 = 1$. The result is shown in Fig. 2.22. Little difference is found for a low longitudinal dispersivity. For large α_L, there is a noticeable difference. This is even more apparant in a mass balance table like Tables 2.4 and 2.5. Mass values are obtained dimensionless as $C_0(t) = 1$. Realistic values in kg can be obtained by multiplication with an appropriate inflow concentration. In Table 2.4 we see the situation for $\alpha_L = 2$. In this case, there is a large difference in the results for the two different BC's. However, when $\alpha_L = 0.1$, we obtain a neglectable difference between the two BC's, see Table 2.5.

Experiment 2: pulse input

In Fig. 2.23 we consider the same experiment, but now for a pulse input of 1 day. The same observation holds: only for large dispersion a difference is found

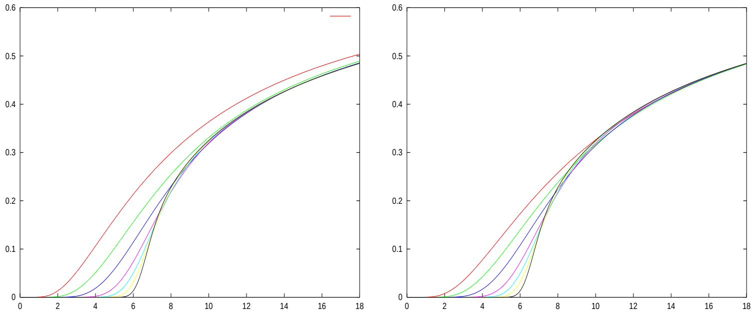


Figure 2.22: BTC with step input for $\alpha_L/L = 0.2$, 0.1, 0.05, 0.02, 0.01, 0.005 and 0.002m. Left for Dirichlet inflow BC with operator splitting method. Right for advective flux inflow BC under OS.

days	M^{in}	$M^{\text{out}}_{\text{adv}}$	M^{pres}	Bal.	t	M^{in}	$M^{\text{out}}_{\text{adv}}$	M^{pres}	Bal.
1	26.07	0.0009	25.35	0.72	1	16.16	0.0004	16.19	-0.03
5	99.40	4.097	94.34	0.97	5	80.79	2.950	78.20	-0.35
10	183.76	26.316	156.79	0.66	10	161.59	21.948	140.53	-0.89
15	266.21	60.2843	205.679	0.25	15	242.38	53.404	192.42	-1.41

Table 2.4: Total mass inflow and outflow, total mass present, and mass balance, in the case of $\alpha_L = 2$ with step input. Left for Dirichlet inflow BC with operator splitting method. Right for advective flux inflow BC under operator splitting (OS).

days	M^{in}	$M^{\text{out}}_{\text{adv}}$	M^{pres}	Bal.	t	M^{in}	$M^{\text{out}}_{\text{adv}}$	M^{pres}	Bal.
1	16.28	0	16.29	-0.01	1	16.16	0	16.19	-0.03
5	80.91	0.033	81.17	-0.29	5	80.79	0.032	81.06	-0.30
10	161.59	13.860	148.83	-0.99	10	161.59	13.827	148.76	-1.00
15	242.5	45.526	198.54	-1.57	15	242.38	45.480	198.48	-1.58

Table 2.5: Total mass inflow and outflow, total mass present, and mass balance, in the case of $\alpha_L = 0.1$ with step input. Left for Dirichlet inflow BC with operator splitting method. Right for advective flux inflow BC under OS.

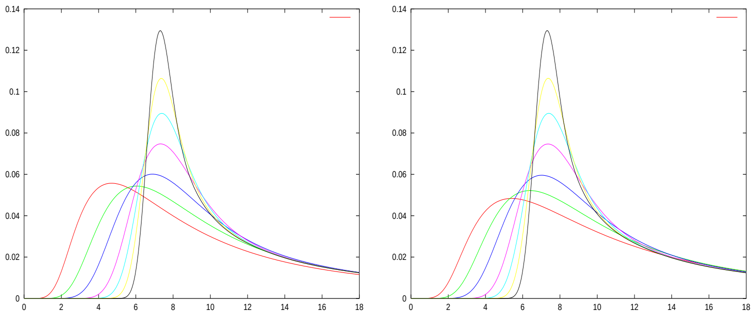


Figure 2.23: BTC with pulse input of 1day for $\alpha_L/L = 0.2$, 0.1, 0.05, 0.02, 0.01, 0.005 and 0.002m. Left for Dirichlet inflow BC with operator splitting method. Right for advective flux inflow BC under OS.

days	M^{in}	$M^{\text{out}}_{\text{adv}}$	M^{pres}	Bal.	t	M^{in}	$M^{\text{out}}_{\text{adv}}$	M^{pres}	Bal.
1	26.07	0.0009	25.35	0.72	1	16.16	0.0004	16.19	-0.03
5	17.42	2.192	15.25	-0.03	5	16.16	1.657	14.62	-0.12
10	16.16	5.651	11.07	-0.07	10	16.16	5.036	11.24	-0.12
15	16.16	7.405	9.09	-0.08	15	16.16	6.982	9.29	-0.11

Table 2.6: Total mass inflow and outflow, total mass present, and mass balance, in the case of $\alpha_L = 2$ with pulse input of 1 day. Left for Dirichlet inflow BC with operator splitting method. Right for advective flux inflow BC under OS.

between the two BC's. This follows clearly from Table 2.6, where the mass balance is given for $\alpha_L = 2$. From the values for the Dirichlet type of BC, we see there is an inflow diffusive flux up to $t = 1$day, after which there is a clear negative inflow diffusive flux causing the total inflow mass to decrease. In Fig. 2.23 this gives rise to the higher peak value and lower tail for $\alpha_L = 2$ in the Dirichlet type BC, as opposed to the case of an advective flux BC.

2.6.3 Adsorption

We now give some solutions of the dual-well direct problem where the adsorption coefficients are varied. We limit ourselves to Freundlich adsorption. Only break through curves are given, as the plots in the xy domain provide little extra

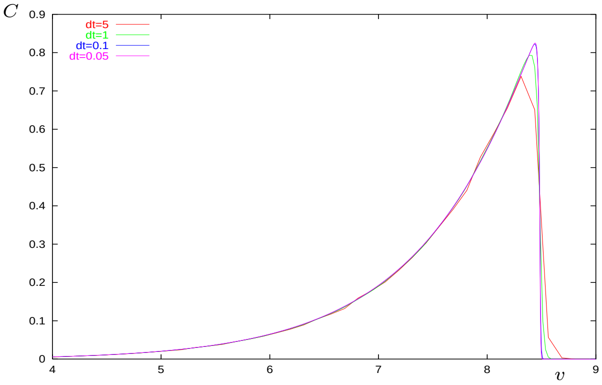


Figure 2.24: Solution of equilibrium mode problem for $D = 0.001$, $K_0 = 1$, $p = 0.75$, with $\frac{\tau}{\Delta y} = 40$ fixed and with τ respectively being 5, 1, 0.1, 0.05, at final time $T = 15$.

information compared to the linear solutions.

Convergence tests in 1-D

First we investigate the convergence and the dependence of the solution on the discretization. For this we use (2.62) in 1D, $\Omega = (0, L)$, in stead of in the dual-well setting. We take $a = 0$, $G = 1$, $b = D$, $g = 1$, $F(C) = C + K_0 C^p$, together with homogeneous Neumann boundary conditions and a Riemann initial condition: $u_0(x) = 1$ if $x \leq 1$, $u_0(x) = 0$ otherwise. This setting corresponds to the developed numerical approximation in one single strip. It provides a good view upon convergence as the plots are transparant and can be related to simple examples of transport problems.

From theoretical considerations we know that for fixed $\tau/\Delta y$ the solution converges as $\tau \to 0$. Here, y is the transformed variable from (2.53). This is behaviour we observed in Fig. 2.24, where $D = 0.001$, $K_0 = 1$ and $p = 0.75$. Starting from $\tau = 0.1$, the concentration profile does not change noticeably anymore as convergence is reached. The two telltale signs of adsorption can be observed: the development of a tail to the left, and the sharp front that remains at the right although diffusion is present.

In Fig. 2.25 we illustrate the influence of the different errors by presenting the same example for 3 different values of the diffusion: $D = 0.1$, 0.01 and

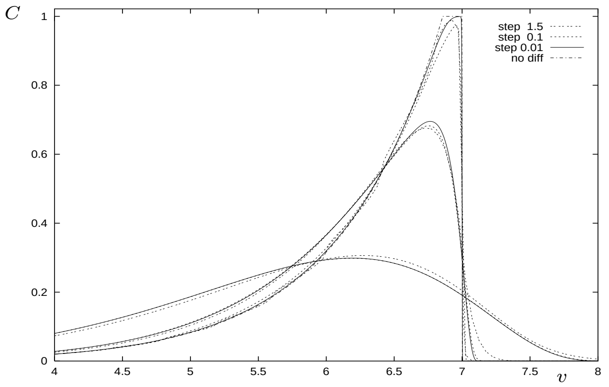


Figure 2.25: Solution of an equilibrium mode problem with $K_0 = 1$, $p = 0.75$, at time $T = 12$, for 3 values of the diffusion: $D = 0.1$, 0.01 and 0.0001, each with 3 different discretizations: $\Delta y = 0.05$, $\tau = 1.5$ (dashed line) ; $\Delta y = 0.02$, $\tau = 0.1$ (dotted line); and $\Delta y = 0.002$, $\tau = 0.01$ (solid line). Also the analytical solution for $D = 0$ is given.

0.0001. The first error is the operator splitting error, due to the splitting of the PDE in different pieces. This error decreases when τ does. As the transport part is exact, no error is introduced in this part. Instead, we have a projection error due to the projection of the transport solution to piecewise constant functions. This error decreases when Δy does. Next, we have the error from the diffusive part, which decreases with smaller timesteps and smaller gridsize. We observe that for $\tau = 1.5$ the operator splitting error produces quite large errors, even in the large diffusion case (where numerical dispersion of the approximations is smaller than the diffusion). For $D = 0.1$ and $D = 0.01$ convergence has been reached by the method, as can be deduced from the small change when passing from $\tau = 0.1$ to $\tau = 0.01$. This is not the case for $D = 0.0001$, where convergence is not reached yet. This is entirely due to the projection error, as in this case we have almost no diffusion, which results in sharps shocks that can only be resolved completely in a fine grid.

Variation of adsorption parameters

We investigate the dependence of the BTC of a dual-well experiment on the parameters of the model. A good dependence is crucial for later parameter identification. For these experiments the wells have each a radius of 15cm and their centers are placed 10m apart. The height of the aquifer is 10m, the porosity of the soil is $\theta_0 = 0.2$ and the hydraulic conductivity is 0.864m/day. The head value is 10m at the extraction well and 15m at the injection well. The transversal dispersivity α_T, and the molecular diffusion D_0 are kept 0, as they can be neglected in most cases. We inject the tracer with constant concentration $C_0(t) = C_0$ during 1 day.

We have already shown that the BTC has a good dependence on the longitudinal dispersivity and on the hydraulic conductivity. Now we investigate the influence of the adsorption parameters on the equilibrium mode Freundlich type adsorption. For the nonlinear case, we have to work dimensionless, so we transform (2.62) (where $\alpha_T = 0$ and $D_0 = 0$) to

$$\partial_t \left(\frac{C}{C_0} + \frac{\varrho C_0^{p-1}}{\theta_0} \Phi \left(\frac{C}{C_0} \right) \right) = g \left\{ \partial_v \left(b \partial_v \frac{C}{C_0} \right) \right\} + G \partial_v \frac{C}{C_0}, \tag{2.127}$$

where $\Phi(s) = K_0 s^p$. We investigate the dependence on $\rho = \frac{\varrho C_0^{p-1}}{\theta_0} K_0$ in Fig. 2.26 and on p in Fig. 2.27. The ρ parameter controls the retardation of the BTC. The p-parameter controls the nonlinearity. The highly nonlinear case, $p = 0.25$, can be distinguished from the setups close to linearity (p close to 1).

These results indicate the possibility to use the dual-well experiment for adsorption parameter identification of the subsurface by the method presented.

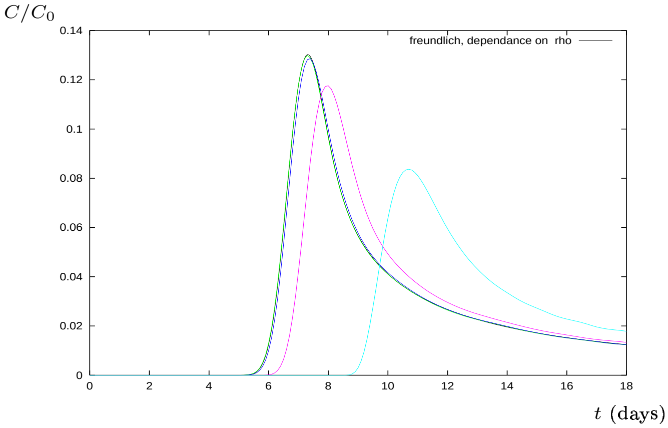


Figure 2.26: Dependence of the BTC for pulse input on the value of ρ with fixed $p = 0.9$ and $\alpha_L = 0.02$m. From left to right we have ρ equal to 0 (linear case), 0.01, 0.1 and 0.5.

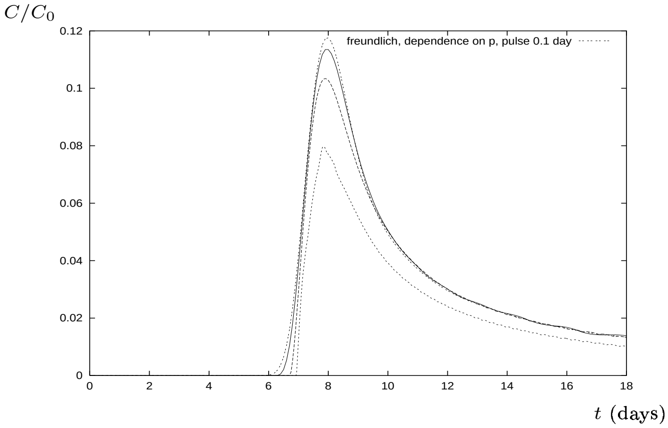


Figure 2.27: Dependence of the BTC for pulse input on the value of p with fixed $\rho = 0.1$ and $\alpha_L = 0.02$m. From top to bottom we have p equal to 0.9, 0.75, 0.5 and 0.25.

Chapter 3

A degenerate diffusion problem in 1D

We now focus on a degenerate diffusion problem. With degenerate we mean that the diffusion coefficient can be 0. Numerical, and even analytical solutions to several problems of this type are known.

We solve this type of problem as it appears in the engineering set-up that we want to model: enrichment of steel with silicon by diffusion annealing. The final aim is to solve an inverse problem which we consider in Part II, Chapter 4.

As a solution technique we choose the *method of lines*, also called semi-discretization technique: a PDE is transformed into an ODE by a suitable spatial discretization. For a comprehensive overview, we refer to [29]. This technique allows for solving many practical problems, including hyperbolic problems when information concerning the characteristics is incorporated in the method. The result illustrated in Fig. 2.21 was obtained by this method. We emphasize that this result was obtained in one of the most complex settings for the mehod of lines: dominant convection. With extra efforts (e.g. by the use of flux limiters, ...) probably better results might be obtained there. However, to reduce computing times, a different solution method was invoked. For the actual problem, no diffuculties with the method of lines are expected.

3.1 Annealing diffusion as a practical example

Electrical steel is an excellent soft magnetic material used for the construction of electrical motors and transformers. Its composition is basically high purity FeSi or FeSiAl alloys. Normally, the alloying content never exceeds 3 wt%. Beyond this concentration the material becomes very brittle due to the concurrence of the ordering phenomena D03 and B2 and it is not possible to perform cold rolling [67].

However, the magnetic properties, namely power losses and magnetostriction are optimized when the alloying content reaches 6.5 wt%.

High Si and Al electrical steel with improved magnetic properties can be produced by hot dipping in a molten Al-25%Si bath followed by diffusion annealing. The hot dipping gives rise to a Si rich layer on top of the substrate, which is subsequently diffused into the bulk at high temperatures, called annealing diffusion. Hence, this higher Si content alloys can only be manufactured if an additional final step is introduced in the production route of electrical steel: enrichment by surface deposition of Si and Al and next its diffusion into the bulk material. Recent research has shown that the magnetic and mechanical properties of the high Si electrical steel produced by diffusion depend strongly on the shape of the diffusion profile obtained after the annealing [4]. The different applications of the electrical steel require different magnetic and mechanical properties. Therefore, we need a diffusion model capable of predicting the diffusion profiles depending on the different conditions of the production process (e.g. annealing temperature, time, Si and Al content of the substrates, microstructure previous to the diffusion annealing) is necessary. We will present this model in the following Sections. The results appeared in [3, 56].

3.1.1 Mathematical model of diffusion annealing

From [66], we know that the Fe-Si interdiffusion is highly dependent on the Si concentration. Therefore, taking into account the dependence of the diffusion D on the Si concentration, C, is indispensible. As we have a ternary system, this is analogue to [60]. Therefore, the diffusion equation to be used is, with $i = 1$ (Si), 2 (Al),

$$\frac{\partial C_i^3(x,t)}{\partial t} = \frac{\partial}{\partial x}\left(D_{i1}^3\left(C_1^3, C_2^3\right)\frac{\partial C_1^3(x,t)}{\partial x} + D_{i2}^3\left(C_1^3, C_2^3\right)\frac{\partial C_2^3(x,t)}{\partial x}\right), \quad (3.1)$$

where $0 \leq x \leq L$, $0 \leq t < \infty$. The superscript 3 indicates the dependent element (Fe). Furthermore, we have the noflow boundary conditions

$$\frac{\partial C_i^3(0,t)}{\partial x} = 0 = \frac{\partial C_i^3(L,t)}{\partial x}, \tag{3.2}$$

along with initial conditions

$$C_i^3(x,0) = C_{0,i}^3(x). \tag{3.3}$$

The Si concentration C_1^3, (mol/mm^3) is obtained from

$$C_1^3 = \frac{x_{\mathrm{Si}}}{x_{\mathrm{Si}} v_{\mathrm{Si}} + x_{\mathrm{Al}} v_{\mathrm{Al}} + (100 - x_{\mathrm{Si}} - x_{\mathrm{Al}}) v_{\mathrm{Fe}}}, \tag{3.4}$$

where x_i is the atomic percent and v_i is the molar volume of each element ($v_{\mathrm{Si}} = 12.0 \times 10^3$, $v_{\mathrm{Al}} = 10.0 \times 10^3$, $v_{\mathrm{Fe}} = 7.10 \times 10^3$ mm^3/mol). The Al concentration is obtained in the same way.

We work in one dimension only. Hence, (3.1)-(3.3) can only be used once the coating has been made homogeneous in the lateral direction. Therefore, for the initial condition for this system we use the measured concentration profile after some minutes of diffusion annealing.

Our main problem when solving (3.1)-(3.3), is the fact that the interdiffusion coefficients D_{ij}^3 are unknown. Therefore, we not only have to solve the equations, but at the same time we need to retrieve these interdiffusion coefficients from the experiments.

First, the coupled system of PDEs is reduced to a single PDE. This is possible because the diffusion path during diffusion annealing is monotone in the Al-Si concentration. We refer to the experiments, see Fig. 3.5, and to other ternary diffusion paths in [21]. It follows that the diffusion path can be assumed to be given by a well defined function

$$C_2^3 = f(C_1^3). \tag{3.5}$$

Given the at% of Si in the steel, we can extract the value of Al along the diffusion path.

Under the above assumption, the PDE system decouples, and we can write formally,

$$\frac{\partial C_1^3(x,t)}{\partial t} = \frac{\partial}{\partial x}\left(D_1^{23}\left(C_1^3\,(x,t)\right) \frac{\partial C_1^3(x,t)}{\partial x} \right), \tag{3.6}$$

where $0 \leq x \leq L$, $0 \leq t < \infty$, and D_1^{23} is the *apparent diffusion coefficient* of Si in Fe along the diffusion path f. Again noflow boundary conditions hold

$$\frac{\partial C_1^3(0,t)}{\partial x} = 0 = \frac{\partial C_1^3(L,t)}{\partial x}, \tag{3.7}$$

along with an initial condition

$$C_1^3(x,0) = C_{0,1}^3(x). \tag{3.8}$$

Remark 3.1.1. *Diffusion in alloys can be subjected to the so-called Kirkendall effect, see [21] chapter 17. This effect causes, due to the difference of diffusion speed of the different components, a solid-state advection (lattice flow). This complicates the analysis. Formally it can be seen as another part of the apparent diffusion coefficient, cf. [21].*

3.1.2 Moving interface

We first introduce the substrate problem.

Definition 3.1.1. *The substrate problem is defined by*

$$\partial_t u - \partial_x \left(D(u)\ \partial_x u\right) = 0 \quad \text{in } (0,L) \times (0,T), \tag{3.9}$$

with $D(u) \in L^\infty$, $D(u) \geq 0$, along with boundary conditions

$$u(0,t) = c_0 \quad \text{or} \quad -D(u(0,t))\partial_x u = 0, \tag{3.10}$$

$$-D(u(L,t))\partial_x u = 0, \quad 0 < t < T, \tag{3.11}$$

and an initial condition

$$u(x,0) = u^0(x), \quad x \in (0,L), \tag{3.12}$$

where the initial function $u^0(x)$ is given by

$$u^0(x) > \delta \quad x \in [0, s_0[, \quad u^0(x) = \delta \quad x \in [s_0, L], \tag{3.13}$$

with $0 < s_0 < L$, $\delta > 0$. We assume further that $u^0(x)$ is smooth up to $x = s_0$.

The function $u^0(x)$ is not analytic in $x = s_0$. The value of δ is the amount of Si in Fe present in the substrate without any annealing performed. As a special case we have the *zero substrate problem* when $\delta = 0$.

A characteristic property of the zero substrate problem combined with degenerate diffusion $D(0) = 0$, is the movement of the *contact point* $x = s(t)$.

Definition 3.1.2. *The contact point is the point $x = s(t)$ where $u(s(t), t) = 0$, with $s(0) = s_0$.*

The speed of the contact point is denoted as $\dot{s}(t)$. We have the following property.

Proposition 3.1.1. *Assume that the speed of the contact point is finite in the substrate problem. Let (3.9) be satisfied in $x = s(t)$ in the limit sense. Then the following holds*

$$\begin{aligned} \dot{s}(t) &= -\lim_{x\to s(t)^-} \left(D'(u)\partial_x u + D(u)\frac{\partial_x^2 u}{\partial_x u} \right) && (3.14) \\ &= -\lim_{x\to s(t)^-} \left(\partial_x \int_0^u \frac{D(z)}{z}\,\mathrm{d}z \right) && (3.15) \end{aligned}$$

Proof. From the definition of the contact point we have that $u(s(t), t)$ is independent on t. Therefore, if $\dot{s}(t)$ is finite,

$$0 = \frac{\mathrm{d}}{\mathrm{d}t} u(s(t), t) = \partial_t u + \dot{s}(t)\partial_x u.$$

Hence,

$$\begin{aligned} \dot{s}(t) &= -\lim_{x\to s(t)-} \frac{\partial_t u}{\partial_x u} = -\lim_{x\to s(t)^-} \left(D'(u)\partial_x u + D(u)\frac{\partial_x^2 u}{\partial_x u} \right) \\ &= -\lim_{x\to s(t)^-} \frac{D(u)\partial_x u}{u} = -\lim_{x\to s(t)^-} \left(\partial_x \int_0^u \frac{D(z)}{z}\,\mathrm{d}z \right). \end{aligned}$$

Here, in the second line we noted that the flux $D(u)\partial_x u$ and the function u both tend to zero at the interface. This allows to use de l'Hospital's rule in reverse (as $D(0) = 0$). □

Remark 3.1.2. *From (3.15) it follows that the function $\frac{D(u)}{u}$ must be integrable.*

Remark 3.1.3. *From (3.14) some deductions can be made on the form of the concentration profile $u(x, t)$ in the contact point $x = s(t)$ in order that for a given diffusion coefficient the speed $\dot{s}(t)$ is indeed finite. If for example $D(u) = u^p$, $p > 0$, then we have $\lim_{x\to s(t)-} D(u) = 0$ in the zero substrate problem. The propagation speed (3.14) is finite in the following cases.*

- *If $\lim_{x\to s(t)-} \frac{\partial_x^2 u}{\partial_x u}$ is finite or behaves as $\frac{1}{u^{p-k}}$, where $0 \le k < p$, the second term of (3.14) is zero.*

- *If $p = 1$, then $D'(u) = c$, a constant. Therefore $\lim_{x \to s(t)-} \partial_x u = f(t)$, with $f(t) > 0$.*
- *If $p > 1$, then $\lim_{x \to s(t)-} D'(u) = 0$. To have a finite speed, it is necessary that $\partial_x u \sim u^{1-p}$. In this way, $\lim_{x \to s(t)^-} D'(u)\partial_x u$ is still a finite function of t. This means however that $\lim_{x \to s(t)^-} \partial_x u = \infty$.*
- *If $p < 1$, the same deduction can be made, now with $\lim_{x \to s(t)^-} \partial_x u = 0$.*

Remark 3.1.4. *From Remark 3.1.3 it follows that $\partial_x u$ should be avoided in numerical computations. Therefore, the form (3.15) should be used. If the speed of the contact point is finite, the function $F(x)$ defined by*

$$F(x) \equiv \tilde{F}(u(x)) = \int_0^{u(x)} \frac{D(z)}{z},$$

will have a finite derivative in $x = s(t)$.

In general, the initial concentration profile will only be a given set of datapoints. Therefore, $u^0(x)$ will be constructed by suitable polynomial interpolation. Hence, $\partial_x u$ and $\partial_x^2 u$ are normally two non zero, bounded, constants. From (3.14) we obtain that $\dot{s}(t) = 0$ for $p > 1$, and $\dot{s}(t) = \infty$ for $0 < p < 1$. This does not allow us to recover $s(t)$, but indicates what will happen numerically: the initial function will transform to a profile with the desired derivative in the contact point.

3.1.3 Analytical solution

For special forms of (3.6) analytical solutions are known. We are interested in particular with the cases where $D(C) = 0$, as this corresponds to a moving interface, which may be observed in the experiments. For the Cauchy problem, in the special case that $D(C,p) = (p+1)C^p$, $p > 0$, a closed form solution exists when the initial profile is $C(x, 0) = E\delta(x)$ (the Dirac measure). It is the Barenblatt-Pattle solution, [65] p.31, taking the form

$$v(x,t) = \begin{cases} t^{-1/(p+2)}(1 - (x/s(t))^2)^{1/p}, & \text{for } |x| < s(t); \\ 0, & \text{for } |x| \geq s(t), \end{cases} \tag{3.16}$$

with the interface given by

$$s(t) = \sqrt{\frac{2(p+1)(p+2)}{p}} t^{1/(p+2)}. \tag{3.17}$$

This solution has a singularity at $x = s(t)$.

This Barenblatt-Pattle solution can be used to test the numerical approximation. At $x = 0$ we have the required homogeneous Neumann boundary condition, and we have this also in $x = L$, as long as the interface does not reach $x = L$. The solution $C(x,t)$ to (3.6)-(3.7), with the initial condition taken to be $v(x,k)$ from (3.16) with k a positive constant and such that $s(k) < L$, is given by $v(x,t-k)$.

3.1.4 Numerical approximation

A numerical approximations that allows us to solve (3.6)-(3.8), for general diffusion coefficients is constructed. This method is used later on to extract the value of $D_1^{23}(C_1^3)$ from the experiments.

No moving interface

For simplicity, in the real experiments we will only consider the case of steels with Si 3 wt%. Hence, no phase changes occur, and we can assume $D_1^{23}(C_1^3) > 0$ in the entire domain. As the diffusion is everywhere positive, no moving interface arises. In the future also experiments with non Si enriched steels might be considered.

Given the apparent diffusion coefficient $D(C_1^3) \approx D_1^{23}(C_1^3)$, system (3.6)-(3.8) can be solved as in [59]. Here, we suggest a different approach. We construct the solution $C_1^3(x,t)$ of (3.6)-(3.8) in an approximative way by reducing it to an initial value problem for a nonlinear system of ODEs by means of a nonequidistant finite difference discretization with respect to the space variable. Next, a stiff ODE solver is used to solve the system of ODEs. The interval $(0, L)$ is partitioned by the set of grid points $\{x_i\}_{i=0}^N$. We denote $C_i(t) \approx C_1^3(x_i,t)$ and let $l_2(x,i)$ stand for the Lagrange polynomial of the second order interpolating the points (x_{i-1}, C_{i-1}), (x_i, C_i) and (x_{i+1}, C_{i+1}). Then, we approximate $\partial_x C$ by $dl_2(x_i,i)/dx \equiv (dl_2(x,i)/dx)_{x=x_i}$ and $\partial_x^2 C$ by $d^2 l_2(x_i,i)/dx^2 \equiv (d^2 l_2(x,i)/dx^2)_{x=x_i}$. To include the Neumann BC's, the governing PDE is extended to the boundary points. It is discretized similarly as in the inner points by the introduction of the fictive points y_{-1} and y_{N+1}, whose concentration values are chosen to satisfy the BC's. Equation (3.6) leads to the system of ODEs

$$\frac{d}{dt}C_i(t) - D(C_i)\frac{d^2}{dx^2}l_2(x_i,i) - D'(C_i)\left[\frac{d}{dx}l_2(x_i,i)\right]^2 = 0, \tag{3.18}$$

for $i = 0, \ldots, N$, where $D'(s) = \frac{dD(s)}{ds}$. This system of nonlinear ODE can be solved using a standard package for stiff ODEs, e.g. LSODA.

Moving interface

We indicate how the case of a moving interface can be best solved numerically. We split $(0, L)$ into two domains $\Omega_1 \equiv (0, s(t))$ and $\Omega_2 \equiv (s(t), L)$. Using Landau's transformation $y = \frac{x}{s(t)}$, the PDE (3.9) on $\Omega_1(t)$ becomes an equation on the fixed domain $(0, 1)$. Denoting the corresponding solution by $\overline{C^1}(y, t)$, it holds that

$$\partial_t C^1 = \partial_t \overline{C^1} - y\frac{\dot{s}(t)}{s(t)}\partial_y \overline{C^1}, \quad \partial_x C^1 = \frac{1}{s(t)}\partial_y \overline{C^1}, \quad \partial_x^2 C^1 = \frac{1}{s^2(t)}\partial_y^2 \overline{C^1}. \tag{3.19}$$

Thus, we are led to the transformed PDE

$$\partial_t \overline{C^1} - \frac{1}{s^2(t)}\partial_y \left(D(C^1)\partial_y \overline{C^1}\right) - y\frac{\dot{s}(t)}{s(t)}\partial_y \overline{C^1} = 0. \tag{3.20}$$

The interval $(0, 1)$ is partitioned by the set of grid points $\{y_i\}_{i=0}^N$, with $y_i = \sum_{l=0}^i \alpha_l$, $(i = 0, \ldots, N)$, where $\alpha_0 = 0$ and $\sum_{l=0}^N \alpha_l = 1$. We can choose these gridpoints so as to obtain a more dense discretization around the point $y = 1$. We denote $C_i^1(t) \approx \overline{C^1}(y_i, t)$ and let $l_2(y, i)$ stand for the Lagrange polynomial of the second order interpolating the points (y_{i-1}, C_{i-1}^1), (y_i, C_i^1) and (y_{i+1}, C_{i+1}^1). Then, we approximate $\partial_y \overline{C^1}$ by $dl_2(y_i, i)/dy \equiv (dl_2(y, i)/dy)_{y=y_i}$ and $\partial_y^2 \overline{C^1}$ by $d^2 l_2(y_i, i)/dy^2 \equiv (d^2 l_2(y, i)/dy^2)_{y=y_i}$. In the case of a Dirichlet BC, the nodal point y_0 need not be considered. In the case of a Neumann BC, we extend the governing PDE to the boundary point and discretize it similarly as for the inner points by the introduction of a fictive point y_{-1}. Equation (3.20) leads to the system of ODEs

$$\frac{d}{dt}C_i^1(t) - \frac{1}{s^2(t)}D(C_i^1)\frac{d^2}{dy^2}l_2(y_i, i) - \frac{1}{s^2(t)}D'(C_i^1)\left[\frac{d}{dy}l_2(y_i, i)\right]^2 - y_i\frac{\dot{s}(t)}{s(t)}\frac{d}{dy}l_2(y_i, i) = 0, \tag{3.21}$$

for $i = 1, \ldots, N-1$, and, in the case of a Neumann BC, also for $i = 0$.

In the second domain $\Omega_2(t)$ the concentration remains constant, $C(x, t) = \delta$. In the interface point $x = s(t)$ the following ODE must be satisfied

$$\dot{s}(t) = -\lim_{y \to 1^-} \frac{\frac{d}{dy}F_l}{s(t)} \tag{3.22}$$

where F_l is the second degree Lagrange polynomial interpolating the points $(y_{N-2}, \tilde{F}(C^1_{N-2}))$, $(y_{N-1}, \tilde{F}(C^1_{N-1}))$ and $(1, \tilde{F}(s(t)) \equiv 0)$. This equation only applies as long as $s(t) < L$. When $s(t) \geq L$, we switch to the case of no moving interface. Next, (3.21)-(3.22) are solved by means of a standard package for stiff ODEs, e.g. LSODA based on a backward finite difference formula.

Power type of diffusion

In the case of a power type diffusion coefficient, $D(s) = p_1 s^{p_2}(1 + p_3 s + p_4 s^2)$, $p_2 > 0$, an additional transformation is performed, see also [14]. This transformation is suggested by Remark 3.1.3 where we noted that $\partial_x u \sim u^{1-p}$. The transformation from u-variable to v-variable by means of $u = v^{1/p}$ leads to the expression $p \sim \partial_x v$. All power degeneracies are removed in this way, which makes $v(x)$ more appropriate for the application of Lagrange interpolation. For example, (3.16) is transformed into an expression for v that is quadratic in x. Using Lagrange polynomials of the second order for the space discretization, the space interpolation will be exact. Hence, the differences between the exact and the numerical solution of e.g. the Barenblatt-Pattle problem will be solely due to time integration errors. Generally, under the transformation $C = v^{1/p}$, we obtain, instead of (3.21), the ODE system

$$\begin{aligned} &\frac{d}{dt}v_i^1(t) - \frac{1}{s^2(t)} D((v_i^1)^{1/p}) \frac{d^2}{dy^2}\tilde{l}_2(y_i,i) - \frac{\left[\frac{d}{dy}\tilde{l}_2(y_i,i)\right]^2}{s^2(t)pv_i^1} \\ &\quad \left(D'((v_i^1)^{1/p})(v_i^1)^{1/p} + (1-p)D((v_i^1)^{1/p})\right) - y_i\frac{\dot{s}(t)}{s(t)}\frac{d}{dy}\tilde{l}_2(y_i,i) = 0, \qquad (3.23) \end{aligned}$$

and likewise for (3.22). Here, $\tilde{l}_2(y_i, i)$ is now the Lagrange polynomial of the second order interpolating the points (y_{i-1}, v^1_{i-1}), (y_i, v^1_i) and (y_{i+1}, v^1_{i+1}).

3.2 Physical experiments

These experiments were carried out by José Barros at LabMet, Department of Metallurgy and Materials Science, Ghent University. They will be extensively discussed, together with hot dipping experiments and mechanical/magnetic experiments of the samples, in his upcoming PhD-thesis. Here we give a short overview of the experiments concerning the diffusion annealing, which are relevant for the model we will construct.

3.2.1 Experimental procedure

The substrates chosen for the production of the high Si and Al alloys were commercial Fe-Si alloys (0 to 3wt%Si). After degreasing the electrical steel plates are subjected to the hot dipping: first the samples are preheated for 45s at 800°C and then they are dipped in a molten Al-25 wt%Si bath at 800°C for times ranging between 5 and 100s. Finally, the samples cool down under a flux of N_2. After hot dipping the specimens were heated in a resistance tube furnace under a N_2 protective atmosphere. The temperatures ranged between 900 and 1100°C.

Samples were carefully polished and the concentration profiles through the thickness of the samples were determined by EDS in a SEM. EDS was also used to analyze element concentration in the different layers present in the coating.

3.2.2 Coating composition and formation

The coating formation is a reaction-diffusion process. Together with intermetallic growth there is substrate dissolution in the molten bath. Both processes are dependent on the chemical composition of the substrate and on the dipping parameters, such as sample temperature previous to dipping, dipping time and cooling rate after dipping, as discussed in [5]. Figure 3.1 and 3.2 show the appearance of the coating after the hot dipping and fast cooling (450 °C/min) and slow cooling (30 °C/min), respectively. The chemical composition of the different intermetallic layers can be found in Table 3.1. In the fast cooled samples the first layer in contact with the substrate is τ_1. A very irregular τ_4 layer grows over the τ_1. Finally, the external layer is an eutectic Al-Si matrix in which there can be found pure Si areas. Additional layers like Fe_3Si, ηFe_2Al_5 and τ_2-τ_3 appear in the slow cooled samples. The samples were dipped during 5s to avoid mass loss by dissolution in the molten bath. The chemical composition through the coating thickness can be found in Fig. 3.3. The diffusion path followed during the coating formation at 800°C is depicted in Fig. 3.4.

3.2.3 Diffusion annealing

The main diffusion annealing parameters are time and temperature. They will determine not only the final concentration profiles of Si and Al but also the resulting texture of the material and therefore its mechanical and magnetic properties. For the diffusion experiments substrates containing 3 wt%Si were chosen in order to ensure ferritic phase in all the range of temperatures. The

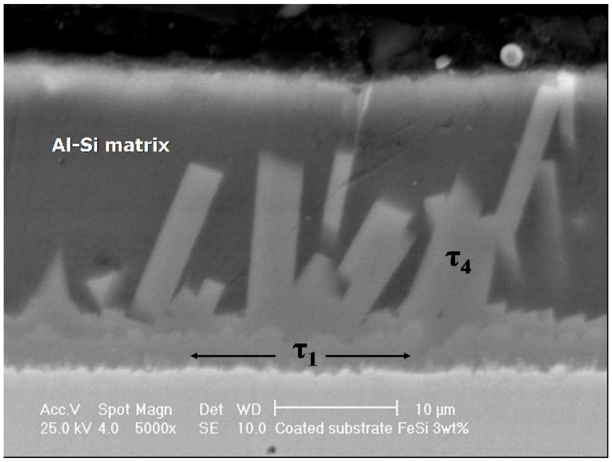

Al-Si matrix
τ_4
τ_1
Acc.V Spot Magn Det WD 10 µm
25.0 kV 4.0 5000x SE 10.0 Coated substrate FeSi 3wt%

Figure 3.1: Coating for FeSi 3wt%, fastly cooled.

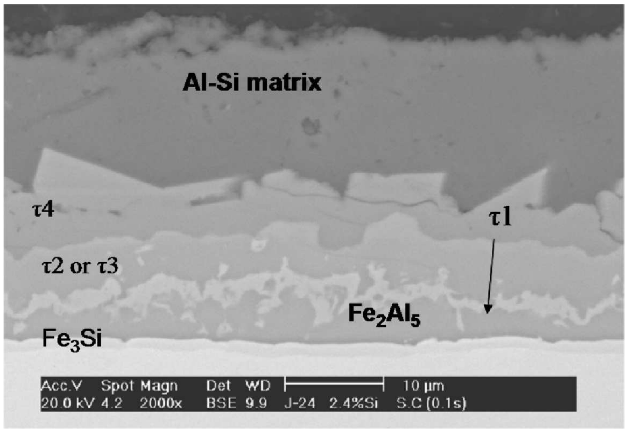

Al-Si matrix
τ4
τ1
τ2 or τ3
Fe_2Al_5
Fe_3Si
Acc.V Spot Magn Det WD 10 µm
20.0 kV 4.2 2000x BSE 9.9 J-24 2.4%Si S.C.(0.1s)

Figure 3.2: Coating for FeSi 2.4wt%, slowly cooled.

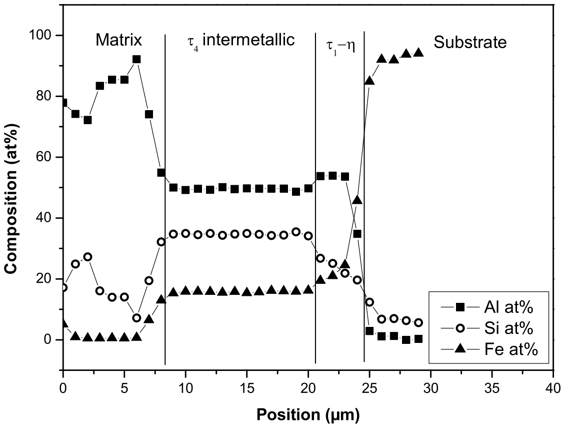


Figure 3.3: Composition through the thickness of the coating for FeSi 3wt%, dipped during 5s.

Phase	Composition	Theoretical at%			Measured at%		
		Al	Fe	Si	Al	Fe	Si
τ_1	$Al_{0.42}Fe_{0.39}Si_{0.19}$	42	39	19	39.2	34.3	26.5
τ_2	$Al_{0.54}Fe_{0.26}Si_{0.20}$	54	26	20	52.3	24.4	23.3
τ_3	$Al_{0.50}Fe_{0.25}Si_{0.25}$	50	25	25	52.3	24.4	23.3
τ_4	$Al_{0.48}Fe_{0.15}Si_{0.37}$	48	15	37	50.2	16.6	33.2
τ_9	$Al_{0.36}Fe_{0.36}Si_{0.28}$	36	36	28	39.2	34.3	26.5
η	Fe_2Al_5	0	71	29	0	69.4	30.6
$D0_3(\beta_1)$	Fe_3Si	-	5	25	-	77.50	22.50

Table 3.1: Theoretical and measured composition of the intermetallic compounds.

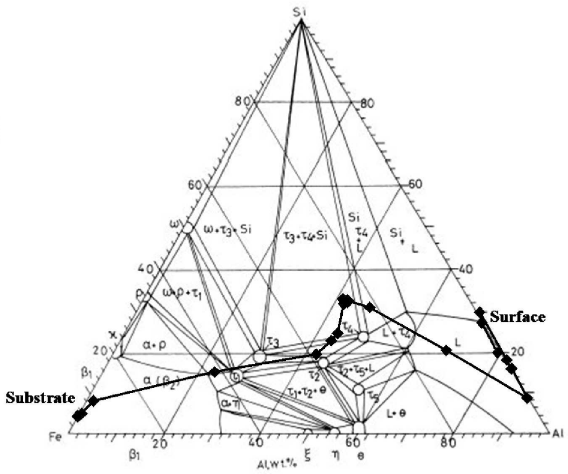


Figure 3.4: Ternary phase diagram for Fe-Al-Si at 800°C, [53], and diffusion path at 800°C during 5s dipping, fastly cooled FeSi 3wt% substrate.

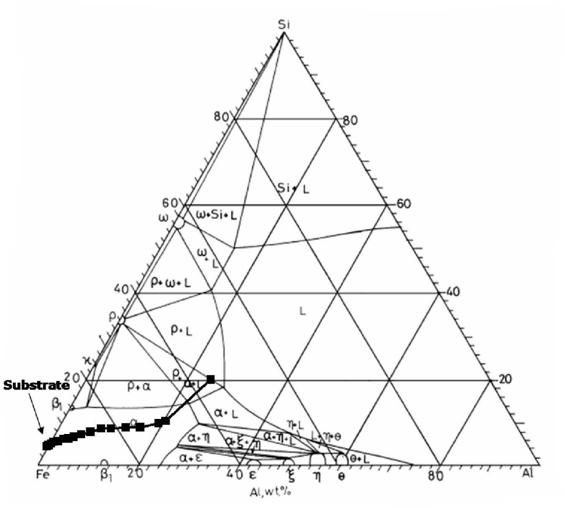


Figure 3.5: Ternary phase diagram for Fe-Al-Si at 1100°C, [53], and diffusion path at 1100°C during 5min, FeSi 3wt% substrate.

annealings were performed at 1100°C. At this annealing temperature the ternary diagram simplifies and, as shown in Fig. 3.5, diffusion takes place mainly in the α-phase. A typical diffusion profile can be seen in Fig. 3.6.

3.3 Numerical experiments

For all experiments a homogeneous Neumann BC at both edges is considered.

3.3.1 Power type diffusion

We consider diffusion of the power like form. This allows us to compare the numerical results with the Barenblatt-Pattle exact solution. In Fig. 3.7 we show a few diffusion profiles starting from a realistic initial condition with $D(C) = 0.4C^{1.59}$. An experimental Si-profile measured after 30 minutes of diffusion

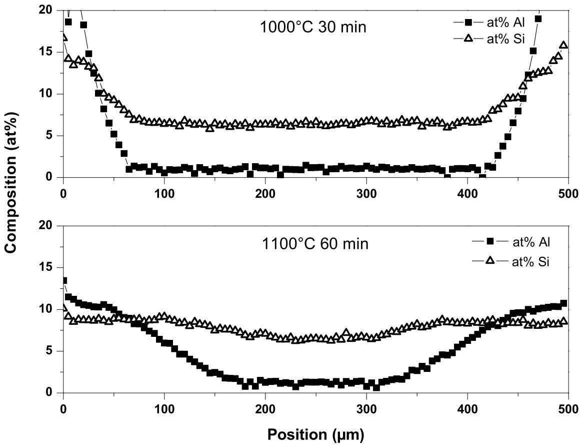


Figure 3.6: Diffusion profiles for samples annealed at 1000°C during 30 min (top) and at 1100°C during 60 min(bottom).

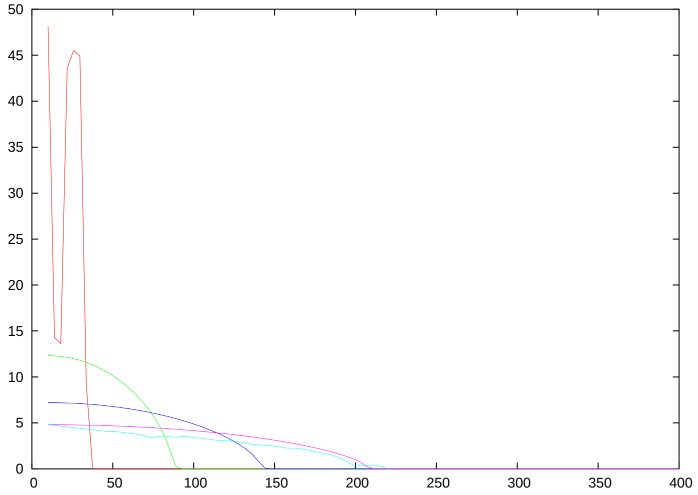


Figure 3.7: Diffusion profiles of Si (at%) into steel (μm). A typical Si-profile after hot dipping is depicted in red. The lines in green, blue and purple show the diffusion after 1, 7 and 30 minutes, respectively for $D(C) = 0.4C^{1.59}$. An experimental Si-profile taken after 30 min of diffusion annealing is depicted in light blue.

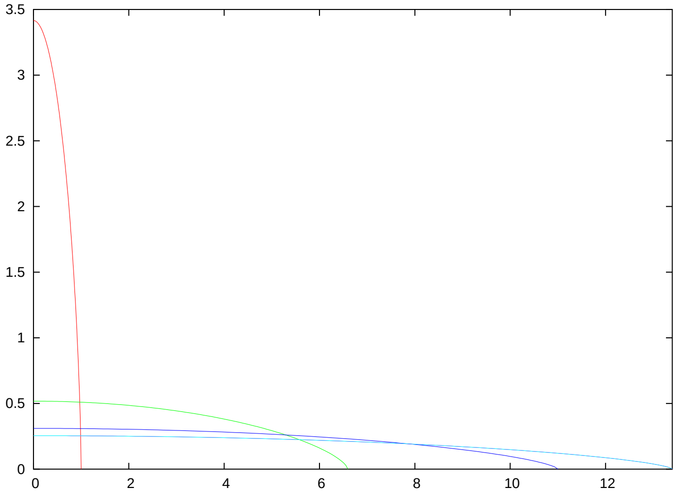


Figure 3.8: Diffusion profiles of the Barenblatt-Pattle solution with $D(C) = 2.5C^{1.5}$. The initial profile is shown in red. The other curves depict profiles after 10, 60 and 120 sec, respectively, as modeled with a nonequidistant moving grid.

annealing is also plotted. This curve matches the modeled curve very well. The model uses a uniform grid with 100 equidistant points. The first 10μm of the 400μm sample is not present as this fraction melted away during the hot dipping.

As a second example, we model the Barenblatt-Pattle problem, where the initial condition is the value of (3.16) at a time $t_0 > 0$. We decompose the domain in 2 parts: left and right of the singularity. On the right part the solution is identically zero. We can use (3.23). In Fig. 3.8 the resulting profiles are shown for the choice $D(C) = 2.5C^{1.5}$. The initial profile is taken so that the edge occurs at $x = 1$. The moving grid consists of 100 points and the grid is more dense at the edges. Compared to the exact solution, the absolute error is nowhere larger than 10^{-6}. In Fig. 3.9 the same experiment is repeated, but now with a fixed equidistant grid on the interval $(0, 13.5)$ consisting of 100 points. Here, the transformation $C = v^{1/p}$ was also performed, but now on (3.18). The largest error is at the moving front, confirming the necessity of extra datapoints

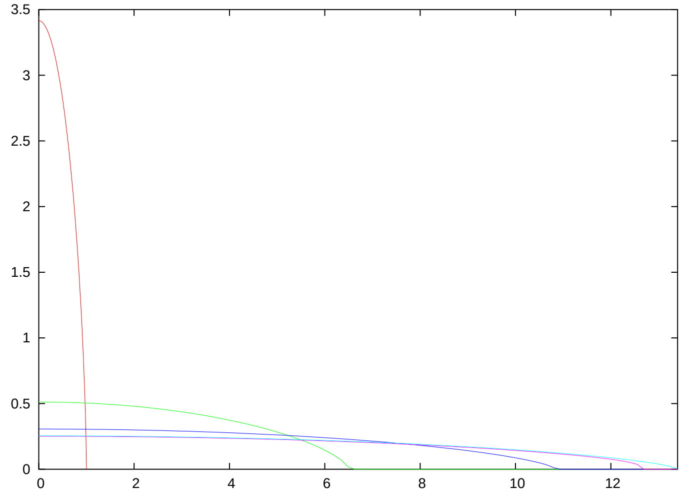


Figure 3.9: Diffusion profiles of the Barenblatt-Pattle solution with $D(C) = 2.5C^{1.5}$. The initial profile is shown in red. The other curves depict the concentration profiles after 10, 60 and 120 sec, respectively, obtained with an equidistant grid over the interval $(0, 13.5)$. For comparison, we also show the exact solution at $t = 120$sec (compare with Fig. 3.8).

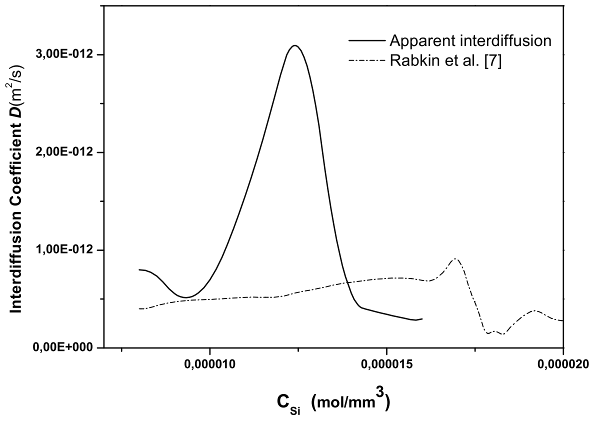


Figure 3.10: Apparent diffusion coefficient at 1100°C (full) and Fe-Si interdiffusion coefficient (dashed), in m^2/s.

around the diffusion front to increase the accuracy there.

3.3.2 B-spline diffusion

In the following numerical experiment we have used 41 gridpoints x_i over the interval $(0\mu\text{m}, 500\mu\text{m})$. The diffusion coefficient is a B-spline through 8 point couples (C_k, D_k), $k = 1, \ldots, 8$ where $C_1 = 0.8e - 5$ and $C_8 = 1.7e - 5$. This B-spline interpolant is given in Fig. 3.10, together with the values of a Fe-Si interdiffusion ((0 at%Al)) taken from [66]. As initial condition for the concentration we have used the experimental data after 5 min of diffusion annealing.

The result of the model with this diffusion coefficient is given in Fig. 3.11 for the 3 timesteps for which we have experimental data. The modeled curves are good approximations of the expermiments except for the 3h-curves. However, for the longest annealing times we detected mass loss specifically of Al. This phenomen is not included in the model and it probably explains the deviation. More experimental work will be needed to quantify this mass loss and decide

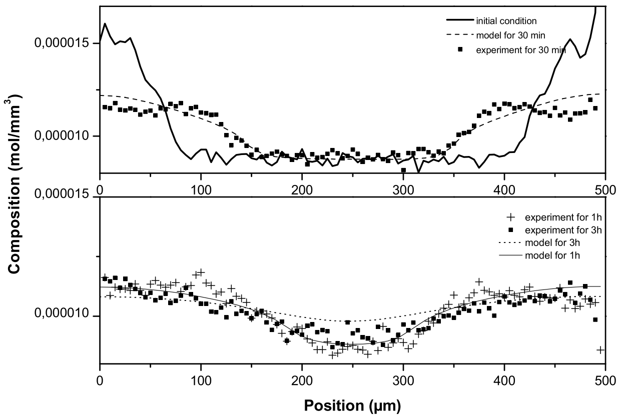


Figure 3.11: Experimental and modeled concentration profiles for 5 min, 30 min, 1h and 3h, respectively, at 1100°C.

how to adapt the model to incorporate this late-time behaviour.

Part II

Inverse Problems

In Part I several mathematical models were presented to solve practical engineering problems. We call such methods **direct problems** or also **forward problems**: given the mathematical model and the necessary data, the problem can be solved. However, in constructing the model, many parameters are used, see for example the dispersivities in (B.13) or the sorption constants in (B.18)-(B.21). These parameters need to be known to use the model. In contrast, **inverse problems** arise when a mathematical model is used to recover the parameters. To this aim, experimental results need to be known.

Inverse problems can also be utilized to determine the optimal way of controlling a process so to get a desired result. In this setup the parameters may be time dependent and can be changed by human intervention, like e.g. pressure or temperature.

The most straithforward way of estimating parameters is curve fitting. The experiment is plotted as a curve, and the direct problem is solved several times to obtain benchmark curves with several parameters. Comparing experimental data with the benchmark, the parameter values are estimated. It is clear that this is only possible if there are few parameters, and the model is stable. Stability here means that a small perturbation in the experimental values, does not lead to very different values of the recovered parameters.

We will present inverse problems for the methods developed in Part I. Our focus will be on the *adjoint method*, also called *costate method.* In Chapter 4 we apply this to diffusion annealing. In Chapter 5 we consider the dual-well problem. For an overview of basic results on numerical methods that can be used to solve inverse problems, and for some background information on inverse methods in general, we refer to Appendix C.

Chapter 4

Inverse problems in annealing diffusion

In Chapter 3 annealing diffusion in a ternary alloy was solved. Briefly summurizing, the diffusion was modeled by (3.6), i.e.

$$\frac{\partial C_1^3(x,t)}{\partial t} = \frac{\partial}{\partial x}\left(D_1^{23}\left(C_1^3\left(x,t\right)\right)\frac{\partial C_1^3(x,t)}{\partial x}\right), \tag{4.1}$$

for $0 < x < L$, $t > 0$. Here, D_1^{23} is the apparent diffusion coefficient of Si in Fe along a given diffusion path in the ternary alloy Si-Al-Fe .

As the diffusion coefficient D_1^{23} is unknown, it must be determined from the experiments. We consider the following general problem.

$$\partial_t C - \nabla \cdot (D(C)\ \nabla C) = 0 \quad \text{in } \Omega \times (0,T), \tag{4.2}$$

along with boundary conditions

$$C = C_0 \quad \text{on } \partial\Omega_1, \quad -D(C)\nabla C \cdot \nu = 0 \quad \text{on } \partial\Omega_2, \quad 0 < t < T, \tag{4.3}$$

and an initial condition

$$C(x,0) = C^0(x), \quad x \in \Omega. \tag{4.4}$$

Here, Ω is an open bounded domain, $\partial\Omega_1$ and $\partial\Omega_2$ are open non-overlapping parts of its Lipshitz boundary $\partial\Omega$ (such that $\partial\Omega = \overline{\partial\Omega_1} \cup \overline{\partial\Omega_2}$), $(0,T)$ is a given time interval and C^0 represents the given initial concentration profile.

Let

$$C^*(x,t) \quad \text{for } x \in \Omega,\ t \in (0,T), \tag{4.5}$$

be a 'given' function, viz. suitably constructed by interpolation of measured values at discrete space-time points (x,t). From (4.5) the function $D(C)$, $C \in (0,1)$, has to be restored. Our aim is to determine the unknown function $D(s)$, $s \in (0,1)$, so that the measured values (4.5) are well approximated by the numerical results from the corresponding direct problem.

To this end we look for a function D in a class of explicit functions parametrized by a vector $\boldsymbol{p} = (p_1, \ldots, p_m)$. More explicitely, we take $D = D(s,\boldsymbol{p})$, where $s \in (0,1)$ and $\boldsymbol{p} \in U_{\text{ad}} \subset \mathbf{R}^m$, with U_{ad} an admissible compact subset of $\mathbb{R}^m$. Possible choices for $D(s,\boldsymbol{p})$ are

- Power law
$$D(s,\boldsymbol{p}) = p_1 s^{p_2}(1 + p_3 s + p_4 s). \tag{4.6}$$

- Linear interpolation
$$D(s,\boldsymbol{p}) = p_i + \frac{p_{i+1} - p_i}{h}(s - ih) \quad \text{for } s \in (ih,(i+1)h), \quad i = 0, \ldots, m-1, \tag{4.7}$$
with $h = \frac{1}{m}$.

- B-spline interpolation
$$\begin{aligned} D(s,\boldsymbol{p}) \quad = \quad & \text{value at } s \text{ of the natural B-spline interpolant,} && (4.8)\\ & \text{through the points } (C_k, p_k),\ k = 1, \ldots, m, && \\ & \text{where the values } C_k \text{ are prescribed.} && (4.9) \end{aligned}$$

We look for an optimal vector-parameter $\widehat{\boldsymbol{p}} = (\hat{p}_1, \ldots, \hat{p}_m)$ such that the cost functional

$$\mathcal{F}(\boldsymbol{p}) \equiv \mathcal{F}(C,\boldsymbol{p}) = \int_0^T \int_\Omega [C(x,t,\boldsymbol{p}) - C^*(x,t)]^2 \, dx \, dt, \tag{4.10}$$

attains its minimum on U_{ad} at $\boldsymbol{p} = \widehat{\boldsymbol{p}}$. Here, $C(x,t,\boldsymbol{p})$ is the solution of (4.2)-(4.4) with D as in (4.6)-(4.8). The vector $\widehat{\boldsymbol{p}}$ is obtained as the limit of a sequence $\{\mathbf{p}_k\}_{k=1}^\infty$ such that $\mathcal{F}(\mathbf{p}_{k+1}, C_{k+1}) < \mathcal{F}(\mathbf{p}_k, C_k)$ and $\mathcal{F}(\mathbf{p}_k, C_k) \to \mathcal{F}(\widehat{\mathbf{p}}, \widehat{C})$, where $C_k = C(x,t,\mathbf{p}_k)$.

The results of this chapter have been published in [3, 40, 56].

4.1 The adjoint problem

4.1.1 Deduction

We deduce the adjoint problem, see Section s:adcome, for Eq. (4.2)-(4.4), with cost functional (4.10). We have the following result.

Theorem 4.1.1. *Let $C(x,t,\boldsymbol{p})$ be the solution of (4.2)-(4.4), where $D(s,\boldsymbol{p})$ is a smooth function, and let $\mathcal{F}$ be defined by (4.10). Let $\overline{\psi}(x,\tau)$ be the solution of the following convection-diffusion equation*

$$\begin{aligned}\partial_\tau\overline{\psi} - \nabla\cdot(D(C(x,T-\tau,\boldsymbol{p}))\nabla\overline{\psi}(x,\tau)) \\ + D'_s(C(x,T-\tau,\boldsymbol{p}))\nabla C(x,T-\tau,\boldsymbol{p})\cdot\nabla\overline{\psi}(x,\tau) = \\ = -2\left(C(x,T-\tau,\boldsymbol{p}) - C^*(x,T-\tau)\right)\end{aligned} \quad (4.11)$$

where $D'_s = \frac{dD(s)}{ds}$, along with boundary and initial conditions given by

$$\begin{aligned}\overline{\psi}(x,\tau) = 0 \quad \textit{for } x\in\partial\Omega_1,\\ D(C,\boldsymbol{p})\nabla\overline{\psi}(x,\tau)\cdot\nu = 0 \quad \textit{for } x\in\partial\Omega_2,\\ \overline{\psi}(x,0) = 0.\end{aligned} \quad (4.12)$$

Then,

$$\nabla_{\boldsymbol{p}}\mathcal{F}(\boldsymbol{p}) = \int_0^T\int_\Omega \nabla_p D(C(x,t,\boldsymbol{p}))\left(\nabla C(x,t,\boldsymbol{p})\cdot\nabla\psi(x,t)\right)\,dx\,dt, \quad (4.13)$$

where

$$\psi(x,t) = \overline{\psi}(x,T-t). \quad (4.14)$$

Proof. Let the perturbed value $\boldsymbol{p}+\boldsymbol{\delta p}$, $(\boldsymbol{\delta p}\equiv(\delta p_1,\ldots,\delta p_m))$, give rise to the solution $C+\delta C$. We are looking for a linear mapping (i.e. Gâteaux differential) between $\delta\mathcal{F} := \mathcal{F}(\boldsymbol{p}+\boldsymbol{\delta p}, C+\delta C) - \mathcal{F}(\boldsymbol{p},C)$ and $\boldsymbol{\delta p}$. Neglecting the second order terms of $\boldsymbol{\delta p}$ we obtain

$$\delta\mathcal{F} = 2\int_0^T\int_\Omega \delta C\left(C(\mathbf{p}) - C^*(x,t)\right)\,dx\,dt. \quad (4.15)$$

To eliminate δC from (4.15) we derive a boundary value problem *in variations* in the following way. Since $C+\delta C$ is the solution of (4.2)-(4.4) corresponding to $\boldsymbol{p}+\boldsymbol{\delta p}$, we have

$$\partial_t(C+\delta C) - \nabla\cdot(D(C+\delta C,\boldsymbol{p}+\boldsymbol{\delta p})\ \nabla(C+\delta C)) = 0 \quad \text{in } \Omega\times(0,T), \quad (4.16)$$

along with boundary conditions

$$C+\delta C = C_0 \quad \text{on } \partial\Omega_1, \quad -D(C+\delta C, \boldsymbol{p}+\boldsymbol{\delta p})\nabla(C+\delta C)\cdot\nu = 0 \quad \text{on } \partial\Omega_2, \tag{4.17}$$

and initial condition

$$C(x,0) + \delta C(x,0) = C^0(x). \tag{4.18}$$

Substracting 'equationwise' (4.2)-(4.4) from (4.16)-(4.18) and neglecting again second order terms, we find the equation *in variations*

$$\partial_t(\delta C) - \nabla\cdot[D(C,\boldsymbol{p})\ \nabla\delta C + D_s'(C,\boldsymbol{p})\delta C\nabla C + (\nabla_p D(C,\boldsymbol{p})\cdot\boldsymbol{\delta p})\nabla C] = 0 \tag{4.19}$$

in $\Omega\times(0,T)$, along with boundary conditions

$$\begin{gathered} \delta C = 0 \quad \text{on } \partial\Omega_1, \\ [D(C,\boldsymbol{p})\nabla(\delta C) + D_s'(C,\boldsymbol{p})\delta C\nabla C + (\nabla_p D(C,\boldsymbol{p})\cdot\boldsymbol{\delta p})\nabla C]\cdot\nu = 0 \quad \text{on } \partial\Omega_2, \end{gathered} \tag{4.20}$$

and initial condition

$$\delta C(x,0) = 0. \tag{4.21}$$

Multiplying (4.19) by a smooth function $\psi(x,t)$, to be specified below, integrating over $\Omega\times(0,T)$ and using (4.20) and (4.21), we get

$$\begin{aligned} &\int_\Omega \delta C\psi\,dx\Big|_0^T - \int_0^T\int_\Omega \delta C\partial_t\psi\,dx\,dt \\ &- \int_0^T\int_{\Omega_1} [D(C,\boldsymbol{p})\nabla(\delta C) + D_s'(C,\boldsymbol{p})\nabla C\delta C + (\nabla_p D(C,\boldsymbol{p})\cdot\boldsymbol{\delta p})\nabla C]\cdot\nu\,\psi\,dx\,dt \\ &\qquad + \int_0^T\int_\Omega \Big[\underbrace{D(C,\boldsymbol{p})(\nabla(\delta C)\cdot\nabla\psi)}_{(I)} + D_s'(C,\boldsymbol{p})(\nabla C\cdot\nabla\psi)\delta C \\ &\qquad\qquad + (\nabla_p D(C,\boldsymbol{p})\cdot\boldsymbol{\delta p})(\nabla C\cdot\nabla\psi)\Big]\,dx\,dt = 0. \end{aligned} \tag{4.22}$$

We restrict the auxiliary function $\psi(x,t)$ so as to obey the conditions

$$\psi(x,t) = 0 \quad \text{for } x\in\partial\Omega_1, \quad D(C,\boldsymbol{p})\nabla\psi(x,t)\cdot\nu = 0 \quad \text{for } x\in\partial\Omega_2, \quad \psi(x,T) = 0. \tag{4.23}$$

Integration by parts of the term (I) in (4.22) leads to

$$-\int_0^T \int_\Omega \delta C \left[\partial_t \psi + \nabla \cdot \big(D(C,\boldsymbol{p})\nabla\psi\big) - D_s'(C,\boldsymbol{p})\big(\nabla C \cdot \nabla\psi\big)\right] \, dx \, dt$$
$$+ \int_0^T \int_\Omega \nabla_p D(C,\boldsymbol{p})(\nabla C \cdot \nabla\psi) \, dx \, dt \cdot \boldsymbol{\delta p} = 0. \qquad (4.24)$$

Now, the function $\psi(x,t)$ can be chosen in a unique way from the requirement that the parabolic equation

$$\partial_t \psi + \nabla \cdot (D(C,\boldsymbol{p})\nabla\psi) - D_s'(C,\boldsymbol{p})\nabla C \cdot \nabla\psi = 2\left(C(\boldsymbol{p}) - C^*(x,t)\right), \qquad (4.25)$$

is satisfied together with (4.23). For this choice of ψ, the relations (4.15) and (4.24) yield

$$\delta\mathcal{F} = \int_0^T \int_\Omega \nabla_p D(C,\boldsymbol{p})\nabla C \cdot \nabla\psi \, dx \, dt \; \boldsymbol{\delta p}. \qquad (4.26)$$

By putting $\tau = T - t$, the problem consisting of (4.23) and (4.25) is reduced to the parabolic problem (4.11)-(4.12).

From the expression (4.26) we obtain the gradient of the cost functional as in (4.13). □

Remark 4.1.1. *The proof above has to be adapted in the case of a moving interface. In this case the solution $C(x,t)$ has only C^2-regularity up to the interface; at the interface the original* PDE *is only fulfilled from the left. Therefore, integration by parts can only be done up to the interface of C, respectively $C + \delta C$. However, the border terms still drop out as there is no flux over the interface, and the other terms can be extended over the entire domain. The following is important to note in the interface problem:*

- *BC (4.12) at the symmetry boundary only makes sense when the interface has reached the boundary. Otherwise, this BC is superfluous. This is consistent with (4.12) as we have a pure reaction problem when the boundary is not reached yet by the interface (which, of course, makes satisfying the BC impossible).*

- *The function $\overline{\psi}(x,\tau)$ used in the proof also only needs to be smooth up to the interface, as integration by parts in space is only done up to this point. We comment on this later.*

4.1.2 Numerical approximation

Equation (4.11) is a convection-diffusion equation with respect to $\overline{\psi}$ and can be solved with appropriate methods. For simplicity, we chose again the method of lines, which was used for the direct problem, see (3.18) and (3.21).

We follow the ansatz as in the 1D problem with moving interface of Chapter 3. To solve (4.11), we first solve the direct problem, and subsequently the adjoint system (4.11)-(4.12) for $\overline{\psi}$. By the same Landau's transformation and space discretizations the values $C_i^1 = C^1(y_i, T-\tau)$ and $C_j^2 = C^2(y_j, T-\tau)$ can be used to construct values for $D(C, \boldsymbol{p})$ and C. Note that the profiles of $\overline{\psi}$ and C will be, generally, qualitatively different, hence the nonequidistant grid will not necessarely lead to a better approximation.

Denote $F_i^1(\tau) \approx \overline{\psi^1}(y_i, \tau)$ on $\Omega_1(\tau)$, and similarly $F_j^2(\tau) \approx \overline{\psi^2}(y_j, \tau)$ on $\Omega_2(\tau)$. Next, denote by $p_2^k(y, i)$ the second order Lagrange polynomial interpolating these values in the points (y_{i-1}, F_{i-1}^k), (y_i, F_i^k) and (y_{i+1}, F_{i+1}^k) for $k = 1$ and 2. Furthermore, set $C_i^* \approx C^*(y_i, T-\tau)$. Two ODE-systems are obtained:

$$\frac{d}{d\tau}F_i^1(\tau) - \frac{1}{s^2(T-\tau)}D(C_i^1)\frac{d^2}{dy^2}p_2^1(y_i, i) + y_i\frac{\dot{s}(T-\tau)}{s(T-\tau)}\frac{d}{dy}p_2^1(y_i, i) = -2(C_i^1 - C_i^*), \tag{4.27}$$

for $i = 1, \ldots, N-1$, and also for $i = 0$ in the case of a Neumann BC, and

$$\begin{aligned} \frac{d}{d\tau}F_j^2(\tau) - \frac{1}{(X - s(T-\tau))^2}D(C_j^2)\frac{d^2}{dy^2}p_2^2(y_j, j) \\ - y_j\frac{\dot{s}(T-\tau)}{X - s(T-\tau)}\frac{d}{dy}p_2^2(y_j, j) = -2(C_j^1 - C_j^*), \end{aligned} \tag{4.28}$$

for $j = 1, \ldots, M-1$, and also for $j = 0$ in the case of a Neumann BC.

The interface equation is the same as in the direct problem, (3.22), allowing the elimination of $F_N^1 = F_M^2$ from (4.27)-(4.28). From the solution we only need to keep track of the values $d_1F_i^1 = \frac{d}{dy}p_2^1(y_i, i)$ and $d_1F_j^2 = \frac{d}{dy}p_2^2(y_j, j)$ at equidistant time values $t = t_1, \ldots, t_m$.

4.1.3 Computation of $\nabla_p\mathcal{F}$

The analytic form of $\nabla_p\mathcal{F}$ is given by (4.13). This expression will be approximated using the numerical values of C, ∇C and $\nabla\psi$, obtained in the previous sections. We emphasize that the values for C and ψ are calculated in the same grid points x_i and the same time points t_m, the latter being equidistant with time step Δt. Then, considering $\psi(x,t) = \overline{\psi}(x,\tau) = \overline{\psi}(x,T-t)$, we obtain

$$\nabla_p\mathcal{F}(\boldsymbol{p}) \approx \sum_{k=0}^{m-1} \Delta t \left[\sum_{i=0}^{N} \frac{(\alpha_i^1 + \alpha_{i+1}^1)}{2} \nabla_p D(C_i^1(k)) d_1 C_i^1(k) d_1 F_i^1(m-k)) \frac{1}{s(t_k)} \right.$$
$$\left. + \sum_{j=0}^{M} \frac{(\alpha_j^2 + \alpha_{j+1}^2)}{2} \nabla_p D(C_j^2(k)) d_1 C_j^2(k) d_1 F_j^2(m-k)) \frac{-1}{X - s(t_k)} \right], \quad (4.29)$$

where $\alpha_0^1 = \alpha_{N+1}^1 = 0 = \alpha_0^2 = \alpha_{M+1}^2$.

4.2 Convergence for $\partial_\tau\psi(x,\tau) - a(x,\tau)\partial_x^2\psi(x,\tau) = f(x,\tau)$, a being degenerate

Equation (4.11) is a degenerate convection-diffusion pure reaction problem in the case of zero substrate problem, where we have a moving interface with on the right of it $C = 0$, so that $D(C) = 0$. In this specific case it is not a priori clear wether the system (4.11)-(4.12) has a solution. In this Section we shall prove the existence of a solution.

The one-dimentional differential equation (4.11) has the form

$$\partial_t\psi(x,t) + a(x,t)\partial_x^2\psi(x,t) = f(x,t), \qquad (4.30)$$

along with boundary conditions

$$\partial_x\psi(0,t) = 0, \quad a(L,t)\partial_x\psi(L,t) = 0, \qquad (4.31)$$

and stopping condition

$$\partial_x\psi(x,T) = 0, \qquad (4.32)$$

where $a(x,t)$ $(= D(C,\boldsymbol{p}))$ and $f(x,t)$ $(= 2\,(C(\boldsymbol{p}) - C^*(x,t)))$, are known functions.

The original problem, $C(x,t)$, has a moving front at $s(t)$, $s(0) = s_0 > 0$, of which it is known from porous media equations that it starts to move at $t = T_1 \geq 0$, and afterwards is strictly increasing. Due to the behaviour of C and the degeneracy of D, we have that the function $a(x,t)$ has in $(s_0, s(T))$ the same moving support as C. Moreover, if $C(x_0,t_0) > 0$, then the solution is C^∞-smooth in a neighbourhood of x_0 for $t > t_0$. On the other hand, $\nabla a = D'(C)\nabla C$ which can become undefined at the interface.

This can be explained as follows. If we consider the Barenblatt-Pattle solution, (3.16), we have that, $D'(C)$ is 0 (for $p_2 > 1$), or ∞ (for $0 < p_2 < 1$). The second par of ∇at, ∇C depends on the solution. For the Barenplatt-Pattle solution $\nabla C = -\infty$ for $x \rightarrow s(t)$, (for $p_2 > 1$) or $\nabla C = 0$ for $x \rightarrow s(t)$ (for $0 < p_2 < 1$), and we get always $\nabla a = c$, with $-\infty < c < 0$. In general, although the solution is smooth up to the interface, and the speed of the interface $\dot{s}(t)$ is finite, it need not be the case that ∇a is bounded. Consider for example the PDE, [65],

$$\partial_t w = \partial_x \left[(aw^2 - bw)\partial_x w\right],$$

which has the traveling-wave solution

$$w = \sqrt{2c_1 x + 2ac_1^2 t + c_2},$$

and the moving edge

$$s(t) = -\frac{2ac_1^2 t + c_2}{2c_1}.$$

In this case, $D(w) = aw^2 - bw$ and $D'(w)$ is nonzero at the interface $w = 0$, whereas $\partial_x w|_{w=0} = \text{sign}(c_1)\infty$. Therefore, $\nabla a = -\infty$, and we conclude that in general we cannot assume regularity of a. This degeneracy does not allow the use of energy type a priori estimates (no integration by parts), so one of the main tools in the analysis of existence is not applicable.

To summarize, we can assume the following for a:

$$\begin{cases} 1. \quad a(x,t) \text{ has strictly increasing support up to } s(t), \\ \qquad\qquad 0 < s_0 \leq s(t) \leq s(T) < L, \quad 0 \leq \dot{s} \leq K \\ 2. \quad a(x,t) = 0 \text{ for } x \geq s(t), \\ 3. \quad \partial_x a(s(t),t) = c(t), \text{ with } -\infty \leq c(t) < 0. \\ \qquad \text{In the case } c(t) = -\infty, \text{ we call the problem degenerate.} \end{cases} \tag{4.33}$$

Here, K is a constant. For simplicity we take $s(T) < L$. In the case we need to model up to times $s(T) \geq L$, we can use the results given here up to $s(t_e) = L$, and then use the known results for diffusion-reaction problems to obtain results for $t > t_e$.

The numerical discretization given in Section 4.1.2 is “impractical” for a convergence proof. We start therefore with a more suitable discretization based on the same principles.

4.2.1 Discretization

As in Section 4.1.2 the domain with moving interface at $x = s(t)$ is transformed to two fixed domains by Landau's mapping. Eq. (4.30) splits then in two convection-diffusion equations:

$$\partial_\tau \widetilde{\psi}_I + y\frac{\dot{s}(T-\tau)}{s(T-\tau)}\partial_y \widetilde{\psi}_I - \frac{\widetilde{a}(y,\tau)}{s^2(T-\tau)}\partial_y^2 \widetilde{\psi}_I = -\widetilde{f}_I(y,\tau), \quad y \in (0,1), \tag{4.34}$$

along with boundary and initial conditions

$$\partial_y \widetilde{\psi}_I(0,\tau) = 0, \quad \widetilde{\psi}_I(y,0) = 0, \tag{4.35}$$

and

$$\partial_\tau \widetilde{\psi}_{II} - y\frac{\dot{s}(T-\tau)}{L - s(T-\tau)}\partial_y \widetilde{\psi}_{II} = -\widetilde{f}_{II}(y,\tau), \quad y \in (0,1), \tag{4.36}$$

along with boundary and initial conditions

$$\partial_y \widetilde{\psi}_{II}(0,\tau) = 0, \quad \widetilde{\psi}_{II}(y,0) = 0, \tag{4.37}$$

and along with the continuity condition

$$\widetilde{\psi}_I(1,\tau) = \widetilde{\psi}_{II}(1,\tau). \tag{4.38}$$

We used $\tau = T - t$ and the fact that $a(x,t) = 0$ when $x > s(t)$. Here $\widetilde{a}(y,\tau) = a(ys(T-\tau), T-\tau)$, and analogously for $\widetilde{f}_I(y,\tau)$, and $\widetilde{f}_{II}(y,\tau)$.

Remark 4.2.1. *The use of a continuity condition between the two equations at $y = 1$ is formal. We cannot guarantee continuity. Therefore, in our approximation scheme we shall always regularize the function $a(x,t)$ locally at $x = s(t)$ by a_Δ, so that the continuity of solution at $x = s(t)$ is assured. This is called* ***Δ-regularization****. In the limiting process (discretization parameters converging to zero) we cannot guarantee this property. The a priori estimates which we will deduce later don't exclude the creation of a shock at $x = s(t)$ which can depend on the order of the degeneracy of a (i.e. $\partial_x a(s(t),t) = -\infty$), the speed $\dot{s}(t)$ and the regularity property of f in the neighbourhood of $x = s(t)$. For this reason, the continuity $\psi_I(1,\tau) = \psi_{II}(1,\tau)$ is only considered formally, and we include it into the approximation scheme, proving convergence to the original equation. However, we will not consider this continuity for the definition of our variational solution. See also Remark 4.1.1.*

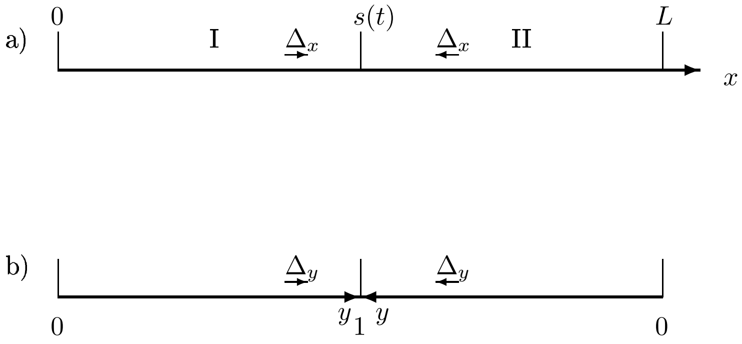


Figure 4.1: a) Domain decomposition in two parts: $(0, L) = (0, s(t)) \cup (s(t), L)$. b) Mapping to the fixed uniform y interval.

Remark 4.2.2. *The BC (4.37) in $y = 0$ is formal, as (4.37) only consists of pure reaction in $y = 0$ and the flow is towards $y = 0$ from the interior of the domain. See also Remark 4.1.1*

We regularize the function a by a_Δ so that $a_\Delta \to a$ if $\Delta \to 0$, the Δ-**regularization**. Its exact definition will be given in Section 4.2.4. It suffices to state that as a consequence we can use the continuity of the solution at $x = s(t)$, and we need to provide the extra terms arising from this a_Δ in the numerical approximation of (4.36). After regularization, (4.36) reads as

$$\partial_\tau \widetilde{\psi}_{II} - y\frac{\dot{s}(T-\tau)}{L-s(T-\tau)}\partial_y \widetilde{\psi}_{II} - \frac{\widetilde{a}_\Delta(y,\tau)}{(L-s(T-\tau))^2}\partial_y^2 \widetilde{\psi}_{II} = -\widetilde{f}_{II}(y,\tau). \tag{4.39}$$

We approximate (4.34) and (4.36/4.39) by using central difference in the elliptic part and *upwind* type differences in the convective terms. For simplicity, we consider a uniform space discretization $\{y_i\}_{i=0}^M$, $\Delta_y = y_i - y_{i-1}$, $\forall i = 1, \ldots, M$. The fixed nodal point y_i corresponds to the moving grid point $x_i(t) = y_i s(t)$ in the x variable in part I of the domain, and to $x_i(t) = L - y_i(L - s(t))$ in part II. We have that $\Delta_x(t) = x_i(t) - x_{i-1}(t) = \Delta_y s(t)$ in part I, and $\Delta_x(t) = x_i(t) - x_{i-1}(t) = -\Delta_y(L - s(t))$ in part II, see Fig. 4.1.

Let $V_i(\tau) \approx \widetilde{\psi}_I(y_i, \tau)$, $\forall i = 0, \ldots, M$, and $V_{M+i}(\tau) := W_{M-i}$, where $W_i \approx \widetilde{\psi}_{II}(y_i, \tau)$, $\forall i = 0, \ldots, M$, see Fig. 4.2. Then our approximation scheme is of

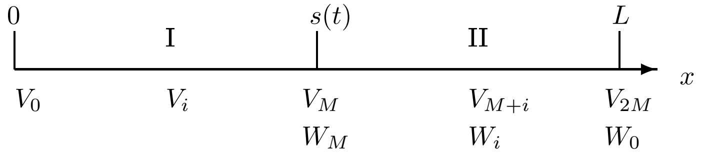


Figure 4.2: Numerical approximation on the space grid.

the form

$$\dot{V}_i + \frac{\dot{s}}{s} y_i \delta^- V_i - \frac{a_i}{s^2 \Delta_y}(\delta^+ V_i - \delta^- V_i) = f_i, \quad i = 1, \ldots, M-1, \tag{4.40}$$

where

$$\delta^- V_i = \frac{V_i - V_{i-1}}{\Delta_y}, \quad \delta^+ V_i = \frac{V_{i+1} - V_i}{\Delta_y},$$

and

$$\dot{V}_{M+i} + \frac{\dot{s}}{L-s} y_{M-i} \delta^- V_{M+i} - \frac{a_{M+i}}{(L-s)^2 \Delta_y}(\delta^+ V_{M+i} - \delta^- V_{M+i}) = f_{M+i}, \tag{4.41}$$

$i = 1, \ldots, M-1$, where we used shorthand notation $\dot{s} = \dot{s}(T-\tau)$, $s = s(T-\tau)$, and

$$a_i = \widetilde{a}_\Delta(y_i, \tau), \quad \text{in Part I,}$$
$$a_{M+i} = \widetilde{a}_\Delta(y_{M-i}, \tau), \quad \text{in Part II,}$$

and

$$f_i = -\widetilde{f}_I(y_i, \tau), \quad f_{M+i} = -\widetilde{f}_{II}(y_{M-i}, \tau), \qquad i = 1, \ldots, M-1.$$

Eq. (4.41) is obtained by *upwind* discretization of (4.39) in terms of W:

$$\dot{W}_i - \frac{\dot{s}}{L-s} y_i \delta^+ W_i - \frac{a_i^*}{(L-s)^2 \Delta_y}(\delta^+ W_i - \delta^- W_i) = f_i^*, \quad i = 1, \ldots, M-1, \tag{4.42}$$

with $a_i^* = a_{M+i}$ and $f_i^* = f_{M+i}$. The values of V_0 and V_{2M} follow from the boundary condition. The approximation of (4.30) in the point $x = s(t)$ follows by using upwinding for the convection part of (4.34) and (4.39), and by determining $\partial_x^2 \psi$ from the approximate values V at the nonuniformily spaced points x_M,

x_{M-1} and x_{M+1}, corresponding with y_M and y_{M-1} in part I ($x_{M-1} \leftrightarrow y_{M-1}$) and part II ($x_{M+1} \leftrightarrow y_{M-1}$). This gives

$$\dot{V}_M + \frac{\dot{s}}{s}\delta^- V_M - \frac{2a_M}{L\Delta_y}\left(\frac{1}{L-s}\delta^+ V_M - \frac{1}{s}\delta^- V_M\right) = f_M, \tag{4.43}$$

where we have used $\frac{1}{2}\Delta_y s + \frac{1}{2}\Delta_y(L-s) = \frac{1}{2}L\Delta_y$. Now (4.40)-(4.43) represent the discretization of (4.30) with respect to moving nodal points, using central differences for ∂_x^2 and upwind for the corresponding convective term which arises from the moving gridpoints.

For the time discretization we use an implicit Euler scheme. Consider the time point $\tau_n = nh$, $n = 1, \dots, N_h$, $N_h h = T$, and denote $V_j^n \approx V_j(\tau_n)$.

We introduce the standard Rothe functions. Define the time continuous functions

$$V_{i,h}(\tau) := V_i^n + \frac{\tau - \tau_n}{h}\left(V_i^{n+1} - V_i^n\right), \tag{4.44}$$

$\forall \tau \in (\tau_n, \tau_{n+1})$, $\forall n = 0, 1, \dots, N_h$, $\forall i = 0, 1, \dots 2M$.

In agreement with the method of lines, the limit function when $h \to 0$ is considered first, and it is shown that

$$V_{h,i}(\tau) \to V_i(\tau), \text{ uniformly on } (0,T). \tag{4.45}$$

To distinguish the two considered domains, we formally define

$$W_{M-i}(\tau) := V_{M+i}(\tau), \quad \text{with} \quad i = 0, \dots, M. \tag{4.46}$$

Define also,

$$V^\Delta(y,\tau) := V_{i-1,h}(\tau) + \frac{y - y_{i-1}}{\Delta_y}\left(V_{i,h}(\tau) - V_{i-1,h}(\tau)\right), \tag{4.47}$$

$$W^\Delta(y,\tau) := W_{i-1}(\tau) + \frac{y - y_{i-1}}{\Delta_y}\left(W_i(\tau) - W_{i-1}(\tau)\right), \tag{4.48}$$

$\forall y \in (y_{i-1}, y_i)$, $\forall i = 1, \dots M$. We show the convergence to a limit function when $\Delta_y \to 0$,

$$V^\Delta(y,\tau)(\tau) \to V(y,\tau), \quad W^\Delta(y,\tau)(\tau) \to W(y,\tau). \tag{4.49}$$

4.2.2 Weak solution

Definition 4.2.1. *A couple $(\psi_I(y,t),\ \psi_{II}(y,t))$ is a variational solution to (4.34) and (4.36), respectively, if $\psi_I \in L_\infty(I\times(0,1))$, $\partial_y\psi \in L_2(I, L_{2,\mathrm{loc}}(0,1))$ and $TV(\psi_{II}) < \infty$, such that the identities*

$$\int_I\int_0^1 \psi_I\partial_\tau\phi\,\mathrm{d}y\mathrm{d}\tau - \int_I\int_0^1 \frac{1}{s^2}\partial_y\psi_I\partial_y\left[a(ys(T-\tau),T-\tau)\phi\right]\,\mathrm{d}y\mathrm{d}\tau \\ - \int_I\int_0^1 y\frac{\dot{s}}{s}\partial_y\psi_I\phi\,\mathrm{d}y\mathrm{d}\tau = \int_I\int_0^1 \phi f(ys(T-\tau),T-\tau)\,\mathrm{d}y\mathrm{d}\tau, \quad (4.50)$$

$$\int_I\int_0^1 \psi_{II}\partial_\tau\phi\,\mathrm{d}y\mathrm{d}\tau - \int_I\int_0^1 \frac{\dot{s}(T-\tau)}{L-s(T-\tau)}\psi_{II}\partial_y(y\phi)\,\mathrm{d}y\mathrm{d}\tau \\ = \int_I\int_0^1 \phi f(L-y(L-s(T-\tau)),T-\tau)\,\mathrm{d}y\mathrm{d}\tau, \quad (4.51)$$

hold for all $\phi \in C^\infty(I\times(0,1))$ with support $\phi \subset [0,T)\times[0,1) \equiv \tilde{Q}_T$.

Remark 4.2.3. *The continuity condition $\psi_I(1,t) = \psi_{II}(1,t)$ is not guaranteed. The functions ψ_I and ψ_{II} have a finite total variation and ψ_I is Lipschitz continuous in $y \in (0,Q)$, $\forall Q < 1$, but we are not able to prove the continuity at $y = 1$.*

Definition 4.2.2. *The function*

$$\psi(x,t) = \begin{cases} \psi_I(\frac{x}{s(t)}, T-t), & x \in (0,s(t)),\ t\in(0,T); \\ \psi_{II}(\frac{L-x}{L-s(t)}, T-t), & x\in(s(t),L),\ t\in(0,T), \end{cases} \quad (4.52)$$

is a variational solution to (4.30)-(4.32), iff $\psi_I(y,\tau)$ and $\psi_{II}(y,\tau)$ are variational solutions to (4.34) and (4.36) (see Definition 4.2.1).

Some extra assumptions on a will be needed. Let us assume there is a constant $Q < 1$ such that we have the following properties of a:

$$\max_{0\le i\le L, L\Delta_y\le Q}\left|\frac{a_i - a_{i-1}}{\Delta_y}\right| \le K(Q,\tau), \qquad \min_{0\le i\le L, L\Delta_y\le Q, \tau\in(0,T)} a_i \ge \delta(Q) > 0. \quad (4.53)$$

Our main result is:

Theorem 4.2.1. *Under the assumptions (4.33) and (4.53), and with* $f \in L_\infty(\Omega) \subset H^1(\Omega)$*, there exists a variational solution to (4.30)-(4.32) in the sense of Definition 4.2.2. The approximate solution generated by*

$$\psi^\Delta(x,t) = \begin{cases} V^\Delta(\frac{x}{s(t)}, T-t), & x \in (0, s(t)),\ t > 0; \\ W^\Delta(\frac{L-x}{L-s(t)}, T-t), & x \in (s(t), L),\ t > 0, \end{cases} \tag{4.54}$$

converges pointwise (up to a subsequence) to a variational solution to (4.30)-(4.32).

The proof of this theorem involves the same techniques of Section 2.5. Another approach would be to regularize the PDE over the entire domain (not only the interface), and use standard techniques for convection-diffusion. We will give an overview of this approach in Section 4.2.5. We now prove the necessary a priori estimates, and the TV estimates needed for Theorem 4.2.1.

4.2.3 A priori estimates

Estimates for V_i^n are given first. When there is no risk for notational confusion, we drop the superscript n in the time step.

Lemma 4.2.1. *If* $c_f = \int_0^T \max_y f(y,\tau)\,\mathrm{d}\tau < \infty$*, then there exists a constant* $C(c_f) < \infty$ *so that*

$$\max_{0 \le j \le 2M,\ n=1,\dots,N_h} |V_j^n| \le C, \tag{4.55}$$

uniformly for h $(h \le h_0)$.

If furthermore $\int_0^T \int_0^L |\partial_x f(x,t)|\,\mathrm{d}x\mathrm{d}t < \infty$*, then*

$$\sum_{i=1}^{2M-2} |Z_i^{n+1}| \le c \int_0^T \int_0^L |\partial_x f(x,t)|\,\mathrm{d}x\mathrm{d}t, \tag{4.56}$$

with $Z_i = V_{i+1} - V_i$.

Proof. Let $\max_{0 \le i \le 2M} V_i^{n+1} = V_l^{n+1}$. From (4.40)-(4.43), using an implicit Euler scheme and reordening terms, we get

$$[1 + (\gamma_l + \beta_l)h]\, V_l^{n+1} \le \max_i V_i^n + \gamma_l h V_l^{n+1} + \beta_l V_l^{n+1} + h \max_i f_i,$$

$l = 0, \dots, 2M$. Hence

$$\max_i V_i^{n+1} \le \max_i V_i^n + h \max_i f_i.$$

Similarly,

$$\min_i V_i^{n+1} \geq \min_i V_i^n + h \min_i f_i,$$

and therefore

$$\max_{0\leq i\leq 2M} \left|V_i^{n+1}\right| \leq \max_{0\leq i\leq 2M} \left|V_i^n\right| + h \max_{0\leq i\leq 2M} \left|f_i\right|. \tag{4.57}$$

This is a recurrent inequality with respect to n. Using the initial condition $V_i^0 = 0$, (4.55) follows, which proves the first assertion of the Lemma.

For the second assertion, note that the time discretization gives, after rearranging the terms:

$$\left[1 + \left(\frac{\dot{s}y_{i+1}}{s\Delta_y} + \frac{a_i}{s^2\Delta_y^2} + \frac{a_{i+1}}{s^2\Delta_y^2}\right)h\right] Z_i^{n+1} = Z_i^n + \left(\frac{\dot{s}y_i}{s\Delta_y} + \frac{a_i}{s^2\Delta_y^2}\right)hZ_{i-1}^{n+1} \\ + \frac{a_{i+1}}{s^2\Delta_y^2}hZ_{i+1}^{n+1} + h(f_{i+1} - f_i), \quad i = 1,\ldots,M-2, \tag{4.58}$$

$$\left[1 + \left(\frac{\dot{s}}{s\Delta_y} + \frac{a_{M-1}}{s^2\Delta_y^2} + \frac{2a_M}{Ls\Delta_y^2}\right)h\right] Z_{M-1}^{n+1} = Z_{M-1}^n \\ + \left(\frac{\dot{s}y_{M-1}}{s\Delta_y} + \frac{a_{M-1}}{s^2\Delta_y^2}\right)hZ_{M-2}^{n+1} + \frac{2a_M}{L(L-s)\Delta_y^2}hZ_M^{n+1} + h(f_M - f_{M-1}), \tag{4.59}$$

$$\left[1 + \left(\frac{\dot{s}y_{M-1}}{(L-s)\Delta_y} + \frac{2a_M}{L(L-s)\Delta_y^2} + \frac{a_{M+1}}{(L-s)^2\Delta_y^2}\right)h\right] Z_M^{n+1} = Z_M^n \\ + \left(\frac{\dot{s}}{s\Delta_y} + \frac{2a_M}{Ls\Delta_y^2}\right)hZ_{M-1}^{n+1} + \frac{a_{M+1}}{(L-s)^2\Delta_y^2}hZ_{M+1}^{n+1} + h(f_{M+1} - f_M), \tag{4.60}$$

and

$$\left[1 + \left(\frac{\dot{s}y_{M-i-1}}{(L-s)\Delta_y} + \frac{a_{M+i}}{(L-s)^2\Delta_y^2} + \frac{a_{M+i+1}}{(L-s)^2\Delta_y^2}\right)h\right] Z_{M+i}^{n+1} = Z_{M+i}^n \\ + \left(\frac{\dot{s}y_{M-i}}{(L-s)\Delta_y} + \frac{a_{M+i}}{(L-s)^2\Delta_y^2}\right)hZ_{M+i-1}^{n+1} \\ + \frac{a_{M+i+1}}{(L-s)^2\Delta_y^2}hZ_{M+i+1}^{n+1} + h(f_{M+i+1} - f_{M+i}), \quad i = 1,\ldots,M-2, \tag{4.61}$$

along with the boundary conditions

$$Z_0 = 0, \qquad Z_{2M-1} = 0. \tag{4.62}$$

All the coefficients in (4.58)-(4.61) are non-negative. We take absolute values in (4.58)-(4.61) and then sum up for $i = 1, \ldots, 2M-2$. Due to the specific structure of the coefficients we obtain the recurrent inequality

$$\begin{aligned}\sum_{i=1}^{2M-2} \left|Z_i^{n+1}\right| \leq \sum_{i=1}^{2M-2} |Z_i^n| + h\left(\frac{\dot{s}y_1}{s\Delta_y} + \frac{a_1}{s^2\Delta^2}\right)\left|Z_0^{n+1}\right| \\ + h\frac{a_{2M-1}}{\Delta_y^2(L-s)^2}\left|Z_{2M-1}^{n+1}\right| + h\sum_{i=1}^{2M-2}\left|\frac{f_i - f_{i-1}}{\Delta_y}\right|\Delta_y.\end{aligned}$$

This recurrent inequality implies (4.56). We used the initial condition $Z_i^0 = 0$ for all $i = 0, \ldots, 2M-1$, and (4.62), i.e. $Z_0^l = 0 = Z_{2M-1}^l$ for all $l = 1, \ldots, N_h$. □

We now consider the limit function.

Lemma 4.2.2. *In the limit $h \to 0$ one obtains*

$$V_{i,h}(\tau) \to V_i(\tau) \textit{ uniformly on } (0,T), \tag{4.63}$$

with $V_{i,h}$ from (4.44).

The function $V_i(\tau)$ satisfies (4.40)-(4.43) and also the a priori estimates

$$\max_{\tau\in(0,T)} |V_i(\tau)| \leq C, \quad \forall i = 0, \ldots, 2M, \tag{4.64}$$

and

$$\sum_{i=0}^{2M-1} |V_{i+1}(\tau) - V_i(\tau)| \leq C, \tag{4.65}$$

uniformly with respect to Δ_y.

Proof. The convergence follows from the definition of $V_{i,h}$ and the boundedness of V_i, obtained in Lemma 4.2.1, combined with the Ascoli-Arzelà Theorem (Theorem A.2.1). More precisely, we have

$$|\partial_\tau V_{i,h}(\tau)| = \left|\frac{V_i^{n+1} - V_i^n}{h}\right| \leq C(\Delta_y, a, f)\max_i |V_i^n| \quad \forall h \leq h_0, \tag{4.66}$$

where the last inequality comes from the fact that the lhs is the approximation of $\dot{V}_i$, allowing us to replace it using (4.40), so that all terms can be estimated for fixed Δy.

Lemma 4.2.1 then implies (4.64)-(4.65). Furthermore, time discretisation of (4.40) implies that

$$\begin{aligned}\frac{V_i^{n+1}-V_i^n}{h} &= -\frac{\dot{s}(\tau_{n+1})}{s(\tau_{n+1})}\frac{y_i}{\Delta_y}(V_i^{n+1}-V_{i-1}^{n+1}) \\ &\quad + \frac{a_i(\tau_{n+1})}{s(\tau_{n+1})^2\Delta_y^2}\left((V_{i+1}^{n+1}-V_i^{n+1})-(V_i^{n+1}-V_{i-1}^{n+1})\right)+f_i(\tau_{n+1}), \quad (4.67)\end{aligned}$$

for all $n = 0, 1, \ldots, N_h$, $(N_h h = T)$. We now take the limit $h_l \to 0$ so that $\tau_{n_l} = h_l n_l \to \tau$. The rhs of (4.67) has a limit and consequently

$$\begin{aligned}\lim_{l\to\infty}\frac{V_i^{n_l+1}-V_i^{n_l}}{h_l} &= \dot{V}_i(\tau) = -\frac{\dot{s}(\tau)}{s(\tau)}\frac{y_i}{\Delta_y}(V_i(\tau)-V_{i-1}(\tau)) \\ &\quad + \frac{a_i(\tau)h}{s(\tau)^2\Delta_y^2}\left((V_{i+1}(\tau)-V_i(\tau))-(V_i(\tau)-V_{i-1}(\tau))\right)+hf_i(\tau).\end{aligned}$$

Similar limits hold for (4.41) and (4.43), which proves the Lemma. □

We now turn our attention to V^Δ and W^Δ. We have the following a priori estimates with respect to the space dependency.

Lemma 4.2.3. *The a priori estimates*

$$\int_0^1 \left|\partial_y V^\Delta(y,\tau)\right| \,\mathrm{d}y \le C, \quad \int_0^1 \left|\partial_y W^\Delta(y,\tau)\right| \,\mathrm{d}y \le C, \qquad (4.68)$$

hold uniformly with respect to Δ, $\tau \in (0,T)$. Furthermore, under the assumptions (4.53) we have in the first domain an energy type a priori estimate,

$$\delta(Q)\int_0^\tau \sum_{j=1}^{L}\Delta_y\int_0^{j\Delta_y}\left(\partial_y V^\Delta(y,\tau)\right)^2 \mathrm{d}y\,\mathrm{d}\tau \le K(Q), \qquad (4.69)$$

uniformly for $\{\Delta\}$, and L with $L\Delta_y \le Q < 1$ (K independant of Δ_y).

Proof. From Lemma 4.2.2 and (4.63) we obtain

$$\sum_{i=0}^{M-1}\left|\frac{V_{i+1}(\tau)-V_i(\tau)}{\Delta_y}\right|\Delta_y = \int_0^1\left|\partial_y V^\Delta(y,\tau)\right|\,\mathrm{d}y \le C,$$

and a similar result for W, which proves (4.68).

Now we write (4.40) as

$$\dot{V}_i + \frac{\dot{s}}{s} y_i \delta V_i - \frac{a_i}{s^2 \Delta_y}(\delta V_{i+1} - \delta V_i) = f_i, \quad i = 1, \dots, M-1, \tag{4.70}$$

where we denoted $\delta V_i = \delta V_i^- = \frac{V_i - V_{i-1}}{\Delta_y}$. To proof the energy estimate, we first multiply (4.70) by $V_i \Delta_y$. Summation for $i = 1, \dots, j \le M$ gives

$$\frac{1}{2}\sum_{i=1}^{j} \partial_\tau (V_i)^2 \Delta_y + \sum_{i=1}^{j} \frac{\dot{s}}{s} y_i \delta V_i V_i \Delta_y - \sum_{i=1}^{j} \frac{a_i}{s^2}(\delta V_{i+1} - \delta V_i) V_i = \sum_{i=1}^{j} f_i V_i \Delta_y. \tag{4.71}$$

Note that (4.71) can be written in the form

$$I_{1,j} + I_{2,j} - I_{3,j} = I_{4,j}. \tag{4.72}$$

If the given function f is bounded, then by Lemma 4.2.2 and (4.68) we estimate

$$|I_{2,j}| \le C \int_0^1 \left|\partial_y V^\Delta\right| \,\mathrm{d}y \le C, \qquad |I_{4,j}| \le C. \tag{4.73}$$

Using Abel's summation, (A.2), for the term $I_{3,j}$ gives

$$\begin{aligned} I_{3,j} \quad &\equiv \quad \frac{1}{s^2}\sum_{i=1}^{j} V_i (a_i \delta V_{i+1} - a_{i-1}\delta V_i) - \frac{1}{s^2}\sum_{i=1}^{j} \frac{a_i - a_{i-1}}{\Delta_y} \delta V_i V_i \Delta_y \\ &= \quad -\frac{1}{s^2}\sum_{i=1}^{j} a_{i-1} (\delta V_i)^2 \Delta_y + \left[a_j V_j \delta V_{j+1} - a_0 V_0 \delta V_1\right] \frac{1}{s^2} \\ &\quad\quad -\frac{1}{s^2}\sum_{i=1}^{j} \frac{a_i - a_{i-1}}{\Delta_y} \delta V_i V_i \Delta_y. \end{aligned} \tag{4.74}$$

Now, multiply (4.71) by Δ_y and sum for $j = 1, \dots, L$, with $L\Delta_y \le Q < 1$. Noticing that $\delta V_1 = 0$ and invoking (4.73) and (4.74), we obtain

$$\begin{aligned} &\frac{1}{2}\sum_{j=1}^{L}\sum_{i=1}^{j} \partial_\tau (V_i)^2 \Delta_y^2 + \frac{1}{s^2}\sum_{j=1}^{L} \Delta_y \sum_{i=1}^{j} a_{i-1} (\delta V_i)^2 \Delta_y \\ &\le C + \frac{1}{s^2}\sum_{j=1}^{L} \Delta_y \sum_{i=1}^{j} \left|\frac{a_i - a_{i-1}}{\Delta_y}\right| |\delta V_i| \, V_i \Delta_y + \frac{1}{s^2}\sum_{j=1}^{L} a_j V_j \, |\delta V_{j+1}| \, \Delta_y. \end{aligned} \tag{4.75}$$

The third term of the rhs can be estimated in the same way as $I_{2,j}$. To estimate (4.75) further we apply the properties of a given in (4.53). We find

$$\frac{1}{2}\sum_{j=1}^{L}\sum_{i=1}^{j}\partial_\tau (V_i)^2\Delta_y^2 + \delta(Q)\sum_{j=1}^{L}\Delta_y\sum_{i=1}^{j}(\delta V_i)^2\Delta_y \le K(Q,\tau). \tag{4.76}$$

Integrating (4.76) over $(0,\tau)$, and noticing that

$$\int_0^T K(Q,\tau)\,\mathrm{d}\tau = K(Q) < \infty, \tag{4.77}$$

leads to

$$\frac{1}{2}\int_0^\tau \sum_{j=1}^{L}\Delta_y\sum_{i=1}^{j}\Delta_y (V_i(\tau))^2\,\mathrm{d}\tau + \delta(Q)\int_0^\tau \sum_{j=1}^{L}\Delta_y\sum_{i=1}^{j}(\delta V_i)^2\Delta_y\,\mathrm{d}\tau \le K(Q), \tag{4.78}$$

which implies the required result (4.69). □

Remark 4.2.4. *From Lemma 4.2.3 energy estimates are seen to be only possible in the first domain, away from the degeneracy point. In this part, we can take the limit $\Delta_y \to 0$ and get convergence. Unfortunately, the L_1 estimates (4.68) are not that usefull, as the L_1-space is not reflexive, and hence compactness does not follow from an estimate of the type $\int_0^Q |\partial_x f|\,\mathrm{d}x < C$. We will have to work towards Kolmogorov compactness to prove convergence over the entire domain.*

Remark 4.2.5. *This is the first time the structure of $\partial_x a(x,t)$ plays a role in the form of the assumptions (4.53). Here, the degeneracy at $x = s(t)$ has been overcome by leaving out the neighbourhood of $x = s(t)$, see Fig. 4.3. Only within this reduced region energy type estimates can be constructed.*

We now deduce analogous results with respect to the τ-variable.

Lemma 4.2.4. *Let the assumptions (4.53) be satisfied. Then,*

$$\int_0^{T-z}\int_0^Q\int_0^x \left(V^\Delta(y,\tau+z) - V^\Delta(y,\tau)\right)^2 \mathrm{d}y\mathrm{d}x\mathrm{d}\tau \le C(Q)z, \tag{4.79}$$

uniformly for $\{\Delta\}$ with $Q < 1$, and

$$\int_0^T\int_0^Q \left|\partial_\tau W^\Delta(y,\tau)\right| \mathrm{d}y\mathrm{d}\tau \le C, \tag{4.80}$$

uniformly for $\{\Delta\}$, with $Q < 1$.

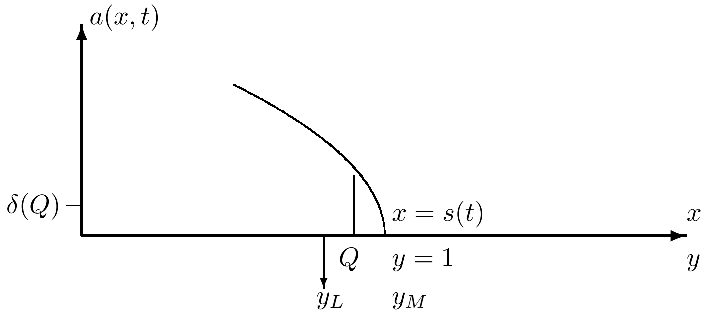


Figure 4.3: The region around the edge $x = s(t)$ is left out in order to obtain a priori error estimates of the energy type.

Proof. We integrate (4.40) over $(\tau, \tau+z)$. Multiplying by $(V_i(\tau+z) - V_i(\tau))\,\Delta_y$ and summing for $i = 1, \ldots, j$ yield

$$\begin{aligned}
&\sum_{i=1}^{j} (V_i(\tau+z) - V_i(\tau))^2\,\Delta_y = \\
&\qquad -\sum_{i=1}^{j} y_i \Delta_y \int_{\tau}^{\tau+z} \frac{\dot{s}(T-r)}{s(T-r)} \delta V_i(r)\,\mathrm{d}r\,(V_i(\tau+z) - V_i(\tau)) \\
&\quad +\sum_{i=1}^{j} \int_{\tau}^{\tau+z} \frac{a_i(r)}{(s(T-r))^2} \left[\delta V_{i+1}(r) - \delta V_i(r)\right]\,\mathrm{d}r\,(V_i(\tau+z) - V_i(\tau)) \\
&\quad +\sum_{i=1}^{j} \Delta_y \int_{\tau}^{\tau+z} f_i(r)\,\mathrm{d}r\,(V_i(\tau+z) - V_i(\tau)) \equiv I_{1,j} + I_{2,j} + I_{3,j}. \qquad (4.81)
\end{aligned}$$

By Lemma 4.2.2 and Lemma 4.2.3 it follows that

$$|I_{1,j}| \le Cz, \qquad |I_{3,j}| \le C \sum_{i=1}^{j} \Delta_y \int_{\tau}^{\tau+z} f_i(r)\,\mathrm{d}r \le Cz. \qquad (4.82)$$

Abel's summation for $I_{2,j}$ implies for the part with $V_i(\tau+z)$,

$$\begin{aligned}
I_{2,j}^1 &= \sum_{i=1}^{j}\int_\tau^{\tau+z} \frac{a_i(r)}{(s(T-r))^2}\left[\delta V_{i+1}(r)-\delta V_i(r)\right]\,\mathrm{d}r V_i(\tau+z)\\
&= \int_\tau^{\tau+z} \frac{a_j(r)}{(s(T-r))^2}\delta V_{j+1}(r)\,\mathrm{d}r V_j(\tau+z)\\
&\quad -\int_\tau^{\tau+z} \frac{a_0(r)}{(s(T-r))^2}\delta V_1(r)\,\mathrm{d}r V_0(\tau+z)\\
&\quad -\sum_{i=1}^{j}\int_\tau^{\tau+z} \frac{a_i(r)V_i(\tau+z)-a_{i-1}(r)V_{i-1}(\tau+z)}{(s(T-r))^2}\delta V_i(r)\,\mathrm{d}r\\
&= \int_\tau^{\tau+z} \frac{a_j(r)}{(s(T-r))^2}\delta V_{j+1}(r)\,\mathrm{d}r V_j(\tau+z)\\
&\quad -\sum_{i=1}^{j}\int_\tau^{\tau+z} \frac{a_i(r)\delta V_i(\tau+z)}{(s(T-r))^2}\delta V_i(r)\,\mathrm{d}r\Delta_y\\
&\quad +\sum_{i=1}^{j}\Delta_y\int_\tau^{\tau+z} \frac{a_i(r)-a_{i-1}(r)}{\Delta_y(s(T-r))^2}V_{i-1}(\tau+z)\delta V_i(r)\,\mathrm{d}r = J_{1,j}^1+J_{2,j}^1+J_{3,j}^1,
\end{aligned}$$

where we used $\delta V_1 = 0$. By Lemma 4.2.2 and Lemma 4.2.3, we estimate

$$\left|J_{3,j}^1\right| \le CK(Q)z, \quad \text{provided that } j \le L,\ L\Delta_y \le Q. \tag{4.83}$$

By the Cauchy inequality it follows that

$$\begin{aligned}
\left|J_{2,j}^1\right| &\le \frac{1}{2}\sum_{i=1}^{j}\Delta_y\int_\tau^{\tau+z}\frac{a_i(r)}{(s(T-r))^2}\left(\delta V_i(r)\right)^2\,\mathrm{d}r\\
&\quad +\frac{1}{2}\sum_{i=1}^{j}\Delta_y\int_\tau^{\tau+z}\frac{a_i(r)}{(s(T-r))^2}\,\mathrm{d}r\left(\delta V_i(\tau+z)\right)^2\\
&\le \frac{1}{2}C\sum_{i=1}^{j}\Delta_y\int_\tau^{\tau+z}\left(\delta V_i(r)\right)^2\,\mathrm{d}r + Cz\sum_{i=1}^{j}\Delta_y\left(\delta V_i(\tau+z)\right)^2.
\end{aligned}\tag{4.84}$$

For $J_{1,j}^1$ we have

$$\left|J_{1,j}^1\right| \le C\int_\tau^{\tau+z}\left|\delta V_{j+1}(r)\right|\mathrm{d}r. \tag{4.85}$$

We now multiply (4.81) by Δ_y and sum for $j = 1, \dots, L$. Next, we integrate over $(0, T - z)$. The estimates obtained till now that are of the form $C(Q)z$, will, after this operation, still be of the form $C(Q)z$. Then, by means of Lemma 4.2.3 the sum over j (discrete space integration) in (4.85) can be estimated to be bounded above by Cz. The second term of (4.84) can also be estimated by means of Lemma 4.2.3 to be bounded above by $C(Q)z$. We conclude that

$$\int_0^{T-z} \sum_{j=1}^{L} \sum_{i=1}^{j} \left(V_i(\tau + z) - V_i(\tau)\right)^2 \Delta_y^2 \, \mathrm{d}\tau \leq C(Q)z + \\ C \int_0^{T-z} \int_\tau^{\tau+z} \sum_{j=1}^{L} \Delta_y \sum_{i=1}^{j} \Delta_y \left(\delta V_i(r)\right)^2 \, \mathrm{d}r \, \mathrm{d}\tau. \quad (4.86)$$

To estimate the last term of (4.86) we use Lemma 4.2.3 in combination with the estimate (see Fig. 4.4)[1]

$$\int_0^{T-z} \int_\tau^{\tau+z} g(r) \, \mathrm{d}r \, \mathrm{d}\tau \leq 2z \int_0^T g(r)) \, \mathrm{d}r \quad \text{with } g \geq 0, \quad (4.87)$$

which implies that

$$\int_0^{T-z} \sum_{j=1}^{L} \sum_{i=1}^{j} \left(V_i(\tau + z) - V_i(\tau)\right)^2 \Delta_y^2 \, \mathrm{d}\tau \leq C(Q)z, \quad (4.88)$$

uniformly for $\{\Delta\}$ with $\Delta_y L \leq Q < 1$.

As a consequence of (4.47) assertion (4.79) holds.

We now turn our attention to W^Δ. In general, $a_i^* \equiv 0$, see (4.33.2), (4.36). However, we allow for a regularization of a, such that $a_i > 0$ close to the interface $x = s(t)$. Therefore, there is an index L such that, for $i \leq L$ $(\Delta_y \leq \Delta_0)$, $a_i^* \equiv 0$. Consequently, from (4.42) we get

$$\left|\dot{W}_i(\tau)\right| \leq C \left|\delta W_{i+1}\right| + \tilde{f}_i^*, \quad \text{for } i = 1, \dots, L. \quad (4.89)$$

Due to Lemma 4.2.3 we find

$$\begin{aligned} \int_0^T \sum_{i=1}^{L} \left|\dot{W}_i\right| \Delta_y \, \mathrm{d}\tau &\leq C \int_0^T \int_0^1 \left|\partial_y W^\Delta(y, \tau)\right| \, \mathrm{d}y\mathrm{d}\tau + \int_0^T \int_0^1 \left|f^*(y, \tau)\right| \, \mathrm{d}y\mathrm{d}\tau \\ &\leq C, \end{aligned} \quad (4.90)$$

[1] This estimates follows from changing the order of integration, i.e. from Fig. 4.4 we have that $\int_0^{T-z} \int_\tau^{\tau+z} g(r) dr d\tau \leq \int_0^T \int_{r-z}^r g(r) d\tau dr = \int_0^T g(r)(r - (r - z)) d\tau$.

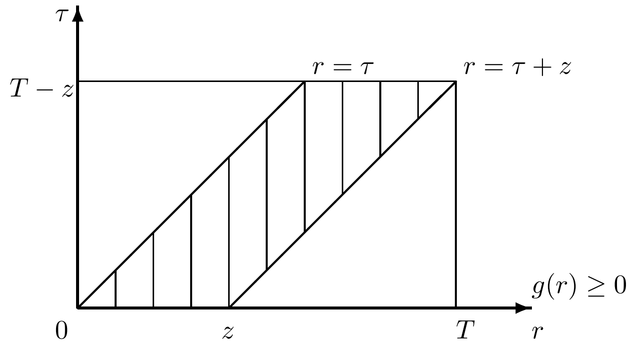


Figure 4.4: Motivation for estimate (4.87)

from which assertion (4.80) follows. □

Remark 4.2.6. *Estimate (4.80) of Lemma 4.2.4 only holds because $a_i^* = 0$. Note further that the integration on $(0,Q)$ in Lemma 4.2.4 allows us to use a discretization of a, denoted by a_Δ, where $a_M \neq 0$ and even $a_{M-k} \neq 0$ as long as $M-k > L$. Taking a_{Δ,M_L} to be the last value of a_Δ that is non zero, we can, for a given Q, choose Δ_0 small enough so that $a_{\Delta,i} = 0$, $y_i < Q$.*

To finish the a priori estimates, we need to rephrase the energy type estimate (4.69), obtained in Lemma 4.2.3, that allows a combination with (4.79).

Lemma 4.2.5. *Let the assumptions of Lemma 4.2.3 be satisfied. Then,*

$$\int_0^T\int_0^Q\int_0^x \left(V^\Delta(y+p,\tau)-V^\Delta(y,\tau)\right)^2 \,\mathrm{d}y\mathrm{d}x\mathrm{d}\tau \leq p^2K(Q+p_0), \tag{4.91}$$

holds uniformly for $\{\Delta\}$, $0<p\leq p_0$, with $Q+p_0<1$.

Proof. Estimate (4.69) can be rewritten as

$$\int_0^T\int_0^Q\int_0^x \left(\partial_y V^\Delta(y,\tau)\right)^2 \,\mathrm{d}y\mathrm{d}x\mathrm{d}\tau \leq K(Q). \tag{4.92}$$

One calculates

$$V^\Delta(y+p,\tau)-V^\Delta(y,\tau) = \int_y^{y+p} \partial_y V^\Delta(y',\tau)\,\mathrm{d}y' = p\int_0^1 \partial_y V^\Delta(y+rp,\tau)\,\mathrm{d}r,$$

where we set $y' = y + rp$ (∂_y denotes the partial derivative to the first component throughout). Hence,

$$\left(V^{\Delta}(y+p,\tau) - V^{\Delta}(y,\tau)\right)^2 \leq p^2 \int_0^1 \left(\partial_y V^{\Delta}(y+rp,\tau)\right)^2 \mathrm{d}r,$$

and consequently,

$$\int_0^T \int_0^Q \int_0^x \left(V^{\Delta}(y+p,\tau) - V^{\Delta}(y,\tau)\right)^2 \mathrm{d}y\mathrm{d}x\mathrm{d}\tau \\ \leq p^2 \int_0^T \int_0^Q \int_0^x \int_0^1 \left(\partial_y V^{\Delta}(y+rp,\tau)\right)^2 \mathrm{d}r\mathrm{d}y\mathrm{d}x\mathrm{d}\tau.$$

By (4.92), the transformation $\xi = y + rp$, $z = p + x$, $r = r$, gives,

$$\int_0^T \int_0^Q \int_0^x \left(V^{\Delta}(y+p,\tau) - V^{\Delta}(y,\tau)\right)^2 \mathrm{d}y\mathrm{d}x\mathrm{d}\tau \\ \leq p^2 \int_0^1 \int_0^T \int_p^{Q+p} \int_{rp}^{z-p+rp} \left(\partial_y V^{\Delta}(\xi,\tau)\right)^2 \mathrm{d}\xi\mathrm{d}z\mathrm{d}\tau\mathrm{d}r \\ \leq p^2 \int_0^1 \int_0^T \int_0^{Q+p} \int_0^z \left(\partial_y V^{\Delta}(\xi,\tau)\right)^2 \mathrm{d}\xi\mathrm{d}z\mathrm{d}\tau\mathrm{d}r \leq p^2 K(Q+p_0),$$

uniformly for $\{\Delta\}$, $0 < p \leq p_0$, with $Q + p_0 < 1$. Here we used $z - p + rp < z$ when $r \in (0,1)$. □

4.2.4 Convergence

We now prove the compactness of $\{V^{\Delta}\}$ and $\{W^{\Delta}\}$ in L_2 and L_1 respectively.

Lemma 4.2.6. *The sequence $\{V^{\Delta_l}\}_{l=1}^{\infty}$ ($|\Delta_l| \to 0$ for $l \to \infty$) is relatively compact in $L_2(I \times (0, Q_V))$ for any $Q_V < 1$.*

The sequence $\{W^{\Delta_l}\}_{l=1}^{\infty}$ is relatively compact in $L_1(I \times (0, Q_W))$ for any $Q_W < 1$.

Proof. For V^{Δ} we can rewrite the estimates (4.79) and (4.91) using the formula

$$\int_0^Q \int_0^x f(s)\,\mathrm{d}s\mathrm{d}x = \int_0^Q (Q-s)f(s)\,\mathrm{d}s.$$

We get

$$\int_0^{T-z}\int_0^{Q'} (Q'-y)\left(V^\Delta(y,\tau+z) - V^\Delta(y,\tau)\right)^2 \,\mathrm{d}y\mathrm{d}\tau \le C(Q')z$$

$$\int_0^{T}\int_0^{Q'} (Q'-y)\left(V^\Delta(y+p,\tau) - V^\Delta(y,\tau)\right)^2 \,\mathrm{d}y\mathrm{d}\tau \le C(Q')p^2,$$

$\forall z \le z_0$, $p \le p_0$. This implies that

$$\int_0^{T-z}\int_0^{Q''} \left(V^\Delta(y,\tau+z) - V^\Delta(y,\tau)\right)^2 \,\mathrm{d}y\mathrm{d}\tau \le C(Q')z, \tag{4.93}$$

with $Q'' < Q'$. Then, by a Kolmogorov compactness argument, the compactness of $\{V^\Delta\}$ in $L_2(I \times Q'')$ follows, where $Q'' < Q' < 1$.

For W^Δ, a straightforward consequence of Lemma 4.2.3 is that

$$\int_0^{T}\int_0^{Q} \left|\partial_y W^\Delta(y,\tau)\right| \,\mathrm{d}y\mathrm{d}\tau \le C.$$

Together with Lemma 4.2.4 this gives the required compactness result by Kolmogorov compactness. □

In Lemma 4.2.6 the constants Q_V and Q_W depend on the chosen regularization of a, combined with the assumptions (4.53). We review the different possibilities.

1. No specific regularization is applied. Due to the space discretization, this implies a discrete regularization (linear interpolation of $a(x_i,t)$). In part I of the domain we have $Q_V < 1$ to satisfy (4.53). In part II we have no limitation, so we can take $Q_W = 1$. Only due to the continuity condition at $x = s(t)$, we have continuity there.

2. The following regularization can be used. Let $\{\Delta_x(t)\} \equiv \{x_i(t)\}_{i=0}^{2M}$ be a set of moving grid points for the numerical approximation of (4.30) and let $x_M(t) \equiv s(t)$ correspond to the degeneracy point of a. Then, define

$$a_\Delta(x,t) = \begin{cases} a(x,t) & \text{for } x < x_{M-1}(t) \\ p_3(x,t) & x_{M-1}(t) < x < x_{M+2}(t) \\ 0 & \text{for } x > x_{M+2}(t), \end{cases} \tag{4.94}$$

where $p_3(x,t)$ is the bicubic spline, which satisfies $p_3(x_{M-1}(t),t) = a(x_{M-1}(t),t)$, $\partial_x p_3(x_{M-1}(t),t) = \partial_x a(x_{M-1}(t),t)$, and analogously at x_{M+2}. Clearly, for this regularization $a_\Delta(x,t) \to a(x,t)$, uniformly in $(0,L)$, for $\Delta_x \to 0$. Now, condition (4.53) is always satisfied, and we can take in part I, $Q_V = 1$. However, in part II we have $Q_W < 1$, such that it corresponds with a value $x_Q > x_{M+2}$. It is always possible to choose Δ_y sufficiently small to fulfill this condition.

3. Another possibility is to apply an epsilon regularization:

Definition 4.2.3 (ϵ-regularization). *The ϵ-regularization of (4.30) is defined by*

$$\partial_t \psi(x,t) + a_\epsilon(x,t)\partial_x^2 \psi(x,t) = f(x,t), \tag{4.95}$$

where $a_\epsilon := a + \epsilon$, $\epsilon > 0$ being a small constant.

Note that (4.95) differs from (4.30) by a small viscosity term, $\epsilon\partial_x^2\psi$, only. Also a_ϵ is degenerate, i.e. $\partial_x a_\epsilon(s(t),t) = -\infty$. Thus, the unboundedness of the derivative of a remains. Therefore, in part I of the domain we need $Q_V < 1$. In part II, the diffusion term does not disappear. Therefore, all estimates for W are invalid. The necessary alterations are considered in Section 4.2.5. To obtain the required solution, the limit $\epsilon \to 0$ will have to be considered.

In all cases, one of the two constants Q remains. This is not a difficulty as the following lemma holds.

Lemma 4.2.7. *Consider a sequence $\{\Delta_l\}_{l=1}^{\infty}$ with $|\Delta_l| \to 0$ for $l \to \infty$. There exists a $V \in L_\infty(I\times(0,1))$ and $W \in L_\infty(I\times(0,1))$, and a subsequence $\{l_k\} \subset \{l\}$ so that for $\Delta_{l_k} \to 0$ one has*

$$V^{\Delta_{l_k}}(y,t) \to V(y,t), \quad W^{\Delta_{l_k}}(y,t) \to W(y,t) \quad \text{for a.e. } (y,t) \in (0,1)\times(0,T).$$

Proof. Consider $\{Q_l\}$, $Q_l \nearrow 1$ for $l \to \infty$. We can choose a $\{l_1\} \subset \{l\}$ so that $V^{\Delta_{l_1}} \to V$ and $W^{\Delta_{l_1}} \to W$ for a.e. $(y,t) \in (0,Q_1)\times(0,T)$. Since $Q_l \nearrow 1$, by the method of diagonalization, we can choose a subsequence $\{\Delta_{l_k}\} \subset \{\Delta_l\}$ so that $V^{\Delta_{l_k}} \to V$ and $W^{\Delta_{l_k}} \to W$ for a.e. $(y,t) \in (0,1)\times(0,T)$. As before, the convergence of V is in L_2 and the one of W is in L_1. The boundedness of V and W follows from Lemma 4.2.2. This completes the proof. □

We are now prepared to prove Theorem 4.2.1.

Proof of Theorem 4.2.1. Due to Definition 4.2.2 it is sufficient to prove that $V^\Delta \to \psi_I$ and $W^\Delta \to \psi_{II}$, where ψ_I and ψ_{II} are the corresponding variational solutions of (4.34) and (4.36), respectively.

We consider a sequence $\{\Delta_l\}_{l=1}^{\infty}$ with gridsize $|\Delta_y^l| \to 0$ for $l \to \infty$. The number of gridpoints is given by M_l. Let $\phi \in C^\infty(\tilde{Q}_T)$. Denote by $\phi_i(\tau) =$

$\frac{1}{\Delta_y^l} \int_{y_i-1/2}^{y_i+1/2} \phi(y,\tau)\,\mathrm{d}y$, $i = 1,\ldots,L$ (we omit the index l). Multiply (4.40) by $\phi_i \Delta_y$ and sum for $i = 1,\ldots,M$. Since $\operatorname{supp}\phi \subset [0,T) \times [0,1)$ there exists an integer l_0 so that $\phi_i \equiv 0$ for $i = M-1, M$ when $l > l_0$. We obtain the equality

$$\sum_{i=1}^{M-1} \dot{V}_i \phi_i \Delta_y + \sum_{i=1}^{M-1} \frac{\dot{s}}{s} y_i \delta V_i \phi_i \Delta_y - \sum_{i=1}^{M-1} \frac{a_i}{s^2} (\delta V_{i+1} - \delta V_i) \phi_i = \sum_{i=1}^{M-1} f_i \phi_i \Delta_y. \quad (4.96)$$

By Abel's summation this reduces to

$$\sum_{i=1}^{M-1} \frac{a_i}{s^2} (\delta V_{i+1} - \delta V_i) \phi_i = - \sum_{i=1}^{M-1} \frac{1}{s^2} \delta V_i \delta(a_i \phi_i) \Delta_y,$$

where we used $\delta V_1 = 0$ and $a_{M-1}\phi_{M-1} = 0$. Integrating (4.96) with respect to τ and performing integration by parts yields

$$\begin{aligned} - \int_0^T \sum_{i=1}^{M-1} V_i \dot{\phi}_i \Delta_y \,\mathrm{d}\tau &+ \int_0^T \sum_{i=1}^{M-1} \frac{\dot{s}}{s} y_i \delta V_i \bar{\phi}^\Delta \Delta_y \,\mathrm{d}\tau \\ &+ \int_0^T \sum_{i=1}^{M-1} \frac{1}{s^2} \delta V_i \partial_y (a\phi)^\Delta \Delta_y \,\mathrm{d}\tau = \int_0^T \sum_{i=1}^{M-1} \bar{f}^\Delta \bar{\phi}^\Delta \Delta_y \,\mathrm{d}\tau, \quad (4.97) \end{aligned}$$

where we used $V_i(0) = 0$ and $\phi_i(T) = 0$. We denoted by $(a\phi)^\Delta$ the piecewise linear function in y defined by means of $a_i \phi_i$ as in (4.47). The function $\bar{\phi}^\Delta$ is piecewise constant in y, so $\bar{\phi}^\Delta = \phi_i$ for $y \in (y_{i-1}, y_i)$. The same holds for $\bar{f}^\Delta$. We can rewrite (4.97) in the form (see (4.47), (4.48)),

$$\begin{aligned} - \int_0^T \int_0^1 \bar{V}^\Delta \partial_\tau \bar{\phi}^\Delta \,\mathrm{d}y\mathrm{d}\tau &+ \int_0^T \int_0^1 \frac{\dot{s}}{s} \bar{y}^\Delta \partial_y V^\Delta \bar{\phi}^\Delta \,\mathrm{d}y\mathrm{d}\tau \\ &+ \int_0^T \int_0^1 \frac{1}{s^2} \partial_y V^\Delta \partial_y (a\phi)^\Delta \,\mathrm{d}y\mathrm{d}\tau = \int_0^T \int_0^1 \bar{f}^\Delta \bar{\phi}^\Delta \,\mathrm{d}y\mathrm{d}\tau. \quad (4.98) \end{aligned}$$

First, we estimate

$$\int_0^T \int_0^1 \left| V^\Delta - \bar{V}^\Delta \right| \mathrm{d}y\mathrm{d}\tau = \frac{1}{2} |\Delta_y| \int_0^T \int_0^1 \left| \partial_y V^\Delta \right| \mathrm{d}y\mathrm{d}\tau \leq C |\Delta_y| \to 0, \quad (4.99)$$

by Lemma 4.2.3.

Since $V^\Delta \to V$ in $L_2((0,T)\times(0,1))$, $V^\Delta \to V$ a.e. in $(0,T)\times(0,1)$ (see Lemma 4.2.7) and $\left|V^\Delta\right| \leq C < \infty$ (from Lemma 4.2.2), and since ϕ is smooth, it follows from (4.99) by invoking the Lebesgue theorem, (A.2.3), that[2]

$$-\int_0^T\int_0^1 \bar{V}^\Delta \partial_\tau \bar{\phi}^\Delta \,\mathrm{d}y\mathrm{d}\tau \underset{\Delta\to 0}{\longrightarrow} -\int_0^T\int_0^1 V\partial_\tau\phi \,\mathrm{d}y\mathrm{d}\tau. \tag{4.100}$$

From (4.69) we obtain, as in Lemma 4.2.6, that

$$\int_0^T\int_0^Q \left(\partial_y V^\Delta\right)^2 \mathrm{d}y\mathrm{d}\tau \leq K(Q) \quad \text{for } Q<1. \tag{4.101}$$

This in turn implies the existence of a subsequence of $\{V^\Delta\}$ for which $\partial_y V^\Delta \rightharpoonup \chi$ in $L_2(I\times(0,Q))$ (weak convergence) (see Appendix, Lemma A.2.4). From the identity

$$\int_0^T\int_0^1 V^\Delta \partial_y v \,\mathrm{d}y\mathrm{d}\tau = -\int_0^T\int_0^1 \partial_y V^\Delta v \,\mathrm{d}y\mathrm{d}\tau,$$

with $v\in C_0^\infty(Q_T)$ and from the convergences

$$\int_0^T\int_0^1 V^\Delta \partial_y v \,\mathrm{d}y\mathrm{d}\tau \underset{\Delta\to 0}{\longrightarrow} \int_0^T\int_0^1 V\partial_y v \,\mathrm{d}y\mathrm{d}\tau,$$

$$\int_0^T\int_0^1 \partial_y V^\Delta v \,\mathrm{d}y\mathrm{d}\tau \underset{\Delta\to 0}{\longrightarrow} \int_0^T\int_0^1 \chi\ v \,\mathrm{d}y\mathrm{d}\tau,$$

we find that

$$\int_0^T\int_0^1 V\partial_y v \,\mathrm{d}y\mathrm{d}\tau = \int_0^T\int_0^1 \chi\ v \,\mathrm{d}y\mathrm{d}\tau \quad \forall v\in C_0^\infty(Q_T).$$

Hence, $\chi = \partial_y V \in L_2(I\times(0,Q))$[3] in the weak sense. We can assume (ϕ is fixed) that $\operatorname{supp}_y \phi \leq Q < 1$, therefore

$$\int_0^T\int_0^1 \frac{\dot{s}}{s}\bar{y}^\Delta \partial_y V^\Delta \bar{\phi}^\Delta \,\mathrm{d}y\mathrm{d}\tau \underset{\Delta\to 0}{\longrightarrow} \int_0^T\int_0^1 \frac{\dot{s}}{s} y\partial_y V\phi \,\mathrm{d}y\mathrm{d}\tau, \tag{4.102}$$

[2] This is seen as follows. We have

$$\begin{aligned}(\bar{V}^\Delta,\partial_\tau\bar{\phi}^\Delta) &= (\bar{V}^\Delta,\partial_\tau\bar{\phi}^\Delta-\partial_\tau\phi)+(\bar{V}^\Delta-V^\Delta,\partial_\tau\phi)+(V^\Delta,\partial_\tau\phi)\\ &\leq ||\bar{V}^\Delta||_{L_2}||\partial_\tau\bar{\phi}^\Delta-\partial_\tau\phi||_{L_2}+||\bar{V}^\Delta-V^\Delta||_{L_1}||\partial_\tau\phi||_{L_\infty}+(V^\Delta,\partial_\tau\phi).\end{aligned}$$

In the limit, the first term tends to 0 by the smoothness of ϕ and the second term by (4.99). It remains to prove that we can pass to the limit under the integral sign. Notice that $V^\Delta \to V$ a.e., and that $|V^\Delta| \leq g$, where g is an integrable function.Thus, Lebesgue's theorem is applicable.

[3] Here we write $y\in(0,Q)$. Note that the y-integrals are on $(0,1)$ due to the compact support of v.

since $\frac{\dot{s}}{s}\bar{y}^\Delta\bar{\phi}^\Delta \to \frac{\dot{s}}{s}y\phi$ in $L_2(Q_T)$ (strongly).

Similarly,

$$\int_0^T\int_0^1 \frac{1}{s^2}\partial_y V^\Delta \partial_y(a\phi)^\Delta \,\mathrm{d}y\mathrm{d}\tau \xrightarrow[\Delta\to 0]{} \int_0^T\int_0^1 \frac{1}{s^2}\partial_y V\partial_y(a\phi)\,\mathrm{d}y\mathrm{d}\tau, \tag{4.103}$$

since $\partial_y(a\phi)^\Delta \to \partial_y(a\phi)$ in $L_2(Q_T)$ (strongly). Furthermore, we have that $\bar{f}^\Delta\bar{\phi}^\Delta \to -f\phi$, due to the fact that $f_i = -f(y_i s(T-\tau), T-\tau)$.

Taking the limit $|\Delta_y| \to 0$ in (4.98) and applying (4.100)-(4.103), we conclude that $V \equiv \psi_I$ satisfies (4.50).

Now, multiply (4.42) by $\phi_i\Delta_y$ and sum for $i = 1, \ldots, M$. Due to the properties of the regularization of a (see (4.94)), we have that $a_i = 0$ for $i \le L$. We can assume that $\operatorname{supp}_y \phi \le Q = L\Delta_y < 1$ because ϕ has compact support. It follows that the contribution resulting from the third term of (4.42) is 0. The argument used for $\partial_y V^\Delta$ cannot be repeated for $\partial_y W^\Delta$, because we only proved that $\partial_y W^\Delta \in L_1$, and L_1 is not reflexive. To this end we proceed as follows. Due to

$$\sum_{i=1}^{M-1} y_i\delta W_{i+1}\phi_i\Delta_y = y_{M-1}\phi_{M-1}W_M - y_0\phi_0 W_1 - \sum_{i=1}^{M-1} W_i\delta(\phi_i y_i)\Delta_y, \tag{4.104}$$

we obtain

$$\sum_{i=1}^{M-1}\dot{W}_i\phi_i\Delta_y + \sum_{i=1}^{M-1}\frac{\dot{s}}{L-s}W_i\delta(\phi_i y_i)\Delta_y = \sum_{i=1}^{M-1} f_i^*\phi_i\Delta_y, \tag{4.105}$$

where we used $\phi_{M-1} = 0$ and $y_0 = 0$ (Note that in general $W_1 \neq 0$). Integrating over τ and performing integration by parts yield

$$-\int_0^T\int_0^1 \bar{W}^\Delta\partial_\tau\bar{\phi}^\Delta\,\mathrm{d}y\mathrm{d}\tau + \int_0^T\int_0^1\frac{\dot{s}}{L-s}\bar{W}^\Delta\partial_y(\phi y)^\Delta\,\mathrm{d}y\mathrm{d}\tau \\ = \int_0^T\int_0^1 \bar{f}^\Delta\bar{\phi}^\Delta\,\mathrm{d}y\mathrm{d}\tau. \tag{4.106}$$

We now take the limit $|\Delta_y| \to 0$ and use $W^\Delta \to W$ in $L_1(I\times(0,Q))$ (see Lemma 4.2.7 and Lemma 4.2.2). Replacing V^Δ and $\bar{V}^\Delta$ by W^Δ and $\bar{W}^\Delta$ in the estimate (4.99) gives $\bar{W}^\Delta \to W$ in $L_1(I\times(0,Q))$. The smoothness of ϕ together with (4.106) implies the identity (4.51). Thus, $W \equiv \psi_{II}(y,\tau)$.

It remains to prove the boundedness of the total variation of ψ_I and ψ_{II}. Passing to the modulus and integrating the equality

$$V^{\Delta}(y+\epsilon,\tau)-V^{\Delta}(y,\tau)=\epsilon\int_0^1 \partial_y V^{\Delta}(y+r\epsilon,\tau)\,\mathrm{d}r, \quad 0<y<1-\epsilon. \tag{4.107}$$

over $(t,y)\in(0,T)\times(0,1-\epsilon)$ gives

$$\int_0^T\int_0^{1-\epsilon}\left|V^{\Delta}(y+\epsilon,\tau)-V^{\Delta}(y,\tau)\right|\leq\epsilon\int_0^1\int_0^T\int_0^1\left|\partial_y V^{\Delta}(y,\tau)\right|\,\mathrm{d}y\mathrm{d}\tau\mathrm{d}r\leq\epsilon C. \tag{4.108}$$

Similarly,

$$\int_0^T\int_0^{1-\epsilon}\left|W^{\Delta}(y+\epsilon,\tau)-W^{\Delta}(y,\tau)\right|\leq\epsilon\int_0^1\int_0^T\int_0^1\left|\partial_y W^{\Delta}(y,\tau)\right|\,\mathrm{d}y\mathrm{d}\tau\mathrm{d}r\leq\epsilon C. \tag{4.109}$$

We let $|\Delta_y|\to 0$ in (4.108)-(4.109). By the pointwise convergences $V^{\Delta}\to V$ and $W^{\Delta}\to W$ and by the Lebesgue theorem, it follows that

$$\limsup_{\epsilon\to 0}\frac{1}{\epsilon}\int_0^T\int_0^{1-\epsilon}|V(y+\epsilon,\tau)-V(y,\tau)|\,\mathrm{d}y\mathrm{d}\tau<C<\infty, \tag{4.110}$$

and a similar result for W. □

Remark 4.2.7. *Despite the obtained theoretical results about the solution of (4.30)-(4.32), there are still open questions concerning uniqueness and smoothness of the solution ψ_{II}. However, the solution ψ_I is (at least) Hölder continuous for $y\in(0,Q)$ in $L_2(0,T)$. This is a consequence of $\partial_y\psi_I\in L_2((0,T)\times(0,Q))$: by the Cauchy-Schwarz inequality we have*

$$\begin{aligned}|\psi_I(y_1,\tau)-\psi_I(y_2,\tau)| &\leq \left|\int_{y_1}^{y_2}\partial_y\psi_I(\xi,\tau)\,\mathrm{d}\xi\right| \\ &\leq |y_1-y_2|^{1/2}\left(\int_{y_1}^{y_2}|\partial_y\psi_I(\xi,\tau)|^2\,\mathrm{d}\xi\right)^{1/2},\end{aligned}$$

and hence,

$$\begin{aligned}\int_0^T|\psi_I(y_1,\tau)-\psi_I(y_2,\tau)|^2\,\mathrm{d}\tau &\leq |y_1-y_2|\int_0^T\int_0^Q|\partial_y\psi_I(\xi,\tau)|^2\,\mathrm{d}\xi\mathrm{d}\tau \\ &\leq C(Q)\,|y_1-y_2|.\end{aligned}$$

4.2.5 Convergence of the ϵ-regularization

We introduced an ϵ-regularization in Definition 4.2.3. As a consequence, the problem in part II of the domain is of the same type as in part I. Therefore, the results for V also hold for W in the ϵ-regularization.

The solution can be guaranteed in the sense of Definition 4.2.1 and Definition 4.2.2, where we replace ψ_I by $\psi_{I,\epsilon}$, ψ_{II} by $\psi_{II,\epsilon}$, a by $a+\epsilon$ and (4.51) by

$$\begin{aligned}&\int_I\int_0^1 \psi_{II,\epsilon}\partial_\tau\phi\,\mathrm{d}y\mathrm{d}\tau - \int_I\int_0^1 y\tfrac{\dot{s}(T-\tau)}{L-s(T-\tau)}\psi_{II,\epsilon}\partial_y(y\phi)\,\mathrm{d}y\mathrm{d}\tau\\ -&\int_I\int_0^1 \tfrac{1}{(L-s)^2}\partial_y\psi_{II,\epsilon}\partial_y\left[a_\epsilon\phi\right]\,\mathrm{d}y\mathrm{d}\tau = \int_I\int_0^1 \phi f(L-y(L-s(T-\tau)),T-\tau)\,\mathrm{d}y\mathrm{d}\tau.\end{aligned} \tag{4.111}$$

The variational solution $\psi_{I,\epsilon}$ and $\psi_{II,\epsilon}$ can be constructed (as a limit) by means of the same approximation scheme as used in Section 4.2.4 where a_ϵ is replaced by $a_\Delta+\epsilon$, see (4.94)

The convergence of the scheme will be in L_2 for both V and W, as opposed to the earlier result, Lemma 4.2.6. However, it remains to prove that for $\epsilon\to 0$ the solution converges to the one of the original problem where W converges in L_1, so $\psi_\epsilon\to\psi$.

The L_1 estimates required for W can be easily recovered. The estimates of Lemma 4.2.3 remains valid. Thus, we have

$$\int_0^T\int_0^1 \left|\partial_y W^{\Delta,\epsilon}(y,\tau)\right|\,\mathrm{d}y\mathrm{d}\tau \le C. \tag{4.112}$$

Furthermore, by the Cauchy-Schwarz inequality we have

$$\begin{aligned}&\int_0^{T-z}\int_0^Q \left|W^{\Delta,\epsilon}(y,\tau+z)-W^{\Delta,\epsilon}(y,\tau)\right|\,\mathrm{d}y\mathrm{d}\tau\\ &\le C\left(\int_0^{T-z}\int_0^Q \left(W^{\Delta,\epsilon}(y,\tau+z)-W^{\Delta,\epsilon}(y,\tau)\right)^2\,\mathrm{d}y\mathrm{d}\tau\right)^{1/2} \le Cz^{1/2},\end{aligned} \tag{4.113}$$

where we used the result (4.93) from Lemma 4.2.6, which is now valid for $W^{\Delta,\epsilon}$.

Taking the limit $|\Delta|\to 0$ and using Lebesgue's dominated convergence theorem (the $\psi^{\Delta,\epsilon}$ are bounded) we have for part I

$$\int_0^{T-z}\int_0^Q \left(\psi_{I,\epsilon}(y,\tau+z)-\psi_{I,\epsilon}(y,\tau)\right)^2\,\mathrm{d}y\mathrm{d}x\mathrm{d}\tau \le C(Q)z, \tag{4.114}$$

$$\int_0^T \int_0^Q \left(\partial_y \psi_{I,\epsilon}(y,\tau)\right)^2 \, \mathrm{d}y\mathrm{d}x\mathrm{d}\tau \le K(Q), \tag{4.115}$$

$$\int_0^T \int_0^Q \left(\psi_{I,\epsilon}(y+p,\tau) - \psi_{I,\epsilon}(y,\tau)\right)^2 \, \mathrm{d}y\mathrm{d}x\mathrm{d}\tau \le p^2 K(Q), \tag{4.116}$$

on account of (4.79)-(4.91). For part II, the inequalities (4.112) and (4.113) give

$$\int_0^T \int_0^1 \left|\partial_y \psi_{II,\epsilon}(y,\tau)\right| \, \mathrm{d}y\mathrm{d}\tau \le C, \tag{4.117}$$

$$\int_0^{T-z} \int_0^Q \left|\psi_{II,\epsilon}(y,\tau+z) - \psi_{II,\epsilon}(y,\tau)\right| \, \mathrm{d}y\mathrm{d}\tau \le C z^{1/2}. \tag{4.118}$$

More in detail, the inequality (4.117) is argued as follows. In (4.112) C is independent of Δ and ϵ. The analogue of (4.101) in part II is now

$$\epsilon \int_0^\tau \int_0^Q (Q-y)(\partial_y W^{\Delta,\epsilon})^2 \, \mathrm{d}y\mathrm{d}\tau \le C, \tag{4.119}$$

as $\delta(Q) = \epsilon$ and $K(Q) < C$ in part II. We deduce that ($\epsilon > 0$ is fixed),

$$\partial_y W^{\Delta,\epsilon} \underset{|\Delta|\to 0}{\rightharpoonup} \partial_y W^\epsilon, \quad \text{in } L_2(I \times (0,Q)),\ \forall Q < 1,$$

and

$$\int_0^T \int_0^Q \partial_y W^\epsilon \phi \, \mathrm{d}y\mathrm{d}\tau = \lim_{|\Delta|\to 0} \int_0^T \int_0^Q \partial_y W^{\Delta,\epsilon} \phi \, \mathrm{d}y\mathrm{d}\tau,$$

$\forall \phi \in L_\infty(Q_T) \subset L_2(I \times (0,Q))$. It holds that

$$\begin{aligned}
\int_0^T \int_0^Q \left|\partial_y W^\epsilon\right| \, \mathrm{d}y\mathrm{d}\tau &= \sup_{|\phi|_\infty \le 1} \int_0^T \int_0^Q \partial_y W^\epsilon \phi \, \mathrm{d}y\mathrm{d}\tau \\
&= \sup_{|\phi|_\infty \le 1} \lim_{|\Delta|\to 0} \int_0^T \int_0^Q \partial_y W^{\Delta,\epsilon} \phi \, \mathrm{d}y\mathrm{d}\tau \\
&\le ||\phi||_\infty \sup_{|\phi|_\infty \le 1} \lim_{|\Delta|\to 0} \int_0^T \int_0^Q \left|\partial_y W^{\Delta,\epsilon}\right| \, \mathrm{d}y\mathrm{d}\tau \\
&\le C.
\end{aligned}$$

By a similar argument as in the previous section (see the proof of Theorem 4.2.1), we have that $\partial_y W^\epsilon = \partial_y \psi_{II,\epsilon}$ in the weak sense.

The above estimates and a Kolmogorov compactness argument imply the compactness of $\{\psi_{I,\epsilon}\}_\epsilon$ in $L_2(I,(0,Q))$, $\forall Q<1$, and the compactness of $\{\psi_{II,\epsilon}\}_\epsilon$ in $L_1(I,(0,Q))$, $\forall Q<1$. Consequently, as in the previous section, we obtain

$$\psi_{I,\epsilon} \xrightarrow[\epsilon\to 0]{} \psi_I, \quad \text{for a.e. } (y,\tau)\in(0,T)\times(0,1),$$

(by the method of diagonalization) and,

$$\psi_{II,\epsilon} \xrightarrow[\epsilon\to 0]{} \psi_{II}, \quad \text{for a.e. } (y,\tau)\in(0,T)\times(0,1).$$

As $\psi_{II,\epsilon}$ is uniformly bounded, we also have $L_1(Q_T)$-convergence. Next, we may easily prove that $\{\psi_I,\psi_{II}\}$ is a variational solution to (4.30)-(4.32) in the sense of Definition 4.2.1 and Definition 4.2.2. This reasoning leads to the following theorem.

Theorem 4.2.2. *Let $\{\psi_{I,\epsilon},\ \psi_{II,\epsilon}\}$ be the variational solution of the ϵ-regularization, obtained as a limit of the numerical approximation $\{V^{\Delta,\epsilon},\ W^{\Delta,\epsilon}\}$, with $a\leftrightarrow a_\Delta+\epsilon$. Then $\{\psi_{I,\epsilon},\psi_{II,\epsilon}\}\to\{\psi_I,\psi_{II}\}$ for $\epsilon\to 0$ in $(L_2((0,T)\times(0,1))\times L_1((0,T)\times(0,1)))$, where $\{\psi_I,\psi_{II}\}$ is a variational solution of (4.30)-(4.32).*

Proof. In addition to the previous arguments it remains to prove that $\{\psi_I,\ \psi_{II}\}$ as a limit of $\{\psi_{I,\epsilon},\ \psi_{II,\epsilon}\}$ is a variational solution.

In the case of ψ_I the proof consists of the following steps. From (4.115) and from $\psi_{I,\epsilon}\to\psi_I$ in $L_2(I\times(0,1))$ it follows that,

$$\partial_y\psi_{I,\epsilon} \rightharpoonup \partial_y\psi_I, \quad \text{in } L_2(I\times(0,Q)) \text{ for any } Q<1.$$

Then, in the identity (cf. (4.50))

$$\int_I\int_0^1 \psi_{I,\epsilon}\partial_\tau\phi\,\mathrm{d}y\mathrm{d}\tau - \int_I\int_0^1 y\frac{\dot{s}}{s}\partial_y\psi_{I,\epsilon}\phi\,\mathrm{d}y\mathrm{d}\tau \\ - \int_I\int_0^1 \frac{1}{s^2}\partial_y\psi_{I,\epsilon}\partial_y\left[a_\epsilon\phi\right]\,\mathrm{d}y\mathrm{d}\tau = \int_I\int_0^1 f\phi\,\mathrm{d}y\mathrm{d}\tau, \quad (4.120)$$

we may pass to the limit $\epsilon\to 0$, noticing that $\partial_y\left[a_\epsilon\phi\right]\to\partial_y\left[a\phi\right]$ in $L_2(I\times(0,Q))$ ($(0,Q)$ is the support of ϕ). It follows that $\psi_{I,\epsilon}$ can be replaced by ψ_I giving (4.50).

In the case of part II, take the limit $\epsilon\to 0$ in (4.111), where $a_\epsilon\equiv\epsilon$. Estimate the last term on the lhs by

$$J_\epsilon := \left|\epsilon\int_I\int_0^1 \frac{1}{(L-s)^2}\partial_y\psi_{II,\epsilon}\partial_y\phi\,\mathrm{d}y\mathrm{d}\tau\right| \le \epsilon C\int_I\int_0^Q \left|\partial_y\psi_{II,\epsilon}\right|\,\mathrm{d}y\mathrm{d}\tau \le \epsilon C, \quad (4.121)$$

due to (4.117). Moreover, we use the convergences $\psi_{II,\epsilon} \to \psi_{II}$ and $J_\epsilon \to 0$ for $\epsilon \to 0$ in (4.111). This completes the proof. □

Remark 4.2.8. *The problem (4.30) has been regularized by a viscocity term $\epsilon\partial_x^2\psi$. The corresponding viscocity solution $\{\psi_{I,\epsilon},\ \psi_{II,\epsilon}\}$ converges to the variational one, which we define as a viscosity solution. In the theory of Kružkov the uniqueness of the entropy (viscocity) solution is proved. However, our elliptic part creates the convective term which is singular (because of $\partial_x a(s(t), t) = -\infty$) and this is not included there. In any case our numerical approximation converges to the viscosity solution (as defined here) under the assumption that the weak solution to (4.30)-(4.32) is unique. Our convergence result is only up to a subsequence of $\{\Delta_l\}$. If the variational solution is unique, then the original sequence $\{V^{\Delta_l}, W^{\Delta_l}\}$ is converging for $l \to \infty$.*

4.3 Discrete time measurements

An adjoint method has been developed for inverse annealing diffusion problems. The adjoint method works well, as is confirmed by experiment 4.4.1, and from the experiments in the next Chapter.

However, in practice for diffusion annealing an extra difficulty is encountered. The cost to get experimental values $C^*(x,t)$ over the entire setup time from 3 to 6 hours is prohibitive. A version is needed that takes discrete measurement times into account, i.e. $C^*(x,t_i)$, $i = 1,\ldots,L$. Typically, measurements are obtained at times 5min, 30min, 1h, 3h. This does not allow to construct e.g. a piecewise linear approximation in time of $f(x,t)$.

The above theory can be adapted by using in stead (4.10) a cost functional given by

$$\begin{aligned}\widetilde{\mathcal{F}} &= \sum_{i=1}^{M}\int_\Omega [C(x,t_i,\mathbf{p}) - C^*(x,t_i)]^2\,\mathrm{d}x \\ &= \int_0^T\int_\Omega [\sum_{i=1}^{M}\delta(t-t_i)][C(x,t,\mathbf{p}) - C^*(x,t)]^2\,\mathrm{d}x\mathrm{d}t,\end{aligned}$$

where δ is the Dirac function. Theorem 4.1.1 changes slightly, in that the rhs of (4.11) is now given by $-\widetilde{f}(x,T-\tau)$ with

$$\widetilde{f}(x,t) = \sum_{i=1}^{M} 2\left(C(x,t_i,\mathbf{p}) - C^*(x,t_i)\right)\delta(t-t_i).$$

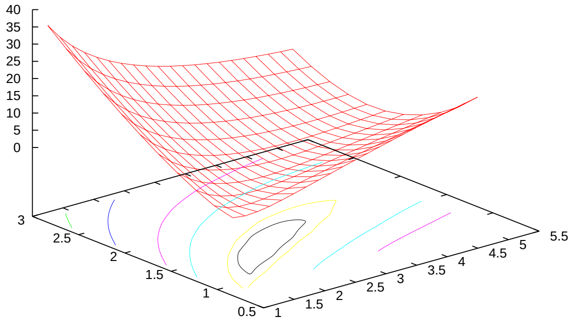


Figure 4.5: Function $\chi^2(\boldsymbol{p})$ for $\hat{\boldsymbol{p}} = (2.5, 1.5)$. Left is the p_2-axis, right the p_1-axis. The contour lines are for the values 0.5, 1, 5, 10, 20 and 30.

This makes the implementation of the costate method more complicated, as the adjoint equation has become a convection-diffusion equation, with reaction only at specific timesteps. We plan to implement the procedure for the case of discrete time measurements in the near future. In this thesis we apply the Levenberg-Marquardt method as an inverse method since this can be easily implemented, see Section C.4.4. In the numerical experiments, we demonstrate the effectiveness of the continuous time version of the developed costate method.

4.4 Numerical experiments

4.4.1 Experiment 1: Adjoint method

We consider again the Barenblatt-Pattle problem. We take the diffusion coefficient $D(C) := (p+1)C^p$. The function $C^*(x,t)$, (4.5), is taken to be the exact solution given by (3.16). Starting from an initial diffusion coefficient $D(C, \boldsymbol{p}) = p_1 C^{p_2}$, $\boldsymbol{p} = (p_1, p_2)$, the method is expected to converge to the exact solution $\hat{\boldsymbol{p}} = (p+1, p)$.

To evaluate the algorithm, we compare the calculated value of $\nabla_p \mathcal{F}(\boldsymbol{p})$, (4.29), in some points, with the plot of the $\chi^2(\boldsymbol{p})$-function defined by

$$\chi^2(\boldsymbol{p}) = \sum_{j=1}^{M}\sum_{i=0}^{N}\left(C(x_i,t_j,\boldsymbol{p}) - C(x_i,t_j,\hat{\boldsymbol{p}})\right)^2,$$

where $C(x_i, t_j)$ is the approximated solution to (4.2)-(4.4) that is obtained with our numerical model. This comparison technique is valid, as χ^2 will have the same qualitative behaviour as $\mathcal{F}$, (4.10).

In Fig. 4.5 we have plotted $\chi^2(\boldsymbol{p})$ for $M = 60$ and $N = 100$, in a sample of width $X = 15$, that underwent diffusion during 60 sec. We can distinguish two regions: an ellipsoidal valley floor where $\chi^2(\boldsymbol{p}) < 1$, with midpoint being the minimum $\boldsymbol{p} = (2.5, 1.5)$, and the region where $1 < \chi^2(\boldsymbol{p})$. In the valley floor, the gradient should be very small. Therefore, small errors can lead to a direction of the gradient which does not point away from the minimum, as it should. Most inverse methods will fail once the parameter values are within this region. In the remaining part of the region, the gradient should be perpendicular to the contour lines, and pointing away from the minimum.

$N = 101$, $\Delta t = 0.1$s	p_1	p_2	$-\nabla_p\mathcal{F}$	RMS
1	2.5	1	(-2.54, 8.73)	0.0556
2	2.5	1.5	(0.222, -0.511)	0.0214
3	2.5	2	(2.27, -5.82)	0.0495
4	2.5	2.5	(3.20, -7.07)	0.0826
5	3.5	1.5	(-1.27, 5.18)	0.0364
6	4.5	1.5	(-1.43, 7.55)	0.0538
7	4	1	(-2.35, 12.8)	0.0855
$N = 202$, $\Delta t = 0.05$s	p_1	p_2	$-\nabla_p\mathcal{F}$	RMS
8	2.5	1.5	(0.124, -0.410)	0.0159
9	2.5	2.5	(3.26, -7.37)	0.0804

Table 4.1: Values of $-\nabla_p\mathcal{F}$ and root-mean square error for several parameter-sets.

In Table 4.1 several values of $-\nabla_p\mathcal{F}$ are presented. These were obtained by means of the developed numerical method, where $\Delta t = 0.1$sec and where an equidistant grid with 101 gridpoints was used. Here, Δt does not refer to the used time discretization, as this is done automatically in our stiff ODE solver, but only to the discretization of $-\nabla_p\mathcal{F}$, (4.29). All values correspond with the expected ones from Fig. 4.5. This indicates that all inverse methods which are based on the obtained value for $\nabla_p \mathcal{F}$ will converge to $\hat{\boldsymbol{p}}$.

To illustrate the effect of grid refinement, we also present in Table 4.1 the case of an equidistant grid with 202 gridpoints. The change in $-\nabla_p \mathcal{F}$ is small, but the root-mean square error (RMS) at $(2.5, 1.5)$ clearly decreases.

4.4.2 Experiment 2: Levenberg-Marquardt

In Fig. 3.10 an apparent diffusion coefficient is given that was obtained using the Levenberg-Marquardt method on the data depicted in Fig. 3.11. The diffusion coefficient was constructed with 8 couples (C_k, p_k), where $C_1 = 0.8\,10^{-5}$ and $C_8 = 1.7, 10^{-5}$. As initial values for the parameters p_k we took the diffusion values from [66], the Si-Fe interdiffusion (0 at%Al). This Si-Fe interdiffusion is also shown in Fig. 3.10.

Chapter 5

Inverse problems for the dual-well experiment

5.1 Subsurface parameter identification

In Chapter 2, an efficient numerical method has been developed to solve convection dominated diffusion problems, and this was applied to the the dual-well. The importance of dual-well experiments is the determination of the properties of the subsurface, see Appendix B for an overview.

We demonstrated that the measured BTC shows a clear response to the value of a number of parameters, mainly the longitudinal dispersivity α_L, the conductivity k and the parameters arising in the adsorption isotherms. In subsurface modeling, common experiments to determine dispersivities or other parameters are the following.

- The column experiment. Water with tracer concentration C_0 is injected into a column. The concentration is measured at the other end. This is a laboratory scale experiment of range 1m.

- Single well injection-extration method. This is a global scale experiment of range 2 to 4 m. A radioactive tracer like ^{131}I or ^{34}Br is injected into a well for a time t_0. Then fresh water is injected to push for some time, after which water is extracted from the well. Probes at different depths record the concentration over time.

- Dual-well or multi-well tests. At a range of 4 to 20 m, a single well cannot control the velocity field, so dual-well tests or multi-well tests are used. In the dual-well extraction-injection tests, two wells are drilled about 10 meters apart. Water is pumped from one well and injected into the other at the same rate, so as to form a steady flow field. After steady state conditions are achieved, a tracer is added into the injection well, either continously or in a pulse. By sampling from the extraction well, we obtain a curve of the tracer concentration versus time. Observation wells can be added to measure the concentration in different places around the extraction well. An extension of this method form the multi-well tests, that use more than two wells.
- Global multi-well tests. On a range of 20 to 100 m, multi-well systems can still be used, but loose their accuracy, as the natural flow field will be dominant. This can be solved by drilling two parallel rows of wells, perpendicular to the natural flow field. A steady and uniform one-dimensional flow will appear between the two lines, after which a tracer can be injected into the aquifer from a well in the middle of the row.

In all of these experiments, one obtains a so called break through curve (BTC) from which all information needs to be extracted.

For the inverse problem we shall measure the time evolution of the concentration $C^{(1)}(t)$ in the extraction well depending on the concentration evolution $C^{(2)}(t)$ in the injection well, see Chapter 2. The average concentration at outflow is the advective tracer mass flow per second in the well over the volume of water flowing per second. The mass in the subsurface is the mass in the xy-plane times the height and porisity, $\theta_0 h_{\text{eff}}$, see Section 2.2. Therefore, denoting by p the parameters on which the model, and hence the break through curve, depends,

$$\begin{aligned}
C_p^{(1)}(t) &= \frac{\int_{\delta B_{r_1}(-d,0)} \theta_0 h_{\text{eff}} C_p(x,y,t)(\boldsymbol{n}\cdot\boldsymbol{v})\,ds}{\int_{\delta B_{r_1}(-d,0)} \theta_0 h_{\text{eff}}(\boldsymbol{n}\cdot\boldsymbol{v})\,ds} \\
&= \frac{\int_{\delta B_{r_1}(-d,0)} C_p(x,y,t)(\boldsymbol{n}\cdot\nabla\Phi)\,ds}{\int_{\delta B_{r_1}(-d,0)} (\boldsymbol{n}\cdot\nabla\Phi)\,ds} \\
&= \frac{\int_0^\pi \frac{2(\cosh v^{(1)}-\cos u)}{\delta} C_p(u,v^{(1)},t)(\partial_v\tilde{\Phi}(v))\sqrt{(x'(u))^2+(y'(u))^2}du}{\int_0^\pi \frac{2(\cosh v^{(1)}-\cos u)}{\delta}(\partial_v\tilde{\Phi}(v))\sqrt{(x'(u))^2+(y'(u))^2}du} \\
&= \frac{1}{\pi}\int_0^\pi C_p(u,v^{(1)},t)du.
\end{aligned} \tag{5.1}$$

Here $\boldsymbol{n}$ denotes the outward normal vector on the boundary $\partial B_{r_1}(-d,0)$ of the circle with radius r_1 centred at $(-d,0)$. The index p is the model parameter vector, e.g. $\boldsymbol{p}=(k,\alpha_L,\alpha_T)$. If $\boldsymbol{p}$ is given, $C_p^{(1)}(t)$ can be computed from (2.44)-(2.48). Thus (5.1) is the only measurement information that we shall use in the inverse problem. Note that this type of measurement is very different from to those in Chapter 4, where experimental values over the entire domain where given. Here, only an integral over a part of the boundary is known.

The approximation scheme developed in Chapter 2 was accurate and fast. These are the main characteristics needed to solve inverse problems. Accurate but slow schemes will fail to obtain answers in a reasonable time, whereas fast but less accurate methods will provide non-reliable answers.

We now present how the developed scheme can be optimally used for parameter identification, invoking the techniques explained in Appendix C. We start with the Levenberg-Marquardt method applied to the dual-well. Next, we develop a costate method for this set-up. These results have been published in [58, 41]. Recently, the adjoint technique has been extended to non-equilibrium adsorption and to the use of a different penalty function by J. Kačur and J. Babušíková in [36].

5.2 Use of the Levenberg-Marquardt method

For the inverse problem we first use the Levenberg-Marquardt method. As our parameters are in $\mathbb{R}^+\backslash\{0\}$ we optimize the logarithm of the parameters instead of the parameters themselves. This has the benefit that the parameters cannot attain negative values, and that the stepsize is always relative. Thus, we introduce a penalty function

$$\mathcal{F}(\boldsymbol{p})=\sum_{i=1}^{N}(C_p^{(1)}(t_i)-\hat{C}^{(1)}(t_i))^2,$$

where $\hat{C}^{(1)}$ is the measured BTC and $C_p^{(1)}$ is the concentration in the extraction well corresponding to a parameter set $\boldsymbol{p}=(p_1,\dots,p_n)$ and obtained numerically from (2.44)-(2.48). Starting from an initial parameter set $\boldsymbol{p}^k$, a new set is given by

$$p_j^{k+1}=p_j^k+\left[\frac{1}{J_k^TJ_k+\lambda I}(J_k^TD_k)\right]_j p_j^k. \tag{5.2}$$

Here, $(J_k)_{ij} = p_j \partial_{p_j} C^{(1)}_{p^k}(t_i)$ the logarithmic Jacobian, $(D_k)_i = C^{(1)}_p(t_i) - \hat{C}^{(1)}(t_i)$ and λ is a parameter which is initially chosen equal to $\mathrm{Tr}(J_k^T J_k)/\mathrm{Tr}(I)$, see [70]. If the new parameter set gives rice to a smaller penalty function value, it is retained, and λ is divided by an ever increasing integer so as to get quadratic convergence. In the other case, the parameter set is discarded, and another set is sought with a larger λ value. The process is stopped when λ becomes larger than a preset value $\lambda_{\max}$.

The main advantage of the Levenberg-Marquardt method is its robustness. However, the numerical determination of the Jacobian can have a heavy price, especially when the solution of the direct problem is elaborate, and there are many parameters. For every parameter p_i, an extra direct run with $p_i + \Delta p_i$ is needed to obtain $(J_k)_{ij}$.

5.3 Use of the adjoint equation method

As mentioned before, the adjoint equation method determines the gradient of the cost functional in terms of the parameters. This gradient can then be used in an iteration based method as for example conjugate gradient methods, De Broyden methods, etc.

For our convenience we present the deduction of the adjoint system in a slightly less complex setting. We assume $D_0 = 0$ and $\alpha_T = 0$. This is feasible as these two parameters are very small compared to the other parameters, and therefore have only a small influence on the BTC. Observe that it is not feasible to determine these two parameters in the experiment used to obtain sorption isotherms. Therefore, they do not play a role in the construction of the adjoint system.

From (2.62), the problem to be solved is

$$\partial_t \left(C + \mathcal{B}(\boldsymbol{p}_1, C)\right) = g(u,v)\partial_v(2\boldsymbol{p}_2 \lambda A \partial_v C)\} + g(u,v) A \gamma \partial_v C, \tag{5.3}$$

where $\boldsymbol{p}_1$ are the parameters in the time derivative, e.g. for Freundlich (K_0, q), and $\boldsymbol{p}_2$, e.g. (α_L), are the parameters in the diffusion part. Furthermore, $\mathcal{B}(\boldsymbol{p}_1, C) = \frac{\varrho}{\theta_0}\Psi(C)$ and g is a known function depending on u and v:

$$g = \frac{4\lambda^2}{\gamma^3 \theta_0 h_{\text{eff}}(v)}, \quad \lambda = \cosh v - \cos u$$

The boundary conditions become, see Chapter 2,

$$\left(\frac{2\alpha_L\lambda A}{\gamma}\right)\partial_v C + AC = AC_0(t) \quad \text{on } \Gamma_1, \tag{5.4}$$

$$\partial_u C = 0 \text{ on } \Gamma_2 \cup \Gamma_4, \quad \partial_v C = 0 \quad \text{on } \Gamma_3. \tag{5.5}$$

The parameters $\boldsymbol{p} = (\boldsymbol{p}_1, \boldsymbol{p}_2)$ must be retrieved by the inverse method.

We consider the penalty function,

$$\mathcal{F}(\boldsymbol{p}) = \int_0^T (C_p^{(1)}(t) - \hat{C}^{(1)}(t))^2 \,\mathrm{d}t \approx \sum_{i=1}^{N} \Delta t_i (C_p^{(1)}(t_i) - \hat{C}^{(1)}(t_i))^2. \tag{5.6}$$

We now deduce the corresponding costate method.

5.3.1 Deduction of the adjoint system

We prove the following.

Theorem 5.3.1. *Let $C(u,v,t,\boldsymbol{p})$ be the solution to the problem (5.3)-(5.5), where $\mathcal{B}$ is a smooth function, $t \in (0,T)$ and $\mathcal{F}$ is defined by (5.6). Let $\overline{\psi}(x,\tau)$ be the solution of the following convection-diffusion equation*

$$\begin{aligned}(1 + \mathcal{B}'(\boldsymbol{p}_1, C(u,v,T-\tau)))\,\partial_\tau\overline{\psi}(u,v,\tau) - \partial_v 2\lambda A p_2 \partial_v(g(u,v)\overline{\psi}(u,v,\tau))& \\ + A\gamma\partial_v(g(u,v)\overline{\psi}(u,v,\tau)) = 0,&\end{aligned} \tag{5.7}$$

where $\mathcal{B}' = \frac{d\mathcal{B}(p,s)}{ds}$, $\boldsymbol{p} = (\boldsymbol{p}_1, p_2)$ and $p_2 = \alpha_L$, along with the boundary conditions

$$\partial_v(g\overline{\psi}) = 0, \ \textit{on } \Gamma_1, \tag{5.8}$$

$$\partial_u\overline{\psi} = 0, \ \textit{along } \Gamma_2 \textit{ and } \Gamma_4, \tag{5.9}$$

$$2\lambda A p_2 \partial_v(g\overline{\psi}) - gA\gamma\overline{\psi} = \frac{1}{\pi^2}\int_0^\pi \left(2\left(C(u,v,T-\tau) - \hat{C}^{(1)}(T-\tau)\right)\right) \mathrm{d}u, \tag{5.10}$$

on Γ_3, together with the initial condition

$$\overline{\psi}(u,v,0) = 0, \quad (u,v) \in [u^{(1)}, u^{(2)}] \times [v^{(1)}, v^{(2)}]. \tag{5.11}$$

Then, one has

$$\begin{aligned}\nabla_p\mathcal{F} = \Bigg(-\int_0^T\int_\Omega \partial_t\psi\nabla_{p_1}\mathcal{B}\,\mathrm{d}V\mathrm{d}t - \int_\Omega \nabla_{p_1}\mathcal{B}(0)\psi(0)\,\mathrm{d}V,& \\ \int_0^T\int_\Omega 2\lambda A(\partial_v C)\partial_v(g(u,v)\psi)\,\mathrm{d}V\mathrm{d}t\Bigg),&\end{aligned} \tag{5.12}$$

where $\psi(u,v,t) = \overline{\psi}(u,v,T-t)$.

Proof. As a first step we deduce the variation $\delta\mathcal{F}$ of the cost functional:

$$\begin{aligned}
\delta\mathcal{F} &= F(\boldsymbol{p}+\delta\boldsymbol{p}) - F(\boldsymbol{p}) \\
&= \int_0^T \left[(C^{(1)}_{p+\delta p}(t) - \hat{C}^{(1)}(t))^2 - (C^{(1)}_{p}(t) - \hat{C}^{(1)}(t))^2\right] \mathrm{d}t \\
&= \int_0^T \left[{C^{(1)}_{p+\delta p}}^2(t) - {C^{(1)}_{p}}^2(t) - 2\hat{C}^{(1)}(t)\left(C^{(1)}_{p+\delta p} - C^{(1)}_{p}(t)\right)\right] \mathrm{d}t \\
&= \int_0^T \left(C^{(1)}_{p+\delta p}(t) - C^{(1)}_{p}(t)\right)\left(C^{(1)}_{p+\delta p} + C^{(1)}_{p}(t) - 2\hat{C}^{(1)}(t)\right) \mathrm{d}t \\
&= \int_0^T \left[\frac{1}{\pi}\int_0^\pi \delta C_p(u,v^{(1)})\,\mathrm{d}u\right]\left[\frac{1}{\pi}\int_0^\pi \left(\delta C_p(u,v^{(1)}) + \right.\right. \\
&\qquad\qquad \left.\left. 2\left(C_p(u,v^{(1)}) - \hat{C}^{(1)}(t)\right)\right)\mathrm{d}u\right] \mathrm{d}t.
\end{aligned}$$

Here, we have omitted that all concentrations depend on t. Dropping higher order terms, we have

$$\delta\mathcal{F} = \int_0^T \int_0^\pi K(t)\delta C_p(u,v^{(1)},t)\,\mathrm{d}u\mathrm{d}t, \tag{5.13}$$

up to first order, where

$$K(t) = \frac{1}{\pi^2}\int_0^\pi \left(2\left(C_p(u,v^{(1)},t) - \hat{C}^{(1)}(t)\right)\right)\mathrm{d}u.$$

In the next step the equation in variations is used to eliminate δC from (5.13). First note that the solution corresponding to the parameters $\boldsymbol{p}+\delta\boldsymbol{p}$ is $C+\delta C$ and satisfies (5.3)-(5.13). Substract equation (5.3) for $C+\delta C$ from the equation for C. Using

$$\mathcal{B}(\boldsymbol{p}_1+\delta\boldsymbol{p}_1, C+\delta C) = \mathcal{B}(p_1,C) + \mathcal{B}'(p_1,C)\delta C + \nabla_{p_1}\mathcal{B}\cdot\delta\boldsymbol{p}_1 + \text{H.O.T},$$

where $\mathcal{B}' = \frac{\partial\mathcal{B}}{\partial C}$, and neglecting higher order terms, we arrive at

$$\begin{aligned}
\partial_t\left\{\delta C + \mathcal{B}'(\boldsymbol{p}_1,C)\delta C + \nabla_{p_1}\mathcal{B}\cdot\delta\boldsymbol{p}_1\right\} &= g(u,v)\partial_v\left(2\lambda A\left[\delta p_2\partial_v C + p_2\partial_v\delta C\right]\right) \\
&\quad + g(u,v)A\gamma\partial_v\delta C, \qquad (5.14)
\end{aligned}$$

with boundary conditions

$$\frac{2\lambda A}{\gamma}\left[p_2\partial_v\delta C+\delta p_2\partial_v C\right]+A\delta C=0 \text{ on } \Gamma_1, \quad \partial_\nu\delta C=0, \text{ elsewhere,} \tag{5.15}$$

and initial condition

$$\delta C(u,v,0)=0. \tag{5.16}$$

Eq. (5.14)-(5.16) constitute *the problem in variations.* We multiply (5.14) by a smooth function $\psi(u,v,t)$, to be specified below, and integrate over $\Omega\times(0,T)$. Integration by parts gives

$$\begin{aligned}
&\int_\Omega \psi\left(\delta C+\mathcal{B}'(\boldsymbol{p}_1,C)\delta C+\nabla_{p_1}\mathcal{B}\cdot\delta\boldsymbol{p}_1\right)\Big|_0^T \mathrm{d}V\\
&\qquad -\int_0^T\int_\Omega(\partial_t\psi)\left[\delta C+\mathcal{B}'(\boldsymbol{p}_1,C)\delta C+\nabla_{p_1}\mathcal{B}\cdot\delta\boldsymbol{p}_1\right]\mathrm{d}V\mathrm{d}t\\
= &\quad -\int_0^T\int_\Omega 2\lambda A\left[\delta p_2\partial_v C+p_2\partial_v\delta C\right]\partial_v(g(u,v)\psi)\,\mathrm{d}V\mathrm{d}t\\
&\qquad +\int_0^T\int_0^\pi g(u,v)2\lambda A\left[\delta p_2\partial_v C+p_2\partial_v\delta C\right]\psi\Big|_{v^{(1)}}^{v^{(2)}}\mathrm{d}u\mathrm{d}t\\
&\qquad -\int_0^T\int_\Omega A\gamma\delta C\partial_v(g(u,v)\psi)\,\mathrm{d}V\mathrm{d}t+\int_0^T\int_0^\pi g(u,v)\psi A\gamma\delta C\Big|_{v^{(1)}}^{v^{(2)}}\mathrm{d}u\mathrm{d}t.
\end{aligned}$$

We perform integration by parts on the term that contains $\partial_v\delta C$. We invoke the BC and IC. Furthermore, we restrict the auxiliary function $\psi(u,v,t)$ so as to obey the condition

$$\psi(u,v,T)=0.$$

This leads to

$$\begin{aligned}
&\int_0^T \int_\Omega \delta C \partial_t \psi \left(1 + \mathcal{B}'(\boldsymbol{p}_1, C)\right) \, \mathrm{d}V \mathrm{d}t + \int_0^T \int_\Omega \partial_t \psi \nabla_{p_1} \mathcal{B} \, \mathrm{d}V \mathrm{d}t \cdot \delta \boldsymbol{p}_1 \\
&\quad + \int_\Omega \nabla_{p_1} \mathcal{B}(0) \psi(0) \, \mathrm{d}V \cdot \delta \boldsymbol{p}_1 - \int_0^T \int_\Omega 2\lambda A (\partial_v C) \partial_v (g(u,v)\psi) \, \mathrm{d}V \mathrm{d}t \; \delta p_2 \\
&\quad + \int_0^T \int_\Omega \delta C \left[\partial_v 2\lambda A p_2 \partial_v (g(u,v)\psi)\right] \, \mathrm{d}V \mathrm{d}t \qquad (5.17) \\
&\quad - \int_0^T \int_0^\pi \delta C 2\lambda A p_2 \partial_v (g(u,v)\psi) \Bigg|_{v^{(1)}}^{v^{(2)}} \mathrm{d}u \mathrm{d}t \\
&\quad + \int_0^T \int_0^\pi g(u,v) \gamma A (-\delta C) \psi \Bigg|_{v = v^{(2)}} \mathrm{d}u \mathrm{d}t \\
&\quad - \int_0^T \int_\Omega \delta C \partial_v (g(u,v)\psi) A \gamma \, \mathrm{d}V \mathrm{d}t + \int_0^T \int_0^\pi \delta C g(u,v) \psi A \gamma \Bigg|_{v^{(1)}}^{v^{(2)}} \mathrm{d}u \mathrm{d}t = 0.
\end{aligned}$$

By choosing appropriate constraints on the function ψ this can be further simplified. Indeed, if ψ is taken so that it satisfies

$$\left(1 + \mathcal{B}'(\boldsymbol{p}_1, C)\right) \partial_t \psi + \partial_v 2\lambda A p_2 \partial_v (g(u,v)\psi) - \partial_v (g(u,v)\psi) A \gamma = 0, \qquad (5.18)$$

together with the boundary condition:

$$\partial_v (g(u, v^{(2)})\psi) = 0, \quad \text{on } \Gamma_1 \qquad (5.19)$$

then (5.17) simplifies to

$$\begin{aligned}
&\int_0^T \int_\Omega \partial_t \psi \nabla_{p_1} \mathcal{B} \, \mathrm{d}V \mathrm{d}t \cdot \delta \boldsymbol{p}_1 + \int_\Omega \nabla_{p_1} \mathcal{B}(0) \psi(0) \, \mathrm{d}V \cdot \delta \boldsymbol{p}_1 \\
&\quad - \int_0^T \int_\Omega 2\lambda A (\partial_v C) \partial_v (g(u,v)\psi) \, \mathrm{d}V \mathrm{d}t \; \delta p_2 \qquad (5.20) \\
&\quad + \int_0^T \int_0^\pi \delta C(u, v^{(1)}, t) \left[2\lambda A p_2 \partial_v (g(u, v^{(1)})\psi) - g(u, v^{(1)}) A \gamma \psi\right] \mathrm{d}u \mathrm{d}t = 0.
\end{aligned}$$

Comparing this equation with (5.13), we see that the last integral is equal to $\delta \mathcal{F}$, if a last boundary condition is imposed on ψ on the outflow boundary, viz.

$$\left[2\lambda A p_2 \partial_v (g(u, v^{(1)})\psi) - g(u, v^{(1)}) A \gamma \psi(u, v^{(1)}, t)\right] = K(t) \quad \text{on } \Gamma_3 \qquad (5.21)$$

where

$$K(t) = \frac{1}{\pi^2} \int_0^{\pi} \left(2 \left(C_p(u, v^{(1)}, t) - \hat{C}^{(1)}(t) \right) \right) \mathrm{d}u.$$

Consequently, the boundary condition on Γ_3 is independent on u. Note that up to now we have no BC on Γ_2 and Γ_4. This was to be expected since convection and diffusion happen along the characteristics (v-lines) due to the used transformation and due to the fact that $\alpha_T = 0 = D_0$. To complete the system, the following two BC are added:

$$\partial_u \psi = 0, \text{ along } \Gamma_2 \text{ and } \Gamma_4. \tag{5.22}$$

The choice is based on the same symmetry reasons that are used for the BC of the direct problem on these edges. □

5.3.2 Numerical approximation

Problem (5.7)-(5.11) is a variable coefficient convection-diffusion problem, with coefficients depending on space and time. As flow and diffusion are along v-lines, the problem is split in independent v-strips. Each subproblem will then be solved independently as a 1-dimensional convection-diffusion problem. We choose the method of lines for the numerical approximation. Setting $u = u_i$, consider a partition $\{v_j\}$, $(j = 0, \ldots, M)$, of the v-strip, where $v_0 = v^{(1)}$ and $v_M = v^{(2)}$. Denote $\psi_j(\tau) \approx \overline{\psi}(u_i, v_j, \tau)$ and let $l_2(v, j)$ stand for the Lagrange polynomial of the second order interpolating the points (v_{j-1}, ψ_{j-1}), (v_j, ψ_j) and (v_{j+1}, ψ_{j+1}). Then, approximate $\partial_v \overline{\psi}$ by $dl_2(v_j, j)/dy = (dl_2(v, j)/dy)_{v=v_j}$ and $\partial_v^2 \overline{\psi}$ by $d^2 l_2(v_j, j)/dy^2 = (d^2 l_2(v, j)/dy^2)_{v=v_j}$. For the Robin BCs, the governing PDE is extended to the boundary points and is approximated similarly as in the inner points. To this end, introduce fictive points v_{-1} and v_{M+1} and assign to them the values that ψ obtained from the discretization of the Robin BC by means of the second order Lagrange interpolant in the border points. This leads to the system of ODEs

$$\begin{aligned} \frac{d}{d\tau} \psi_j(\tau) - \frac{A}{1 + \mathcal{B}'} \left(2p_2 \left(\partial_v(\lambda g) + \lambda \partial_v g \right) - g\gamma \right) \frac{dl_2(v_j, j)}{dy} - \frac{2Ap_2 \lambda g}{1 + \mathcal{B}'} \frac{d^2 l_2(v_j, j)}{dy^2} \\ = \frac{A}{1 + \mathcal{B}'} \left(2p_2 \partial_v (\lambda \partial_v g) - \gamma \partial_v g \right) \psi_j(\tau), \end{aligned} \tag{5.23}$$

with $j = 0, \ldots, M$, $(u, v) = (u_i, v_j)$, along with

$$2\lambda A p_2 g \frac{dl_2(v_0, 0)}{dy} + (2\lambda A p_2 \partial_v g - gA\gamma)\psi_0(\tau) = K(T - \tau), \text{ with } (u, v) = (u_i, v_0), \tag{5.24}$$

and

$$g\frac{dl_2(v_M, M)}{dy} + (\partial_v g)\psi_M(\tau) = 0, \text{ with } (u, v) = (u_i, v_M), \tag{5.25}$$

This gives M+3 equations for the M+3 unknowns, $(j = -1, 0, \ldots, M+1)$. The system (5.23)-(5.25) is solved by a standard package for stiff ODE, e.g. LSODA. For the numerical realization of (5.12), let us first note that for Freundlich adsorption it holds that $\nabla_{p_1}\mathcal{B} = (C^q, K_0 C^q \ln(C))$. In the case of initial condition (2.48) one has $\nabla_{p_1}\mathcal{B}(0) = 0$, simplifying (5.12). Then, numerical approximations for C, ψ, $\partial_t\psi$, $\partial_v\psi$ and $\partial_v C$ lead to $\nabla\mathcal{F}$.

We refer to Appendix C for the construction of iteration schemes based on $\nabla\mathcal{F}$, like the conjugate gradient method that we will use in the experiments. Different methods to determine this gradient can be used for the purpose of comparison and they are also discussed there.

5.4 Numerical experiments

5.4.1 Without adsorption

For these linear inverse problems we only apply the Levenberg-Marquardt method, as the direct computation is very fast, and the number of parameters is low. We shall consider a “standard” example with the following defining data: the wells have each radius $r_1 = r_2 = 15$cm and their centers are placed 10m from each other ($d = 5$m, $c = 0$). The height of the aquifer is $H = 10$m, the porosity of the soil is $\theta_0 = 0.2$ and the hydraulic conductivity is $k = 10^{-5}$m/s $=$ 0.864m/day. The longitudinal dispersivity α_L, its transversal counterpart α_T, as well as the prescribed head value at the extraction well and at the injection well will take various values in several experiments.

The computational grid, for 40×40 equally spaced nodal points in the transformed (u, v) variables, is plotted (transformed back to Cartesian coordinates) in Fig. 2.10.

For the inverse problem experiments, a pulse type of injection is used: the concentration is constant $C^{(2)}(t) = C_0$ for 1 day and then is set to 0. For ease of presentation, we take $C_0 = 1$. The tracer is assumed not to decay and molecular

k	α_L	α_T	RMS
1.7280	0.20000	0.02000	0.06152
1.3903	0.21460	0.02008	0.04759
1.0360	0.24149	0.02017	0.02242
0.8140	0.25024	0.02019	0.008917
0.8604	0.18218	0.02013	0.005374
0.8629	0.10900	0.02002	0.0008367
0.8661	0.09925	0.01995	8.744e-05
0.8660	0.09902	0.01965	8.202e-05
0.8656	0.09922	0.01773	6.702e-05
0.8645	0.09978	0.01241	2.226e-05
0.8640	0.1000	0.01014	2.126e-06
0.8640	0.1000	0.01000	1.462e-07

Table 5.1: Successive estimated parameter values for the pulse input injection case, using only the extraction well averages, use of standard scheme.

diffusion is neglected. Resulting tracer concentrations in the extraction well for this type of injection have been shown in Fig. 2.13 and Fig. 2.15.

The Levenberg-Marquardt method is based on repetitive execution of the direct problem. If for this direct problem the standard scheme is applied, see Section 2.3, we use an 80×200 grid with operator splitting time step 0.05 days for the experiment and the inverse algorithm. If we use the benchmark scheme, see Section 2.3.4, then α_T is set equal to zero, and the method is applied on 80 strips, with $y_{\text{div}} = 200$ and time step 0.05 days. We always use the BTC over 18 days in the inverse experiments.

Experiment 1

We consider a BTC obtained for $\alpha_L = 0.1$, $\alpha_T = 0.01$, $h_1 = 4$m and $h_2 = 15$m. Recall that $k = 0.864$m/day. In Table 5.1 we display the successive values of the estimated parameters by the Levenberg-Marquardt method, when starting from the inaccurate initial values $k = 1.728$m/day, $\alpha_L = 0.2$m, $\alpha_T = 0.02$m. In Table 5.2 we display the same inverse experiment, except that now the benchmark method is used in the Levenberg-Marquardt method.

Observe that in general the inverse method gives convergence first to k, and to a lesser extend also to α_L. It starts to converge to the final α_T only

k	α_L	RMS
1.7280	0.20000	0.06210
1.4414	0.2106	0.05073
1.0994	0.2307	0.02822
0.8332	0.2509	0.008319
0.8496	0.1942	0.005873
0.8616	0.1171	0.001336
0.8639	0.1007	5.832e-05
0.8640	0.09999	1.037e-06
0.8640	0.10000	7.919e-08

Table 5.2: Successive estimated parameter values for the pulse input injection case, using only the extraction well averages, use of benchmark scheme.

after recovering these values. Therefore, we suggest the following strategy: first apply the benchmark method to determine k and α_L in a first approximation. Then, use the slower general method to refine k and α_L and determine α_T simultaneously.

Experiment 2

We investigate the stability of the solution of the ill-posed inverse problem by adding artificial noise to the measurement data. This noise is normally distributed with a standard deviation σ. Fig. 5.1 shows the result of this perturbation on the BTC for $\sigma = 0.001$ and $\sigma = 0.01$. In Table 5.3 we display the successive values of the estimated parameters, when starting from the inaccurate initial values $k = 1$m/day, $\alpha_L = 0.2$m, with $\sigma = 0.01$, for the benchmark method. We recover the exact values within 10%. The same result is obtained with the general method. The transversal dispersivity cannot be recovered, not even when $\sigma = 0.001$. Starting from the inaccurate initial values $k = 1$m/day, $\alpha_L = 0.2$m, $\alpha_T = 0.2$m, with $\sigma = 0.001$, we recovered $k = 0.8635$m/day, $\alpha_L = 0.1006$m and $\alpha_T = 0.004884$m. Using only the BTC at the extraction well does not allow us to recover the transversal dispersivity.

Experiment 3

A different approach has been used in Table 5.4, where we display the successive values of the estimated parameters using the perturbed BTC ($\sigma = 0.001$) from

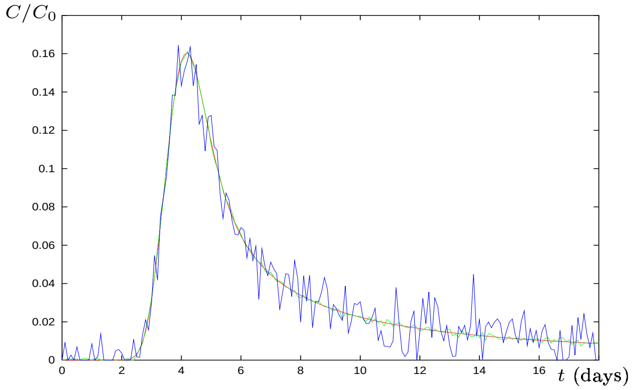


Figure 5.1: BTC of pulse input for $\alpha_L = 0.1$m and the result of artificial noise with $\sigma = 0.001$ and $\sigma = 0.01$.

k	α_L	RMS
1.0000	0.20000	0.02030
0.8889	0.2002	0.01124
0.8505	0.1971	0.01000
0.8499	0.1883	0.009788
0.8538	0.1638	0.009237
0.8600	0.1248	0.008622
0.8624	0.1108	0.008538
0.8627	0.1095	0.008536

Table 5.3: Successive estimated parameter values for the pulse input injection case, using the perturbed BTC ($\sigma = 0.01$), use of benchmark scheme.

k	α_L	α_T	RMS
1.0000	0.2000	0.02000	0.01844
0.8656	0.2000	0.01999	0.00648
0.8500	0.1938	0.01994	0.005953
0.8539	0.1782	0.01981	0.005193
0.8595	0.1420	0.01934	0.003219
0.8651	0.1059	0.01775	0.001109
0.8654	0.0997	0.01414	0.0009662
0.8648	0.0999	0.01094	0.0009640
0.8646	0.1000	0.01003	0.0009635
0.8646	0.1000	0.00996	0.0009634

Table 5.4: Successive estimated parameter values for the pulse input injection case, using the perturbed extraction well averages ($\sigma = 0.001$) and the perturbed concentration values at an extra point (−5.6m, 0m) ($\sigma = 0.0002$).

the extraction well and the perturbed breakthrough data ($\sigma = 0.0002$) at an extra point. Note that the maximum concentration reached at this extra point is only 0.0007. Convergence is reached for the three parameters. The convergence is also faster: the parameter α_T starts to converge before the other two parameters have reached stable values. The RMS error in Table 5.4 is calculated at the extraction well only. If in this experiment $\sigma = 0.01$ for the BTC at the extraction well, similar results are obtained.

The better performance of the second method clearly lies in the presence of values in a special, additional point. From the figures illustrating the direct problem solutions, it is clear that the intensity with which the tracer can diffuse beyond the extraction well is highly dependent on the value of α_T. Therefore, taking into accountn additional measurements at the extra point (−5.6m, 0m), situated just beyond the extraction well, contributes essentially to the sensitivity of the cost functional on α_T.

We have also performed similar experiments in the setting of constant injection concentration, but the sensitivity on the parameter α_T did not noticeably increase. Experiments in which different weights are given to different portions of the BTC in the cost functional, do not improve the convergence. Note further that the bigger the α_T value, the faster this value will converge if an extra point is taken into consideration.

p	FD	CD	AM
(0.2, 0.6)	(0.095,-0.0065)	(0.0938,-0.00644)	(0.0874, -0.00678)
(0.094, 0.608)	(0.00741, -0.000568)	(0.00027 , -0.000603)	(0.000116, -0.000529)
(0.1016, 0.729)	Stop. Cost = 0.000091		
(0.1, 0.3)	(0.0484, -0.0088)	(0.0382 , -0.00919)	(0.0349, -0.00899)
(0.0794, 0.305)	(0.0107, -0.00497)	(0.00152 , -0.00531)	(-0.00275, -0.00514)
(0.0742, 0.474)	(-0.0151, -0.00063)	(-0.023, -0.00074)	(-0.024, -0.00066)
(0.0926, 0.516)	(0.0103, -0.00133)	(0.0031 , -0.00136)	(0.0023, -0.00121)
(0.1018, 0.769)	Stop. Cost = 0.000038		

Table 5.5: Gradient by 3 different techniques

5.4.2 With adsorption

In this section we use an experiment where the BTC is the result of the direct model with the following parameters: $d = 10$m, $r_1 = r_2 = 0.15$m, $H = h_1 = 10$m, $h_2 = 15$m, $\theta_0 = 0.2$, $k = 0.864$, $\alpha_L = 0.02$. Moreover, we consider Freundlich adsorption with $\Psi(s) = K_0 s^q = 0.1 s^{0.8}$. There are 100 measurement points at the extration well during the time interval $[0, T] = 18$ days. Operator splitting is done every 0.1 days. At the inflow boundary there is again an injection with $C_0(t) = 1$ for $t \in (0, 1)$ and $C_0(t) = 0$ afterwards. The parameters that we will try to recover inversely are α_L, K_0 and q.

Adjoint method

To illustrate the adjoint equation method (AM) for determining $\nabla\mathcal{F}$ from (5.12), we compare this gradient with gradients that are calculted numerically with the forward (FD), $\nabla_{p_i} = \frac{\mathcal{F}(p_i+\delta p_i)-\mathcal{F}(p_i)}{\delta p_i}$, and center difference (CD), $\nabla_{p_i} = \frac{\mathcal{F}(p_i+\delta p_i)-\mathcal{F}(p_i-\delta p_i)}{2\delta p_i}$. Take $\delta p = 0.01$, so that the FD is of order 0.01 and CD of order 0.0001. FD and CD require repectively one and two extra solutions of the direct problem per parameter in order to obtain the gradient. In Table 5.5 we present $\nabla\mathcal{F}$ for different adsorption parameters $p = (K_0, q)$. The first and fourth line are initial values, the other lines give the minima as found by line search based on the conjugate gradient method using as gradient the value obtained by AM; the values in the table are the gradients for these parameter obtained with the 3 different methods. We stopped the iterations when the cost $\mathcal{F} < 0.0001$.

The FD doesn't provide good results as one parameter is retrieved within the given accuracy, after which the error on the gradient is such that the sequence of values for the second parameter can no longer converge. CD is comparable with AM (slightly less good). The AM requires solving a linear PDE; it is obtained

it	p	cost
0	(0.2, 0.6)	0.035
1	(0.134 0.528)	0.0080
2	(0.105 0.608)	0.00075
3	(0.099 0.743)	0.000023

it	p	cost
0	(0.1, 0.3)	0.0065042
1	(0.0981 0.366)	0.0035
2	(0.0961 0.448)	0.0016
3	(0.0952 0.547)	0.00059
4	(0.0961 0.668)	0.00012

Table 5.6: Levenberg-Marquardt iterations starting from 2 parameter sets

in a fraction of the time needed to solve one single direct problem.

Levenberg-Marquardt method

In a second experiment we compared the conjugate gradient method, based on the adjoint gradient method, with the Levenberg-Marquardt method,

The main advantage of the Levenberg-Marquardt method is its robustness. However, the numerical determination of the Jacobian can be computationally costly, especially when the solution of the direct problem is elaborate, which is the case in the present example. For every parameter p_i, an extra direct run with parameter value $p_i + \Delta p_i$ is needed. Therefore, a direct run and an inverse run will need $(n+1)$ times the time needed for solving (5.3), where n is the number of parameters.

Again, we only allow the adsorption parameters to change. Starting the Levenberg-Marquardt method (LM) from $p = (0.2, 0.6)$ and $p = (0.1, 0.3)$, the initial values of Table 5.5, the iterations given in Table 5.6 are obtained. It can be seen that LM is a little less efficient in the number of gradients that have to be determined compared to the AM. Therefore, LM is an adequate method if a small number of parameters is considered. In inverse problems with many parameters the AM is preferable (inverse algorithms based on the gradient that are better than CG can be used). Nevertheless, it may not be forgotten that the line search needed to retrieve the minimum, implies several extra executions of the direct problem when a method that uses line search, like CG, is used.

Part III

On a practical groundwater flow problem

Chapter 6

A practical groundwater flow problem

6.1 Physical background: Tòth's regional flow problem

Complex real world groundwater flow problems are often idealized so to admit analytical solutions, which though "elementary" give a good insight in basic groundwater hydraulics. Tòth's regional flow problem is an example of this, and is referenced in many textbooks, e.g. [9], and [24]. Our goal is to re-approach this problem so as to get a semi-analytical solution in a more realistic domain, Fig. 6.2.

We consider the problem of groundwater flow in a small drainage basin, as first presented by Tòth in [72] and theoretically analysed in [73]. The basin has vertical impenetrable boundaries corresponding to symmetry considerations. The longitudinal component can be neglected as in most small basins the slopes of the valley flanks greatly exceed the longitudinal slopes of the valley floors. Therefore, a two-dimensional model can be adopted in (x, z) coordinates, with x the horizontal coordinate and z the elevation. The basin furthermore has the special property that the water table follows the surface. This is possible when the aquifer has a low conductivity and there is abundant rainfall. First, the domain is assumed to be limited by a horizontal impermeable boundary at the base. Next, we will consider a semi-infinite domain, having no base.

This last assumption will approximate the situation of very deep basins. In the absence of sources, the stationary hydraulic head, $h(x,z)$, satisfies the steady state equation

$$\nabla \cdot (K(z)\nabla h(x,z)) = 0, \text{ in } \Omega$$

where $K(z)$ is the hydraulic conductivity, see (B.5).

There exist several simplified approaches for the problem. Tòth, [72] and [73], has studied the boundary value problem for Laplace's equation ($K(z)$ constant) in a finite vertical, two dimensional, saturated, homogeneous, isotropic region bounded on top by a sloping sinusoidal curve, which represents the watertable. However, he approximates the problem by reducing the domain to a rectangle with the given top boundary values projected onto the top of this rectangle. This assumes that the solution has the same value on the top of the rectangle as it did on the given boundary. In [68] the top boundary condition is taken into account exactly, but a vertically infinite region is considered, the z-coordinate varying from 0 to $-\infty$. The hydraulic conductivity is assumed to decrease exponentially with depth, i.e., $K(z) = ce^{dz}$, which is supported by some experimental data. The assumption of the semi-infinite region, of course, only allows reliable results for deep drainage basins but not for the usual basins of depth from 600 up to 1000 feet, as studied by Tòth. More generally, numerical methods such as finite difference or finite element methods can be used to solve such problems, but these methods require more CPU-time with increasing depth of the basins, and don't provide as much qualitative information as a formal solution.

In this chapter, we first search the hydraulic head h in a non-homogeneous porous medium, in a region bounded between two vertical impermeable boundaries, bounded on top by a sloping sinusoidal curve and by a horizontal impermeable boundary at the base. We extend the result by allowing a Dirichlet boundary condition at the base. The latter is conform with an underlying confined aquifer of high conductivity that interacts with the considered low conductive region.

To deal with the present problem, we first extend the semi-analytical method of [68] and reduce the problem to solving an infinite system of linear equations. In case of a Dirichlet boundary condition at the base we use a Fourier series decomposition of the boundary condition. This involves a Gramm matrix which is positive definite. This system is truncated so as to yield an approximate solution that provides the best match with the given boundary data at the top surface. To test the validity of the method, a simple Galerkin finite element method was implemented. Moreover, an *infinite element method* was implemented to reduce

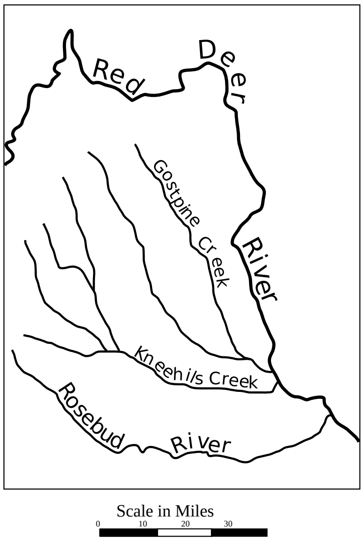


Figure 6.1: Area in Central Alberta with parallelism of creeks.

the computation in case of deep drainage basins.

An outline of this chapter is as follows. In Section 6.2 we present the original analytical solution of Tóth. In Section 6.3, we state the mathematical model for the hydraulic potential. In Section 6.4, we derive a formal solution to the boundary value problem by a suitable Fourier expansion method and infinite linear system techniques. In Section 6.5, we discuss the semi-analytical approach and obtain some numerical results that are compared with results from the literature. In Section 6.6 we briefly sketch a finite element approach and we also deal with an infinite element method.

For continuity with the sources and citations, American units, i.e., miles and feet, will be used almost everywhere in this chapter. The results presented here appeared in [57, 55]; an outline of the numerical algorithms was presented in [54].

6.2 Tòth's regional flow model

In [72], Tòth started the study of small drainage basins in Central Alberta, Canada, Fig. 6.1. The surface of the area generally slopes downward to the east. The surface is very gently rolling and is subdivided by a few main creeks into nearly parallel watersheds and valleys. The distances between adjacent water divides are 6 to 10 miles. The creeks all have tributary coulees that are dry except during periods of surface runoff.

He observed 3 features: a close correlation of the piezometric surface with the topography in general; relatively high or low natural levels in certain wells as compared with the general piezometric surfaces at wells of similar depth; the different character of the change in head with changing well depth, if the wells are grouped according to recharge and discharge areas. Examining the effects of the topography and geology, he suggested the following conditions:

1. No confined or unconfined flow system of large areal extent can be formed. It can be stated that any single watershed seems to constitue a unit system in the groundwater flow, allowing to speak of single local regions that can be studied independantly.

2. Vertical impermeable boundaries can be assumed to exist for all practical purposes at water divides and streams. This is because the topography is approximately symmetrical relative to either a water divide or a valley bottom. Further, recharge is due to infiltrating rain and melt water, wich is equally distributed on both sides of a divide or stream.

3. An abrupt decrease in permeability can be considered to exist at the bottom of the top layer. This top layer is the Paskapoo formation (mostly soft gray, clayey sandstone) of thickness from 0 to 600 or 1000 feet going from east to west across the area. Below this is the Edmonton formation (sandstones and siltstones cemented with bentonitic clay), which is marked by a drop in permeability. Further evidence for this aspect is contact springs that exist where contact zones outcrop. We can therefore treat this boundary as a horizontal impermeable boundary.

In the theoretical approach [73] an analytic solution was given. The surface of the water table was taken to have a fixed slope, superimposed with a sinus function. The region of the groundwater flow was represented however by a rectangular area, Fig. 6.2. This rectangle was made of a horizontal impermeable boundary at its base, by two impermeable boundaries extending downward from

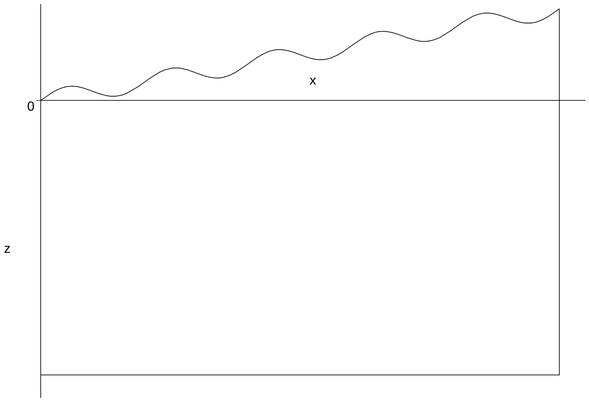


Figure 6.2: The domain.

the stream and the water divide, and by a horizontal line at the elevation of the stream along which the head is supposed to be the same as that for the real water table.

The mathematical model for the steady state case of the hydraulic head h is the following. The domain Ω is $0 < x < L$, $-T < z < 0$. The water table is given by

$$g(x) = \frac{ax}{L} + V \sin\left(\frac{2\pi n x}{L}\right),$$

where $L > 0, T > 0, a \geq 0$, and V are constants and n is a fixed positive integer.

From (B.5) with constant hydraulic conductivity, we obtain

$$\Delta h(x, z) = 0, \tag{6.1}$$

where Δ is the Laplacian. The four boundary conditions are

$$\begin{aligned} \frac{\partial h}{\partial x}|_{x=0} = \frac{\partial h}{\partial x}|_{x=L} &= 0, \\ \frac{\partial h}{\partial z}|_{z=-T} &= 0, \\ h(x, z=0) &= T + g(x), \end{aligned}$$

where the impermeable base is the stratum for the head. The general solution of the Laplace equation can be written as

$$h(x,z) = e^{-kz}(A\cos kx + B\sin kx) + e^{kz}(M\cos kx + N\sin kx),$$

where the arbitrary constants A, B, M, N are determined by the boundary conditions. The final expression reads, [73],

$$\begin{aligned} h(x,z) &= T + \frac{a}{2} + \frac{V}{2\pi n}(1-\cos 2\pi n) + \\ &+2\sum_{m=1}^{\infty}\left[\frac{2\pi n V L\,(1-\cos 2\pi n\cos m\pi)}{(2\pi n)^2-(m\pi)^2} + \frac{aL}{m^2\pi^2}(\cos m\pi - 1)\right]\cdot \\ &\frac{\cos(m\pi x/L)\cosh(m\pi z/L)}{L\cosh(m\pi T/L)}, \end{aligned} \tag{6.2}$$

and satisfies the boundary conditions and the Laplace equation. Tòth gives several examples for different values of the parameters; these can be compared with the (general) examples later in this chapter.

A well known result was obtained for deep basins, see the figures in [73]. Three distinctly different types of flow systems can occupy a basin: local, intermediate and regional systems. A *local system* of groundwater flow has its recharge area at a topographic high and its discharge area at an adjacent topographic low area. An *intermediate system* has its recharge and discharge areas a few topographic highs and lows further away. Finally, the *regional system* has its recharge area at the water divide and its discharge area at the bottom of the valley. This system is present in all deep basins.

In the next section, we present our more general mathematical model. It is an extension of [68]. We will highlight the differences of our results with those of [68] throughout.

6.3 General mathematical model

As in [73] we consider the following governing differential equation for the hydraulic head $h(x,z)$ in the stationary regime in absence of sources

$$\nabla\cdot(e^{dz}\nabla h(x,z)) = 0, \text{in } \Omega. \tag{6.3}$$

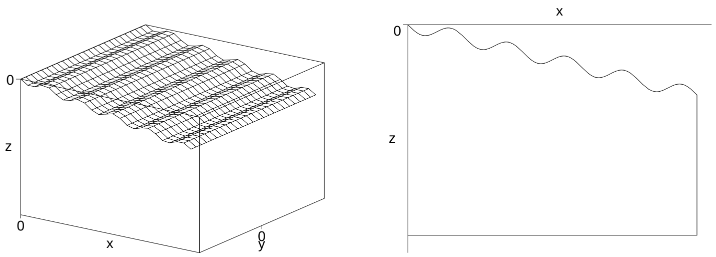


Figure 6.3: The domain

The hydraulic conductivity is $K = ce^{dz}$ (c a positive constant, $d \geq 0$). The domain Ω under consideration is given by (see Fig. 6.3)

$$0 < x < L \quad \text{and} \quad -T < z < g(x) \equiv -\left[\frac{ax}{L} + V\sin\left(\frac{2\pi nx}{L}\right)\right], \qquad (6.4)$$

where $L > 0$, $T > 0$, $a \geq 0$, and V are constants and n is a fixed positive integer.

The boundary conditions are given by

$$\frac{\partial h}{\partial x}|_{x=0} = \frac{\partial h}{\partial x}|_{x=L} = 0, \qquad (6.5)$$

$$h(x,z) = z \quad on \quad z = g(x), \qquad (6.6)$$

and

$$\frac{\partial h}{\partial z}|_{z=-T} = 0, \qquad (6.7)$$

or

$$h(x,z)|_{z=-T} = f(x). \qquad (6.8)$$

Here, as in [68], g is defined by (6.4). Moreover, f is a piecewise smooth function on $[0, L]$. We recall that in [68], the depth T of the soil layer is taken to be infinity, the boundary conditions (6.7-6.8) being replaced by the condition that h is bounded for $z \to -\infty$.

6.4 Analytical solution

By separation of variables a formal analytic solution of the diffusion equation (6.3), i.e.

$$\frac{\partial^2 h}{\partial x^2} + \frac{\partial^2 h}{\partial z^2} + d\frac{\partial h}{\partial z} = 0 \quad \text{in } \Omega, \tag{6.9}$$

under the mentioned boundary conditions is found. We set $h(x,z) = X(x)Z(z)$. We obtain 2 separate second order ODE which can readily be solved taking into account the BCs. With BC(6.5) we derive

$$h(x,z) = \left\{ \begin{array}{l} \eta_{1,0} + \eta_{2,0} z \\ \eta_{1,0} + \eta_{2,0} e^{-dz} \end{array} \right\} + \sum_{m=1}^{\infty} \cos(\frac{m\pi x}{L}) \left(\eta_{1,m} e^{\lambda_m^+ z} + \eta_{2,m} e^{\lambda_m^- z} \right) . \tag{6.10}$$

Here, $\lambda_m^\pm$ is given by

$$\lambda_m^\pm = \frac{1}{2}\left[-d \pm \sqrt{d^2 + 4\frac{m^2\pi^2}{L^2}} \right], \quad m \in \mathbb{N}. \tag{6.11}$$

Moreover, $\eta_{i,m}$ ($i = 1$ and 2; $m = 0, 1, \ldots$) are arbitrary constants. In (6.10) the top line corresponds with the case $d = 0$, the bottom line with the case $d \neq 0$. We'll keep this notation throughout.

The coefficients $\eta_{i,m}$, ($i = 1, 2; m = 0, 1, 2 \ldots$), are determined in such a way that $h(x,z)$ satisfies the remaining BCs (6.6) and (6.7) or (6.8).

Case of Neumann BC (6.7)

The function $h(x,z)$, given by (6.10), coping with (6.7), has the form

$$h(x,z) = \eta_{1,0} + \sum_{m=1}^{\infty} \eta_{1,m} \cos(\frac{m\pi x}{L}) \left[e^{\lambda_m^+ z} - \frac{\lambda_m^+}{\lambda_m^-} e^{(\lambda_m^- - \lambda_m^+)T} e^{\lambda_m^- z} \right]. \tag{6.12}$$

Case of Dirichlet BC (6.8)

Imposing (6.8) on the function (6.10) requires that

$$f(x) = \left\{ \begin{array}{l} \eta_{1,0} - \eta_{2,0} T \\ \eta_{1,0} + \eta_{2,0} e^{dT} \end{array} \right\} + \sum_{m=1}^{\infty} \cos(\frac{m\pi x}{L}) \left(\eta_{1,m} e^{-\lambda_m^+ T} + \eta_{2,m} e^{-\lambda_m^- T} \right) , \tag{6.13}$$

$0 < x < L$. On account of this cosine Fourier series expansion of f, we may obtain the coefficients $\eta_{2,m}$ in terms of $\eta_{1,m}$, $m \in \mathbb{N}$. Consequently, the expression (6.10) reduces to

$$\begin{aligned} h(x,z) \quad = \quad & \frac{D_0}{2} + \sum_{m=1}^{\infty} D_m \cos(\frac{m\pi x}{L}) e^{\lambda_m^-(z+T)} + \left\{ \begin{array}{l} \eta_{2,0}T(1+\frac{z}{T}) \\ -\eta_{2,0}e^{dT}(1-e^{-d(z+T)}) \end{array} \right\} \\ & + \sum_{m=1}^{\infty} \eta_{1,m} \cos(\frac{m\pi x}{L}) \left[e^{\lambda_m^+ z} - e^{(\lambda_m^- - \lambda_m^+)T} e^{\lambda_m^- z} \right] \end{aligned} \tag{6.14}$$

where

$$D_m = \frac{2}{L} \int_0^L f(t) \cos \frac{m\pi t}{L}\, dt,\ m \in \mathbb{N}. \tag{6.15}$$

The expressions (6.12) and (6.14) can be combined into a single one, viz

$$\begin{aligned} h(x,z) \quad = \quad & \frac{D_0}{2} + \sum_{m=1}^{\infty} D_m \cos(\frac{m\pi x}{L}) e^{\lambda_m^-(z+T)} + \left\{ \begin{array}{l} \beta_0(1+a_0\frac{z}{T}) \\ \beta_0(1-a_0e^{-d(z+T)}) \end{array} \right\} \\ & + \sum_{m=1}^{\infty} \beta_m \cos(\frac{m\pi x}{L}) \left[e^{\lambda_m^+ z} - a_m e^{(\lambda_m^- - \lambda_m^+)T} e^{\lambda_m^- z} \right]. \end{aligned} \tag{6.16}$$

Here, β_m, $(m \in \mathbb{N})$, are arbitrary constants. In the case of the BC (6.7) one has

$$a_0 = 0, a_m = \frac{\lambda_m^+}{\lambda_m^-},\ (m \in \mathbb{N}_0),\ D_m = 0,\ (m \in \mathbb{N}).$$

In the case of the BC (6.8) one takes

$$a_m = 1,\ (m \in \mathbb{N}), \text{ and } D_m \text{ given by (6.15).}$$

The coefficients β_m, $(m \in \mathbb{N})$, must be determined by imposing the remaining

BC (6.6), i.e.

$$\begin{aligned}
&-\left[\frac{ax}{L}+V\sin(\frac{2\pi nx}{L})\right]-\frac{D_0}{2}-\sum_{m=1}^{\infty}D_m\cos(\frac{m\pi x}{L})e^{\lambda_m^-(T-\left[\frac{ax}{L}+V\sin(\frac{2\pi nx}{L})\right])}\\
&=\left\{\begin{array}{l}\beta_0(1+a_0\frac{-\left[\frac{ax}{L}+V\sin(\frac{2\pi nx}{L})\right]}{T})\\ \beta_0(1-a_0e^{-d(T-\left[\frac{ax}{L}+V\sin(\frac{2\pi nx}{L})\right])})\end{array}\right\}\\
&\quad+\sum_{m=1}^{\infty}\beta_m\cos(\frac{m\pi x}{L})\left\{\exp\left[-\lambda_m^+\left\{\frac{ax}{L}+V\sin(\frac{2\pi nx}{L})\right\}\right]\right.\\
&\quad\left.-a_me^{(\lambda_m^--\lambda_m^+)T}\exp\left[-\lambda_m^-\left\{\frac{ax}{L}+V\sin(\frac{2\pi nx}{L})\right\}\right]\right\},\ 0<x<L.
\end{aligned}\tag{6.17}$$

This condition is made non-dimensional by setting

$$y=x/L,\ \alpha_m=\beta_m/L,\ \sigma_m^{\pm}=\lambda_m^{\pm}L,\ \tilde{D}_m=D_m/L,\quad(m\in\mathbb{N})$$

and

$$\tilde{a}=a/L,\ \tilde{V}=V/L,\ \tilde{T}=T/L.$$

Introducing $K(y)=\tilde{a}y+\tilde{V}\sin(2\pi ny)$ for $0<y<1$, we get from (6.17) that

$$\begin{aligned}
&-K(y)-\frac{\tilde{D}_0}{2}-\sum_{m=1}^{\infty}\tilde{D}_m\cos(m\pi y)e^{\sigma_m^-\tilde{T}}e^{-\sigma_m^-K(y)}\\
&=\left\{\begin{array}{l}\alpha_0(1+a_0\frac{-K(y)}{\tilde{T}})\\ \alpha_0(1-a_0e^{-d\tilde{T}L}e^{dLK(y)})\end{array}\right\}\\
&\quad+\sum_{m=1}^{\infty}\alpha_m\cos(m\pi y)\left[e^{-\sigma_m^+K(y)}-a_me^{(\sigma_m^--\sigma_m^+)\tilde{T}}e^{-\sigma_m^-K(y)}\right]
\end{aligned}\tag{6.18}$$

Define

$$u_k(y)\ =\ \left\{\begin{array}{ll}1+a_0\frac{-K(y)}{\tilde{T}}, & d=0,\ k=0\\ \cos(k\pi y)\left[e^{-\sigma_k^+K(y)}-a_ke^{(\sigma_k^--\sigma_k^+)\tilde{T}}e^{-\sigma_k^-K(y)}\right], & \text{otherwise.}\end{array}\right.$$

Then, condition (6.18) can be rewritten in the form

$$-K(y)-\frac{\tilde{D}_0}{2}-\sum_{m=1}^{\infty}\tilde{D}_m\cos(m\pi y)e^{\sigma_m^-\tilde{T}}e^{-\sigma_m^-K(y)}=\sum_{m=0}^{\infty}\alpha_mu_m(y),\tag{6.19}$$

where $0 < y < 1$, from which the remaining coefficients α_m, $(m \in \mathbb{N})$, must be determined. Multiplying both sides of (6.19) by $u_k(y)$ and integrating with respect to y over $(0,1)$, we arrive at the formal infinite system of equations

$$\sum_{m=0}^{\infty} b_{km}\alpha_m = c_k,\ k \in \mathbb{N}, \tag{6.20}$$

with

$$b_{km} = \int_0^1 u_k(y)u_m(y)\,dy. \tag{6.21}$$

$$\begin{aligned} c_k = & -\int_0^1 u_k(y)K(y)\,dy - \frac{\tilde{D}_0}{2}\int_0^1 u_k(y)\,dy \\ & -\tilde{D}_m \sum_{m=1}^{\infty}\int_0^1 u_k(y)\cos(m\pi y)e^{\sigma^-_m \tilde{T}}e^{-\sigma^-_m K(y)}\,dy \end{aligned} \tag{6.22}$$

The infinite matrix $\mathbf{B}= [b_{km}]_{k,m=0,1,2,\dots}$ is the Gramm matrix of the set $\{u_k : k \in \mathbb{N}\}$.

The integrals (6.21) and (6.22) can be evaluated *analytically.* To this end we recall the definition of the modified Bessel functions of the 1st kind

$$I_m(x) = (\tfrac{x}{2})^2 \textstyle\sum_{k=0}^{\infty}(\tfrac{x}{2})^{2k}\frac{1}{(k!(k+m)!)},$$

and we invoke the identity (see [22])

$$e^{-p\sin\theta} = I_0(p) + 2\sum_{q-1}^{\infty} I_{2q}(p)\cos(2q\theta) + 2\sum_{q=1}^{\infty} I_{2q-1}(p)\sin((2q-1)\theta), \tag{6.23}$$

$\forall\theta$, in order to deal with $e^{-\sigma^{\pm}_k K(y)}$.

We first evaluate 2 auxiliary integrals, which are then used to express b_{km} and c_k. We make some notational assumptions for brevity. In both expressions the ± symbol denotes the sum of two terms, one for each sign, or the sum of four terms if there are a pair of plus and minus signs (see [68]). For example, $\frac{a\pm b}{(a\pm b)^2+c} = \frac{a+b}{(a+b)^2+c} + \frac{a-b}{(a-b)^2+c}$, and $\frac{\pm a \pm b}{(\pm a \pm b + c)^2} = \frac{a+b}{(a+b+c)^2} + \frac{a-b}{(a-b+c)^2} + \frac{-a+b}{(-a+b+c)^2} + \frac{-a-b}{(-a-b+c)^2}$.

On account of (6.23) we obtain

$$\begin{aligned}
T(m,k,A,B) &\equiv \int_0^1 \cos(m\pi y)\cos(k\pi y) e^{-(A+B)K(y)}\,dy \\
&= \frac{1}{2} I_0(\tilde{V}(A+B)) \frac{\tilde{a}(A+B)(1-e^{-\tilde{a}(A+B)}(-1)^{k+m})}{(m\pm k)^2\pi^2 + \tilde{a}^2(A+B)^2} \qquad (6.24) \\
&+\frac{1}{2}\sum_{q=1}^{\infty}(-1)^q I_{2q}(\tilde{V}(A+B)) \frac{\tilde{a}(A+B)(1-e^{-\tilde{a}(A+B)}(-1)^{k+m})}{\left[(m\pm k)\pi \pm 4qn\pi\right]^2 + \tilde{a}^2(A+B)^2} \\
&+\frac{1}{2}\sum_{q=1}^{\infty}(-1)^q I_{2q-1}(\tilde{V}(A+B)) \times \\
&\quad \frac{\left[(2q-1)2n\pi \pm (m\pm k)\pi\right](1-e^{-\tilde{a}(A+B)}(-1)^{k+m})}{\left[(2q-1)2n\pi \pm (m\pm k)\pi\right]^2 + \tilde{a}^2(A+B)^2},\ A+B \neq 0.
\end{aligned}$$

Similarly, (6.23) leads to

$$\begin{aligned}
S(k,A) &\equiv -\int_0^1 K(y) cos(k\pi y) e^{-AK(y)}\,dy \qquad (6.25) \\
&= -\tilde{a} I_0(\tilde{V}A)\left\{-\frac{A\tilde{a}\,e^{-\tilde{a}A}(-1)^k}{(\tilde{a}A)^2+(k\pi)^2} + \frac{\left[(\tilde{a}A)^2-(k\pi)^2\right]\left(1-(-1)^k e^{-\tilde{a}A}\right)}{\left[(\tilde{a}A)^2+(k\pi)^2\right]^2}\right\} \\
&-\frac{\tilde{V}}{2} I_0(\tilde{V}A)\frac{(2n\pi\pm k\pi)\left(1-(-1)^k e^{-\tilde{a}A}\right)}{(2n\pi\pm k\pi)^2+(\tilde{a}A)^2} \\
&-\tilde{a}\sum_{q=1}^{\infty}(-1)^q I_{2q}(\tilde{V}A)\left\{-\frac{A\tilde{a}\,e^{-\tilde{a}A}(-1)^k}{(\tilde{a}A)^2+(k\pi\pm 4qn\pi)^2}\right. \\
&\left.+\frac{\left[(\tilde{a}A)^2-(k\pi\pm 4qn\pi)^2\right]\left(1-(-1)^k e^{-\tilde{a}A}\right)}{\left[(\tilde{a}A)^2+(k\pi\pm 4qn\pi)^2\right]^2}\right\} \\
&-\frac{\tilde{V}}{2}\sum_{q=1}^{\infty}(-1)^q I_{2q}(\tilde{V}A)\frac{\left[\pm k\pi\pm 4qn\pi+2n\pi\right]\left(1-(-1)^k e^{-\tilde{a}A}\right)}{\left[\pm k\pi\pm 4qn\pi+2n\pi\right]^2+(\tilde{a}A)^2} \\
&-\tilde{a}\sum_{q=1}^{\infty}(-1)^q I_{2q-1}(\tilde{V}A)\left\{-\frac{\left[(2q-1)2n\pi\pm k\pi\right]e^{-\tilde{a}A}(-1)^k}{(\tilde{a}A)^2+\left[(2q-1)2n\pi\pm k\pi\right]^2}\right. \\
&\left.+\frac{2\tilde{a}A\left[(2q-1)2n\pi\pm k\pi\right]\left(1-(-1)^k e^{-\tilde{a}A}\right)}{\left[(\tilde{a}A)^2+((2q-1)2n\pi\pm k\pi)^2\right]^2}\right\}
\end{aligned}$$

$$
\begin{aligned}
&-\frac{\tilde V}{2}\sum_{q=1}^{\infty}(-1)^q I_{2q-1}(\tilde V A)\left\{\frac{\tilde a A\left(1-(-1)^k e^{-\tilde a A}\right)}{\left[(2q-2)2n\pi \pm k\pi\right]^2+(\tilde a A)^2}\right.\\
&\left.-\frac{\tilde a A\left(1-(-1)^k e^{-\tilde a A}\right)}{\left[4qn\pi \pm k\pi\right]^2+(\tilde a A)^2}\right\},\ A\neq 0.
\end{aligned}
$$

Now, the entries of the Gramm matrix entering the infinite linear system (6.20) can be written as:

$$
\begin{aligned}
b_{00} &= 1-2a_0 e^{\sigma_0^-\tilde T}T(0,0,0,\sigma_0^-)+a_0 e^{2\sigma_0^-\tilde T}T(0,0,0,2\sigma_0^-),\ \text{when } d\neq 0,\\
b_{00} &= 1-a_0\left(\tfrac{\tilde a}{\tilde T}+\tfrac{\tilde a^2}{3\tilde T^2}+\tfrac{\tilde V^2}{2\tilde T^2}-\tfrac{\tilde V\tilde a}{\pi\tilde T^2 n}\right),\ \text{when } d=0,\\
b_{0k} &= T(0,k,0,\sigma_k^+)-a_k e^{(\sigma_k^- -\sigma_k^+)\tilde T}T(0,k,0,\sigma_k^-)+\tfrac{a_0}{\tilde T}S(k,\sigma_k^+)\\
&\quad -\tfrac{a_0}{\tilde T}e^{(\sigma_k^- -\sigma_k^+)\tilde T}S(k,\sigma_k^-),\ \text{when } d=0,\ k\neq 0,\\
b_{kk} &= -a_k e^{(\sigma_k^- -\sigma_k^+)\tilde T}+T(k,k,\sigma_k^+,\sigma_k^+)+e^{-4\sigma_k^+\tilde T}T(k,k,\sigma_k^-,\sigma_k^-),\\
&\qquad \text{when } d=0,\ k\neq 0,\\
b_{km} &= T(k,m,\sigma_m^+,\sigma_k^+)-a_m e^{(\sigma_m^- -\sigma_m^+)\tilde T}T(k,m,\sigma_m^-,\sigma_k^+)\\
&\quad -a_k e^{(\sigma_k^- -\sigma_k^+)\tilde T}T(k,m,\sigma_m^+,\sigma_k^-)\\
&\quad +a_k a_m e^{(\sigma_k^- +\sigma_m^- -\sigma_k^+ -\sigma_m^+)\tilde T}T(k,m,\sigma_m^-,\sigma_k^-),\\
&\qquad \text{when } d=0,\ (k-m)k\neq 0,\ \text{and when } d\neq 0,\ k+m\neq 0.
\end{aligned}
$$

Similarly, the right hand side of (6.20) reads

$$
\begin{aligned}
c_0 &= -\frac{1}{2}\tilde a-2a_0 e^{\sigma_0^-\tilde T}S(0,\sigma_0^-)-\tilde D_0\left(1-e^{\sigma_0^-\tilde T}T(0,0,0,\sigma_0^-)\right)\\
&\quad -\sum_{m=0}^{\infty}\tilde D_m e^{\sigma_m^-\tilde T}\left(T(m,0,\sigma_m^-,0)-e^{\sigma_0^-\tilde T}T(m,0,\sigma_m^-,\sigma_0^-)\right),\ d\neq 0,\\
c_0 &= -\tfrac{1}{2}\tilde a+\tfrac{a_0}{\tilde T}\left(\tfrac{\tilde a^2}{3}+\tfrac{\tilde V^2}{2}-\tfrac{\tilde V\tilde a}{\pi n}\right)-\tilde D_0\left(1-\tfrac{\tilde a}{2\tilde T}\right)\\
&\quad -\sum_{m=0}^{\infty}\tilde D_m e^{\sigma_m^-\tilde T}\left(T(0,m,\sigma_m^-,0)+\frac{1}{\tilde T}S(m,\sigma_m^-)\right),\ \text{when } d=0,
\end{aligned}
$$

$$\begin{aligned} c_k \quad = \quad & S(k,\sigma_k^+) - a_k e^{(\sigma_k^- - \sigma_k^+)\tilde{T}} S(k,\sigma_k^-) \\ & -\tilde{D}_0 \left(T(0,k,0,\sigma_k^+) - e^{(\sigma_k^- - \sigma_k^+)\tilde{T}} T(0,k,0,\sigma_k^-) \right) \\ & - \sum_{m=0}^{\infty} \tilde{D}_m e^{\sigma_m^- \tilde{T}} \left(T(m,k,\sigma_m^-,\sigma_k^+) - e^{(\sigma_k^- - \sigma_k^+)\tilde{T}} T(m,k,\sigma_m^-,\sigma_k^-) \right), \end{aligned}$$

with $k = 1, 2, \ldots$.

From (6.20), the unknowns α_m, $m \in \mathbb{N}$ are obtained.

6.5 Approximation of the formal solution and numerical experiments

The solution $h(x,z)$ to the flow problem is given by (6.16). In practice, we must approximate h by a truncated expression of the form

$$\begin{aligned} h_{N,l}(x,z) \quad = \quad & \frac{D_0}{2} + \sum_{m=1}^{l} D_m \cos(\frac{m\pi x}{L}) e^{\lambda_m^-(z+T)} \left\{ \begin{array}{l} \beta_0(1 + a_0 \frac{z}{T}) \\ \beta_0(1 - a_0 e^{-d(z+T)}) \end{array} \right\} \\ & + \sum_{m=1}^{N} \beta_m \cos(\frac{m\pi x}{L}) \left[e^{\lambda_m^+ z} - a_m e^{(\lambda_m^- - \lambda_m^+)T} e^{\lambda_m^- z} \right], \end{aligned} \qquad (6.26)$$

where the integers N and l are parameters and $-T < z < g(x)$, $0 < x < L$. Recall that in case of the homogeneous Neumann BC (6.7), $D_m = 0$ for all $m \in \mathbb{N}$. The expression (6.26) with $D_m \neq 0$ corresponds to the Dirichlet BC (6.8) approximated by

$$h(x,z)\big|_{z=-T} = \widetilde{f}_l(x) \equiv \frac{D_0}{2} + \sum_{m=1}^{l} D_m \cos(\frac{m\pi x}{L}). \qquad (6.27)$$

Notice that not only the Fourier coefficients D_m of f are decreasing functions of m, but also λ_m^-, ($\lambda_m^- < 0$), and moreover $z + T \geq 0$ in (6.26). Thus, in (6.26), l may be taken relatively small. The function $h_{N,l}$ above satisfies the BC (6.7) or (6.27), as well as the BC (6.5). The error at the top surface $z = g(x)$, $0 \leq x \leq L$, of the domain Ω, committed when using $h_{N,l}(x,z)$ instead of $h(x,z)$ reads

$$e_{N,l}(x) = h_{N,l}(x, g(x)) - g(x),\ 0 \leq x \leq L. \qquad (6.28)$$

The index l has to be suppressed in case of BC (6.7).

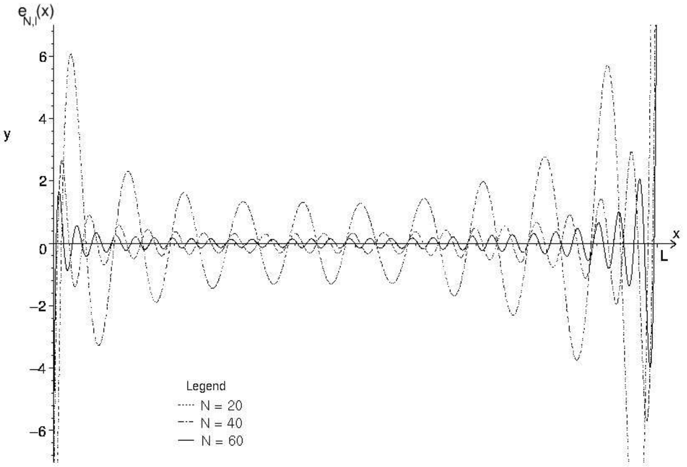


Figure 6.4: The error on the top boundary in case of a Neumann BC for different N-values, with data $a/L = 0.05$, $V = 50$, $d = 0$, $L = 20000$, $T = 2000$.

Minimizing the error $e_{N,l}(x)$ in the L_2-norm leads to the finite linear system for $\alpha_m = \frac{\beta_m}{L}$, $(m = 0, \ldots, N)$:

$$\sum_{m=0}^{N} b_{km}\alpha_m = c_k, \; k = 0, 1, \ldots, N, \tag{6.29}$$

where b_{km} and c_k, (k and $m = 0, \ldots, N$), are given by (6.21) and (6.22), respectively. This system is nothing else than the truncated version of (6.20). It is regular. Indeed, the finite Gramm matrix is symmetric and, as the functions $u_k(y)$, $(k = 1, ..., N)$, are linearly independent, also positive definite. System (6.29) provides the best matching solution in the $L_2(0, L)$-sense on the top boundary for a given N. The error $e_{N,l}(x)$ can easily be evaluated numerically. The parameter N is chosen so that $\|e_{N,l}(x)\|$ is within a required accuracy. Fig. 6.4 depicts $e_{N,l}(x)$ for a specific choice of data with BC (6.7), $(D_m = 0, \; m \in \mathbb{N})$, for different values of N.

Some numerical results of the procedure are now described. These were

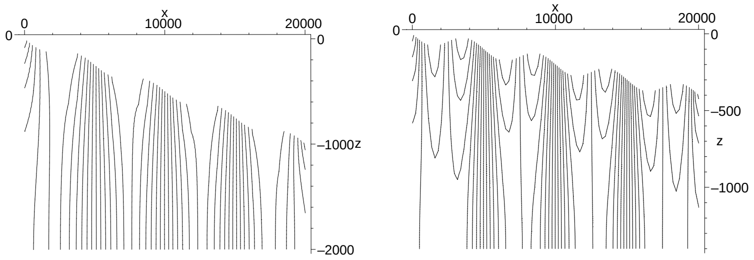


Figure 6.5: Equipotential lines in case of a Neumann BC. Data in left part: $a/L = 0.05$, $V = 50$, $d = 0$, $n = 4$, $L = 20000$, $T = 2000$. Data in right part: $a/L = 0.02$, $V = 50$, $d = 0$, $n = 4$, $L = 20000$, $T = 1400$.

obtained with a standard mathematical package, viz Maple. No programming was necessary. In Fig. 6.5 we present 2 results with the same data as in [73] (Neumann BC at the base). For $N = 60$, the equipotential lines are in full agreement with those of [73]. Moreover, they are obtained over the entire domain, in contrast with [73]. Note that the flowlines of the groundwater flow are perpendicular to these equipotential lines. In the left part of Fig. 6.5 we have regional flow, i.e. flow from the highest part towards the lowest part of the region, while in the right part there is only local flow, i.e. flow from a hill to the nearest valley.

In the case of a shallow basin with decreasing conductivity and a Neumann BC, Fig. 6.6 shows the equipotential lines for $d = 0.00235$ ($N = 100$) and $d = 0.0235$ ($N = 120$) for the same region as considered in the right part of Fig. 6.5. Note that for the second value of d the relative conductivity is reduced from 1 to 0.1 on a depth of 100 feet. A direct consequence of the conductivity decreasing with depth, is the decrease of the region where there is vertical flow, i.e. the equipotential lines are more vertical, indicating horizontal flow. Note, however, that the decreasing conductivity doesn't change the local flow character compared to the case $d = 0$. When $T >> a/2$, the numerical results for the equipotential lines are found to be in good agreement with those from [68], as it should on account of (6.18), when compared to the corresponding relation in [68].

As a last example, we consider a region with a Dirichlet boundary at the

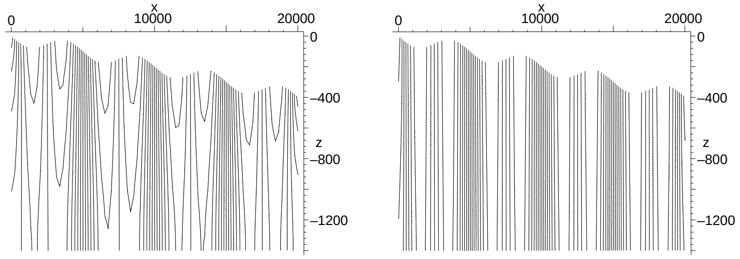


Figure 6.6: Equipotential lines in case of a Neumann BC with $a/L = 0.02$, $V = 50$, $n = 4$, $L = 20000$, $T = 1400$, $d = 0.00235$ (left) and $d = 0.0235$(right).

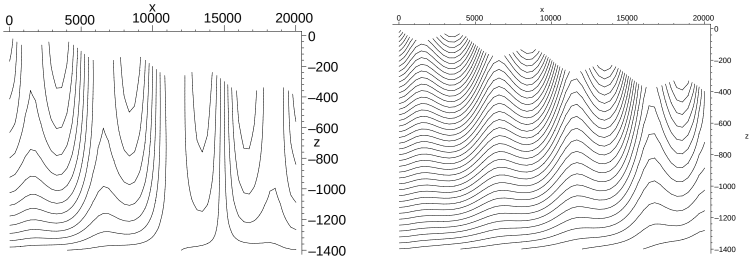


Figure 6.7: Equipotential lines in case of a Dirichlet BC at the base for $a/L = 0.02$, $V = 50$, $n = 4$, $L = 20000$, $T = 1400$. Left: $d = 0.00235$, $u = -50/L$, $v = -250$. Right: $d = 0.0$, $u = -50/L$, $v = -400$.

base. We take the function f appearing in (6.8) to be linear in x:

$$f(x) = ux + v,\ u \text{ and } v \text{ constant.} \tag{6.30}$$

This BC can be interpreted as corresponding to an underlying, highly conductive, aquifer. The function f then represents the Dupuit-Forcheimer flow (see [24]) in this aquifer. The resulting equipotential lines are depicted in Fig. 6.7 for 2 specific choices of the data. For the truncated series (6.26) we have taken $l = 21$ and N sufficiently high to reduce the error to 1 % or less of the main topographical features.

6.6 Numerical approximation methods

6.6.1 Finite element algorithm

Let $\partial\Omega = \overline{\Gamma}_1 \cup \overline{\Gamma}_2$, with $\Gamma_1 \cap \Gamma_2 = \emptyset$, be the boundary of Ω. Here, Γ_2 represents the upper surface of the physical domain. We consider the diffusion equation (6.3) under the BCs

$$\begin{aligned} \frac{\partial h}{\partial n} &= 0 \text{ on } \Gamma_1, \\ h &= g \text{ on } \Gamma_2, \end{aligned}$$

where g is a given sufficiently smooth function defined on Γ_2. For a weak formulation of this BVP, $\partial\Omega$ is only assumed to be Lipschitz continuous. Consider the function space

$$V = \{v \in H^1(\Omega) \mid v = 0 \text{ on } \Gamma_2\}. \tag{6.31}$$

Recall the weak formulation of the problem: find $h \in H^1(\Omega)$ such that

$$a(h, v) \equiv \int_\Omega e^{dz} \left[\frac{dh}{dx}\frac{dv}{dx} + \frac{dh}{dz}\frac{dv}{dz}\right] d\sigma = 0,\ \forall v \in V. \tag{6.32}$$

and

$$h = g, \text{ on } \Gamma_2. \tag{6.33}$$

Let τ_m be a regular triangulation of $\overline{\Omega}_m$, Ω_m being a polygonal domain that approximates Ω. Set $\partial\Omega_m = \overline{\Gamma}_{1m} \cup \overline{\Gamma}_{2m}$, where $\overline{\Gamma}_{1m}$ and $\overline{\Gamma}_{2m}$ are polygonal lines which piecewisely linearly interpolate Γ_1 and Γ_2, respectively. Here, m denotes the mesh parameter. The standard finite element space on $\overline{\Omega}_m$ is

$$X_m \equiv \left\{ v \in C^0(\overline{\Omega}) \mid v|_K \text{ is a polynomial of degree } 1 \; \forall K \in \tau_m \right\}.$$

The approximate (nonconforming) BVP corresponding to (6.32)-(6.33) takes the form: find $h_m \in X_m$ such that

$$\begin{aligned} a(h_m, v) &= 0, \; \forall v \in V_m \equiv \{v \in X_m \mid \; v = 0 \text{ on } \overline{\Gamma}_{2m}\}. \\ h_m &= g_m \text{ on } \overline{\Gamma}_{2m}. \end{aligned}$$

Here, g_m is the piecewise linear Lagrange-interpolant of g on $\overline{\Gamma}_{2m}$.

The results of the FEM are found to be in full agreement with those from the previous section. Actually, the equipotential lines obtained with both methods are nearly identical. An example of the obtained equipotential lines is given in the left part of Fig. 6.11 for a Neumann BC at the base, for a relatively deep region ($L = 8000$, $T = 3000$), which has been divided in 3025 triangular elements.

However, the FEM does require more CPU-time with increasing depth of the basin. Moreover, it cannot be applied to the BVP in [68] on a semi-infinite region (i.e. $T = -\infty$). In Section 6.6.2 we'll show how to cope with this difficulty by an infinite element method. The hydraulic head will be obtained for a semi-infinite region by using a small number of elements in the mesh, leading to minimal CPU-time.

Application of the FEM. The FEM is more versatile than the semi-analytic solution, in that (6.3) can be solved in more general settings like different domains. As an example, we consider areal recharge in a small basin. We set up the following experiment. Consider 2 parallel rivers (at $x = 0$ and $x = L$, x the transversal direction), with an elevation in between. The longitudinal components of the flow can be neglected if the slope of the elevation flanks greatly exceeds the longitudinal slopes of the river floors and the longitudinal slope of the elevation top. Therefore, also in this case a 2-dimensional scheme can be adopted, with x the horizontal coordinate along the elevation flanks and z the depth. We may consider the rivers to act as no-flow boundaries, simularly as in [72, 73]. As before, the groundwater level follows the surface. Below the considered domain, we have an aquifer. This aquifer will undergo an areal recharge due to the overlying basin. Based on the Dupuit-Forcheimer model, we take the bottom boundary condition (6.8) to be

$$f(x) = -\frac{N}{2}x(x-L) + \frac{h_{x=L} - h_{x=0}}{L}x + h_{x=0} \tag{6.34}$$

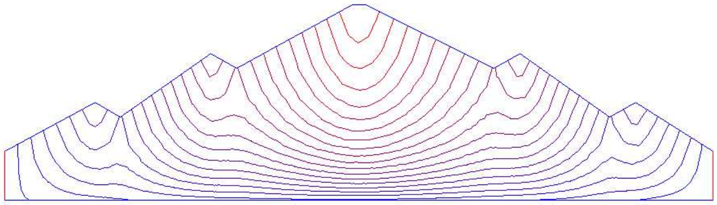


Figure 6.8: Topology elevation between two rivers. Top of elevation is at $z = 0$, $L = 5500m$, underlying aquifer at $z = -400m$. Equipotential lines for $h_{x=0} = -290m$, $h_{x=L} = -300m$, $N = 0.05/L$ and $d = 0.008$.

with $h_{x=0}$ and $h_{x=L}$ the hydraulic head of the aquifer under the rivers, and with N being the areal recharge (dimension L/T), see [24].

For the experiment, we consider the elevation as in Fig. 6.8 and we take as data for this problem $h_{x=0} = -290m$, $h_{x=L} = -300m$, $L = 5500m$, $d = 0.008$ and $N = 0.05/L$. This means that the areal recharge over the length L totalizes 0.05 m/s. The resulting equipotential lines are given in the same figure.

6.6.2 Infinite element algorithm

Comparison of the results of Section 6.5 for BC (6.7), with those from a semi-infinite region, [68], shows that both results are in close agreement when $T >> a/2$. Therefore, to reduce computational efforts, it is interesting to consider a semi-infinite approach in those cases.

An infinite element method, [10] or [78], may be developed, using a constant but unknown far field value of the hydraulic head for $z \to -\infty$, which replaces the boundedness assumption in [68].

Let Ω be the semi-infinite region shown in Fig. 6.9. The boundary is denoted by $\overline{\Gamma}_1 \cup \overline{\Gamma}_2$, with $\Gamma_1 \cap \Gamma_2 = \emptyset$. We have that Γ_1 are vertical lines extending to $-\infty$. Consider the following BCs

$$\begin{aligned} \frac{\partial h}{\partial n} &= 0 \text{ on } \Gamma_1, \\ h &= g \text{ on } \Gamma_2, \\ h \to c_{\text{ff}} \quad &\text{as} \quad z \to -\infty \text{ , } c_{\text{ff}} \text{ constant,} \end{aligned}$$

where again g is a given sufficiently smooth function defined on Γ_2, where Γ_2 is

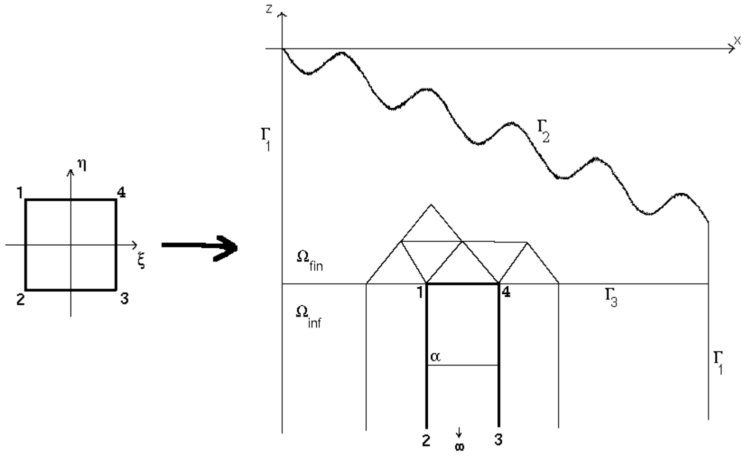


Figure 6.9: The mapping of the infinite rectangular elements.

assumed to be Lipschitz continuous. We introduce an analogue function space V as in (6.31). The weak formulation of the problem is identical to (6.32)-(6.33).

We construct a mesh for $\overline{\Omega}_m$. Here, $\overline{\Omega}_m$ is a semi-infinite polygonal domain that approximates Ω, with boundary $\overline{\Gamma}_1 \cup \overline{\Gamma}_{2m}$, where the polygonal line Γ_{2m} piecewisely linearly interpolates the curved boundary Γ_2 of Ω. The domain Ω_m is splitted into a bounded part Ω_{fin} and an unbounded part Ω_{inf} by means of the horizontal line Γ_3, $z = -T$. For Ω_{fin} we consider a regular triangulation τ_m, while on Ω_{inf} we consider a mesh ρ_m of semi-infinite rectangles matching perfectly with the elements of τ_m, see Fig 6.9. For the IEM we take globally continuous, elementwise polynomials of degree 1 on Ω_{fin}. On Ω_{inf} we use globally continuous, elementwise mapped bilinear polynomials as specified below.

Let K be a generic semi-infinite rectangular element, the corner points of which are numbered as in Fig. 6.9. The abscissa of nodes 1 and 4 are denoted by x_1^K and x_4^K, respectively. We consider a *master* square element $\hat{K}$ with corners $(\pm 1, \pm 1)$ in the $\xi\eta$-plane, as indicated. The mapping $(\xi, \eta) \rightarrow (x, z)$ given by

$$\left\{ \begin{array}{rl} x = & \frac{x_4^K - x_1^K}{2}\xi + \frac{x_4^K + x_1^K}{2} \\ z = & 2(-T-\alpha)\frac{\eta}{1+\eta} + \alpha \end{array} \right. \Leftrightarrow \left\{ \begin{array}{rl} \xi = & \frac{2}{x_4^K - x_1^K}x - \frac{x_4^K + x_1^K}{x_4^K - x_1^K} \\ \eta = & \frac{\alpha - (-2T-\alpha)}{z - (-2T-\alpha)} - 1 \end{array} \right.$$

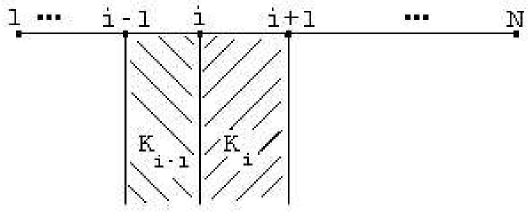


Figure 6.10: The nodes on $\overline{\Gamma}_3$, and the connected semi-infinite rectangles.

transforms $\hat{K}$ into K, with correspondence of the nodes. Here, $\alpha < -T$ is a parameter, the horizontal line $z = \alpha$ inside K corresponding to the midline $\eta = 0$ of $\hat{K}$. We denote this mapping by $\mathcal{F}_K$. On $\hat{K}$ we consider the space $Q_1(\hat{K})$ of bilinear polynomials. For each $\hat{v} \in Q_1(\hat{K})$ define

$$v(x,z) = \hat{v}(\xi,\eta), \text{ when } (x,z) = \mathcal{F}_K(\xi,\eta),\ (\xi,\eta) \in \hat{K} \tag{6.35}$$

The function v is called a ***mapped bilinear function*** on K. Define 2 function spaces on $\overline{\Omega}$

$$\begin{aligned} X_m \;&\equiv\; \big\{v \in C^0(\overline{\Omega}) \mid v|_K \text{ is a polynomial of degree } 1\ \forall K \in \tau_m \text{ and} \\ &\qquad v\,|_K \text{ is a mapped bilinear function with} \qquad (6.36)\\ &\qquad v|_{z\to-\infty} \to c\ \ \forall K \in \rho_m,\ c \text{ constant}\big\}, \\ V_m \;&\equiv\; \{v \in X_m \mid\ v = 0 \text{ on } \overline{\Gamma}_{2m}\}. \qquad (6.37) \end{aligned}$$

Consider the following discrete variational problem: find $h_m \in X_m$ such that

$$\begin{aligned} a(h_m, v) &= 0,\ \forall v \in V_m \\ h_m &= g_m \text{ on } \overline{\Gamma}_{2m}. \end{aligned}$$

Here, g_m is the piecewise linear Lagrange-interpolant of g on $\overline{\Gamma}_{2m}$.

For computational purposes we must identify a suitable basis for the approximation space X_m. To this end three types of nodes are distinguished: (a) the nodes in $\overline{\Omega}_{\text{fin}} \backslash \overline{\Gamma}_3$, corresponding to the triangulation τ_m; (b) the nodes on $\overline{\Gamma}_3$; (c) the single node *at infinity*, corresponding with $z \to -\infty$. For simplicity in notation, let the nodes on $\overline{\Gamma}_3$ be numbered from 1 to N, see Fig 6.10. Moreover, let the M nodes in $\overline{\Omega}_{\text{fin}} \backslash \overline{\Gamma}_3$ be numbered from $N+1$ to $N+M$.

To the M nodes in $\overline{\Omega}_{\text{fin}}\backslash\overline{\Gamma}_3$ we associated the functions h_i, $(i = N{+}1, \ldots, N{+}M)$ on $\overline{\Omega}$ given by

$$\begin{aligned} h_i|_{\overline{\Omega}_{\text{fin}}} &= \text{ standard cardinal basis function on } \overline{\Omega}_{\text{fin}} \text{ corresponding to} \\ &\qquad \text{the node } (x_i, z_i) \text{ in } \tau_h \\ h_i|_{\Omega_{\text{inf}}} &= 0 \end{aligned}$$

The functions h_i, $(i = 1, \ldots, N)$ associated to the N nodes on $\overline{\Gamma}_3$ are defined by

$$\begin{aligned} h_i|_{\overline{\Omega}_{\text{fin}}} &= \text{ standard cardinal basis function on } \overline{\Omega}_{\text{fin}} \text{ corresponding to} \\ &\qquad \text{the node } (x_i, z_i) \text{ in } \tau_h \\ h_i|_{K_i} &= \psi_1^{K_i} \text{ corresponding to } \hat{\psi}_1 \text{ on } \hat{K} \text{ through } \mathcal{F}_{K_i} \\ h_i|_{K_{i-1}} &= \psi_4^{K_{i-1}} \text{ corresponding to } \hat{\psi}_4 \text{ on } \hat{K} \text{ through } \mathcal{F}_{K_{i-1}} \\ h_i &= 0 \text{ elsewhere in } \Omega_{\text{inf}} \end{aligned}$$

Here $\hat{\psi}_1$ and $\hat{\psi}_4$ are standard bilinear basisfunctions in $\hat{K}$, associated to the nodes 1 and 4, respectively. Recall that $\hat{\psi}_1(\xi, \eta) = \frac{1}{4}(1 - \xi)(1 + \eta)$, $\hat{\psi}_4(\xi, \eta) = \frac{1}{4}(1 + \xi)(1 + \eta)$.

Finally, the function h_{inf} defined on $\overline{\Omega}_h$ associated to the ***node at infinity*** is

$$\begin{aligned} h_{\text{inf}}|_{\overline{\Omega}_{\text{fin}}} &= 0 \\ h_{\text{inf}}|_{\overline{\Omega}_{\text{inf}}} &= 1 - \tfrac{T+\alpha}{2T+\alpha+z}. \end{aligned}$$

It is continuous on $\overline{\Gamma}_3$ and has the property: $h_{\text{inf}} \to 1$ for $z \to -\infty$. Observe that the function h_{inf} has been constructed as the image of the function

$$\hat{\psi}_2 + \hat{\psi}_3 = \tfrac{1-\eta}{2} \text{ on } \hat{K} \tag{6.38}$$

under the transformation $\mathcal{F}_{K_i}$, (for all $i = 1, \ldots, N$).

One easily sees that

$$X_h \quad = \quad \text{span}\{h_1, \ldots, h_N, h_{N+1}, \ldots, h_{N+M}, h_{\text{inf}}\}.$$

Moreover, these $N + M + 1$ functions are linearly independent and thus form a basis for X_h. Of course, deleting the basisfunctions associated to nodes on $\overline{\Gamma}_{2h}$, we are left with a basis for the space V_h. In the decomposition

$$u_h = \sum_{i=1}^{N+M} c_i h_i + c_{\text{ff}} h_{\text{inf}}, \tag{6.39}$$

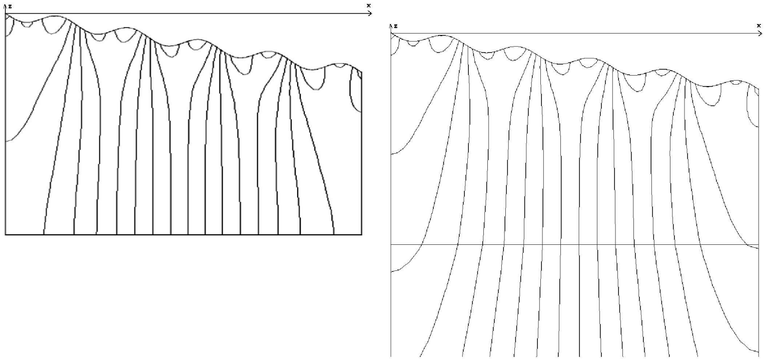


Figure 6.11: Equipotential lines. Left: FEM in case of Neumann BC at the base, with data $L = 8000$, $a/L = 0.1$, $V = 80$, $d = 0.00235$, $T = 3000$. Right: IEM with the same data and $\alpha = -12000$.

it holds that $c_i = u_h(x_i, z_i)$, $i = 1, \ldots, N + M$. Moreover, $c_{\mathrm{ff}} = \lim_{z \to -\infty} u_h$, (independent on x), i.e. c_{ff} represents the constant far field value.

Again, the results of the IEM are in full agreement with those from the literature. In Fig. 6.11 equipotential lines are depicted, to the left obtained with FEM, and to the right obtained with IEM. Both arise for the same data, $L = 8000$, $T = 3000$. The horizontal line in this figure corresponds with $z = -T$. The parameter α has been set at -12000. The domain is divided in 3025 triangles and 40 semi-infinite rectangles. We obtained a far field value $c_{\mathrm{ff}} = -411.971$, which corresponds with the far field value that is obtained with the semi-analytical method of [68], for $N = 100$, namely $c_{\mathrm{ff}} = -411.868$.

6.7 Conclusion

We have developed a semi-analytical method to solve the groundwater flow problem (6.3), under different boundary conditions, in a finite region bounded on top by a sloping sinusoidal boundary. The method is simple and can easily be implemented in a mathematical package. In particular, for the special case of the Laplace equation ($d = 0$), the result has the same qualitative behaviour as the one in [73]. However, our solution is valid on the entire region under consideration, which is not the case in [73]. Furthermore, for the special case of a semi-infinite region ($T = \infty$), the result is in full agreement with [68]. The results over semi-infinite domains coincide with the results obtained for deep regions with an impenetrable lower boundary.

Furthermore, we developed 2 different numerical methods that can be used to readily solve problem (6.3). Both methods are found to give results which are in full agreement with those from the semi-analytical method. They impose no limitation on the form of the top boundary function g. The semi-analytical method could also be adapted to handle this feature by using numerical integration instead of the decomposition in Bessel series.

The second numerical method was an infinite element method. It's worth mentioning that this method provides an alternative for the situation of too many elements that arises in deep regions with the standard FEM. By using an IEM the number of elements could be kept to a minimum, while still providing excellent results. Note that the used standard FEM broke up for deep regions when $d \neq 0$, because of too small entries in the stiffness matrix. Using some form of adaptivity might resolve this difficulty. However, by using IEM, this difficulty can readily be avoided.

Appendix A

Notations, concepts and auxiliary results

A.1 Basic definitions, identities and inequalities

Let $\Omega \subset \mathbb{R}^n$, ($n = 1, 2$ or 3), be an open bounded domain with a Lipschitz-continuous boundary and let $I = (0, T)$, with $T > 0$ finite.

Total variation

There are several definitions possible for the total variation of a function. Most common is the following definition:

Definition A.1.1. *A function $g \in L_\infty(I \times \Omega)$ has a bounded total variation (in τ and x) if*

$$\begin{aligned} TV(g) := \lim_{\epsilon \to 0} \sup \frac{1}{\epsilon} \int_0^T \int_\Omega |g(x+\epsilon, \tau) - g(x, \tau)| \, \mathrm{d}x \mathrm{d}\tau \\ + \lim_{\epsilon \to 0} \sup \frac{1}{\epsilon} \int_0^T \int_\Omega |g(x, \tau+\epsilon) - g(x, \tau)| \, \mathrm{d}x \mathrm{d}\tau < \infty. \quad \text{(A.1)} \end{aligned}$$

Abel's summation

Abel's summation is a discrete version of the formula of integration by parts. We have that

$$\sum_{i=1}^{m} b_i(a_i - a_{i-1}) = b_m a_m - b_0 a_0 - \sum_{i=1}^{m} (b_i - b_{i-1}) a_{i-1}, \tag{A.2}$$

Cauchy-Schwartz

Let H be an innerproduct space with associated innerproduct $(.,.)_H$ and norm $\|.\|_H$. Then one has

$$(x,y)_H \leq \|x\|_H \|y\|_H, \quad \forall x \text{ and } y \in H.$$

Young, Cauchy

The Young inequality is

$$ab \leq \frac{1}{p}a^p + \frac{1}{q}b^q, \quad \text{with } \frac{1}{p} + \frac{1}{q} = 1, \ (p \text{ and } q \in \mathbb{R}_0^+), \quad \forall a, b \in \mathbb{R}^+.$$

The Cauchy-inequality is the special case

$$ab \leq \frac{a^2}{2} + \frac{b^2}{2}, \quad \forall a, b \in \mathbb{R}^+,$$

from which it follows that

$$ab \leq \frac{\alpha}{2}a^2 + \frac{1}{2\alpha}b^2, \quad \forall \alpha > 0, \quad \forall a, b \in \mathbb{R}^+.$$

Hölder

The Hölder inequality is

$$\int_\Omega |fg| \leq \|f\|_{L_p} \|g\|_{L_q}, \quad \frac{1}{p} + \frac{1}{q} = 1, \ (p \text{ and } q \in \mathbb{R}_0^+), \quad \forall f \in L_p(\Omega), \forall g \in L_q(\Omega).$$

Poincaré

There are several version of this inequality, relating a function to it's gradient. Commonly used is:

$$\exists C \equiv C(\Omega) > 0 \text{ (fixed)}: \quad \|u\|_{L_p} \leq C\|\nabla u\|_{L_p}, \quad \forall u \in W_0^{1,p}(\Omega), \quad 1 \leq p < \infty.$$

In one dimension we have

$$\exists C \equiv C(\Omega) > 0 \text{ (fixed)}: \quad \|u\|_{W^{1,p}} \leq C\|u'\|_{L_p}, \quad \forall u \in W_0^{1,p}(\Omega), \quad 1 \leq p < \infty.$$

A.2 Theorems

We begin with the Ascoli-Arzelà Theorem, [77].

Theorem A.2.1 (Ascoli-Arzelà). *Let X be a compact metric space, and $C(X)$ the Banach space of real-valued continuous functions f on X normed by $\|f\| = \sup_{x\in X}|f(x)|$. Then a sequence $\{f_n(x)\} \subseteq C(X)$ is relatively compact in $C(X)$ if the following two conditions are satisfied:*

$$f_n(x) \textit{is equi-bounded (in } n\textit{), i.e.,} \ \sup_{n\geq 1}\sup_{x\in X}|f_n(x)| < \infty,$$

$$f_n(x) \textit{is equi-continuous (in } n\textit{), i.e.,} \ \lim_{\delta\to 0}\sup_{n\geq 1, \mathrm{dist}(x,x')\leq\delta}|f_n(x) - f_n(x')| = 0$$

Now we give the requirements for Kolmogorov compactness, which is an extension of the Ascoli-Arzelà theorem, valid in C, to L_p (see [77], p.275 or [11], p.72).

Theorem A.2.2 (Riesz-Fréchet-Kolmogorov). *Let $\Omega \subset \mathbb{R}^n$ be open and let ω be strongly included*[1] *in Ω. Assume that $\mathcal{F}$ is a bounded subset of $L_p(\Omega)$, with $1 \leq p < \infty$. Suppose that*

$$\forall\epsilon > 0 \quad \exists\delta > 0, \quad \delta < \mathrm{dist}\,(\omega, \mathbf{C}\Omega),$$

such that

$$||f(x+h) - f||_{L_p(\omega)} < \epsilon, \quad \forall h \in \mathbb{R}^n \quad \textit{with} \quad |h| < \delta \quad \textit{and} \quad \forall f \in \mathcal{F}\,(^2). \tag{A.3}$$

Then $\mathcal{F}$ is relatively compact in $L_p(\omega)$.

[1] ω is strongly included in Ω if for the closure of ω in $\mathbb{R}^n$, written $\bar{\omega}$, we have that $\bar{\omega} \subset \Omega$ and $\bar{\omega}$ is compact; we denote this by $\omega \subset\subset \Omega$.

[2] Note that if $x \in \omega$ and $|h| < \delta < \mathrm{dist}\,(\omega, \mathbf{C}\Omega)$, that then $x + h \in \Omega$ and f(x+h) makes sense. This can be seen as an integral equicontinuity condition similar to the one in the Ascoli theorem.

For functions which are also time dependent we can use the triangle inequality

$$||f(x+p,t+q)-f(x,t)|| \leq ||f(x+p,t+q)-f(x,t+q)||+||f(x,t+q)-f(x,t)||,$$

to get Kolmogorov compactness if condition (A.3) is satisfied for every term separately. Note further that requirement (A.3) is weaker than asking that the partial derivative is bounded.

From [11], p. 54

Theorem A.2.3 (Lebesgue dominant convergence theorem). *Let $\{f_n\}$ be a sequence of $L_1(\Omega)$ functions. Suppose that*
a) $f_n(x) \to f(x)$ a.e. in Ω,
b) there exists a function $g \in L_1(\Omega)$ such that for each n, $|f_n(x)| \leq g(x)$ a.e. in Ω.
Then, $f \in L_1(\Omega)$ and $||f_n - f||_{L_1} \to 0$.

Let $h \in L_\infty(\Omega)$. Under de conditions of the theorem one has that $(f_n, h) \to (f, h)$. [Note that $(f_n, h) = (f_n - f, h) + (f, h) \leq ||f_n - f||_{L_1}||h||_{L_\infty} + (f, h)$.]

We end with a result on the extraction of a weak convergent subsequence from a bounded sequence, [17]

Theorem A.2.4 (Eberlein-Shmulyan theorem). *A Banach space X is reflexive if and only if every bounded sequence of X contains a subsequence which converges weakly to an element of X.*

Appendix B

Basic concepts in groundwater flow modeling

B.1 The subsurface

“Subsurface” is the general term to indicate the medium that can be found under the surface. This includes the fixed part of the earth, together with the groundwater and the other substances and organisms within it. The subsurface gives a massive impression, but in reality it is a mixture of particles of soil with holes in between filled with fluid (water, oil) and/or air and gases. Many processes take place: water flows through the empty spaces, bacteria live on the surface of the soil particles, the soil adsorbs contaminants that are in the water, fractures form when the water evaporates.

We focus on groundwater flow and on transport of a single contaminant. Groundwater is present in almost all geological formations. In some materials it flows very slowly, e.g. in clay or stone, in others relatively fast, e.g. in sand or gravel. For most practical applications, the slow groundwater flow can be neglected. Furthermore, it is generally not necessary to study the individual erratic flow of a water molecule through the pores of the subsurface. A more global view will be sufficient for most applications. Then, the movement of groundwater can be described mathematically in a relatively simple way. The main equations of water movement are based on two fundamental principles: Darcy’s law and the conservation of mass.

Material	θ_0 (%)
Gravel, medium	32
Sand, medium	39
Sand, fine	43
Clay	42
Limestone	30
Dolomite	26
Dune sand	45
Loess	49
Shale	6
Tuff	41
Basalt	17

Table B.1: Representative values of porosity θ_0.

B.2 Nomenclatura

The subsurface consists of solid particles and empty spaces, called **pores**. **Porosity** is the amount of pores compared to the the total volume of the underground. We denote it by θ_0. In Table B.1 some representative values are given. The pores are partly filled with air, partly with water. The proportion of these two elements classifies the subsurface. The **water table** is defined as the level to which water will rise in a well drilled into the subsurface. Under this level we have the **saturated zone**. **Capillary water** is the water that is held by surface tension forces just above this water table. We call this zone the **capillary zone** or **fringe**. The **vadose zone** extends from the upper limit of this zone to the lower edge of the **soil-water zone**. Vadose zone water is held in place by hygroscopic, i.e., adsorption to the surface of the soil grains, and capillary forces. Infiltrating water passes downward toward the water table as gravitational flow. The top layer is handled independently, as the water amount in this soil-water zone depends on the weather. In the zone of saturation, the porosity is a direct measure of the water contained per unit volume. When we discuss water flow, this generally only applies to the saturated zone. Only a portion of the water can by removed from this zone by drainage or by pumping from a well. The volume of water released from this zone per unit surface area per unit decline in the water table is called the **specific yield**. More complicated situations may arise, see [9] for details.

Apart from this general vertical division of the subsurface, a division that accounts for the stratified nature of geological formations is often used in groundwater flow modeling. The many layers of the subsurface are subdivided according to their characteristics; three main subdivisions are made. The **aquifer** is a formation that contains sufficient saturated permeable material to yield significant quantities of water to wells or springs. An **aquiclude** is on the opposite side of the spectrum: it is a relatively impermeable confining unit, such as clay. A formation in between these two, such as a sandy clay layer, is called an **aquitard**. This may leak water to adjacent sand aquifers. In many models, one is only concerned about the behaviour of the aquifers. The other layers are viewed only as boundary conditions. In doing so, one must keep in mind that the reality is much more complex. Three dimensional models take all this complexity into account. But the disadvantage is that only in few places the complete stratification of the subsurface is known.

Aquifers form our groundwater resource, and get most attention from modelers. They may be classified as **unconfined** or **confined**. In an unconfined aquifer, there is a water table. A confined aquifer does not have a water table. This can happen e.g. when an aquifer lays in between two aquicludes, and is completely saturated with water.

B.3 Darcy's law and hydraulic conductivity

The first groundwater model was developed in 1856 by Henri Darcy, [19]. In Fig. B.1 we depict his experiment. He investigated the flow of water through columns of sand, by measuring the water levels h_1, h_2. These water levels are refered to as **piezometric surface** or **hydraulic head** or **heads**, for short. We have

$$h = \frac{p}{\gamma} + z,$$

neglecting velocities. Here p is the pressure, γ the specific weight of water, and z the elevation above a horizontal datum. He discovered one of the most important laws in hydrology: *the flow rate through porous media is proportional to the head loss and inversely proportional to the length of the flow path.* The **specific discharge** or **Darcy velocity**, q is,

$$q = -k\frac{dh}{dz}, \tag{B.1}$$

where k is a proportionality constant called the **hydraulic conductivity**. The minus sign indicates that flow is in the direction of decreasing heads. The Darcy

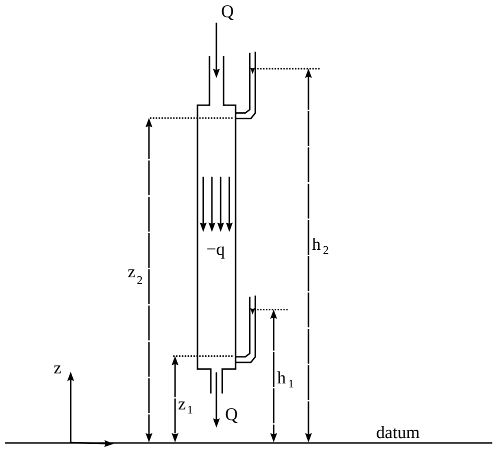


Figure B.1: Darcy's experiment

velocity q is an average discharge velocity through the entire cross section of the column. The actual groundwater molucules are limited to the pore space only, so the **seepage velocity** v_s will be

$$v_s = \frac{q}{\theta_0}. \tag{B.2}$$

The hydraulic conductivity k of a soil or rock depends on a variety of physical factors, and is an indication of an aquifer's ability to transmit water. In Table B.2 we give some representative values. As can be seen, k can vary many orders of magnitude in an aquifer that may contain different types of material. These aquifers are called **heterogeneous aquifers**.

A further complication are variations in one or more directions due to the processes of deposition and layering. This is called **anisotropy**. Mostly, the hydraulic conductivity in the vertical direction is found to be less than the value in the horizontal directions. In three dimensions we write

$$\mathbf{q} = -\mathbf{K}\nabla h, \tag{B.3}$$

where $\mathbf{K}$ is a second order tensor with nine components. For an isotropic medium, this will be $k\mathbf{I}$. Here, k can be a function of the depth z. For a simple anisotropic medium, it will still be a diagonal matrix, but with three different diagonal elements.

Material (unconsolidated)	k (10^{-2} m/sec)
Gravel,	3.0 to 3 × 10^{-2}
Sand, medium	6 × 10^{-2} to 9 × 10^{-5}
Sand, fine	2 × 10^{-2} to 2 × 10^{-5}
Loess	2 × 10^{-3} to 1 × 10^{-7}
Clay	5 × 10^{-7} to 1 × 10^{-9}
Limestone and dolomite	6 × 10^{-4} to 1 × 10^{-7}
Shale	2 × 10^{-7} to 1 × 10^{-11}
Basalt	4 × 10^{-5} to 2 × 10^{-9}
Permeable basalt	2 to 4 × 10^{-5}

Table B.2: Representative values of hydraulic conductivity k.

Darcy's law applies to laminar flow (no large pores) in porous media, and is certainly valid for Reynolds number[1] less than 1, which is applicable in most groundwater systems.

B.4 Mass conservation equation

We take a representative elementary volume. More precisely, we consider a rectangular box with measures Δx, Δy and Δz, centred at the point (x, y, z), and having its boundary planes two by two orthogonal to the X-, Y- and Z-axis, respectively. The law of conservation of mass in a time interval $(t, t + \Delta t)$ requires that

$$\text{Mass in} - \text{Mass out} = \text{change in storage.}$$

In the limit $\Delta x \to 0$, $\Delta y \to 0$, $\Delta z \to 0$ and $\Delta t \to 0$, this readily leads to

$$[-\partial_x(\rho_w q_x) - \partial_y(\rho_w q_y) - \partial_z(\rho_w q_z)] = \partial_t(\rho_w \theta_0), \tag{B.4}$$

in which ρ_w is the density of water in the point (x, y, z) at time t, q_x, q_y and q_z are the components of the Darcy flux and θ_0 is the porosity. If we assume a constant ρ_w, substitute (B.3) and take steady state conditions, we obtain

$$\nabla \cdot (\mathbf{K}\nabla h) = 0. \tag{B.5}$$

[1] The Reynolds number is defined by $\rho q d/\mu$, with d the pore space, μ the viscosity of the pore fluid, ρ the density.

B.5 Dupuit-Forchheimer flow

In many groundwater models, the Dupuit-Forchheimer approximation is used. Dupuis (1863) and Forchheimer (1886) independently suggested that flow lines are predominantly horizontal, and velocities do not vary over the aquifer depth. See [24], Chapter 3, for details, or [9], Section 2.5. We briefly recall the main results here. The Dupuit-Forchheimer approximation assumes that

$$\frac{\partial h}{\partial z} = 0.$$

This leads to one of the most essential simplifications of real-world groundwater flow problems: three-dimensional flow problems reduce to two-dimensional ones. This assumption is normally valid when the length of a flow line is large compared to the aquifer thickness or when the head gradient is not large.

A usefull concept is the fluid potential. We first define the **discharge** Q in the x and y direction as

$$Q_x = \int_0^{h_{\text{eff}}} q_x dz, \quad Q_y = \int_0^{h_{\text{eff}}} q_y dz, \tag{B.6}$$

where h_{eff} is the effective height of the water column, i.e. $h_{\text{eff}} = \min(H, h)$, with H being the height of the aquifer and the head h being measured from the base of the aquifer on. In the Dupuit-Forchheimer approximation the specific discharge $\mathbf{q}$ does not vary over the aquifer height, and hence

$$Q_x = h_{\text{eff}} q_x, \quad Q_y = h_{\text{eff}} q_y.$$

The **flow potential** or **discharge potential** $\Phi(x, y)$ is the function for which holds

$$Q_x = -\frac{\partial \Phi}{\partial x}, \quad Q_y = -\frac{\partial \Phi}{\partial y}. \tag{B.7}$$

Under the steady state conditions (B.5) one has

$$\Delta \Phi = 0.$$

Combining (B.7) and (B.2), we obtain the seepage velocity as

$$\mathbf{v}_s = -\frac{1}{h_{\text{eff}} \theta_0} \nabla \Phi. \tag{B.8}$$

The flow potential is known for many elementary situations, see [24]. Relevant to our work is the relation between the flow potential and hydraulic head. For confined and unconfined flow we respectively have that

$$\begin{aligned} \Phi(x,y) &= kHh(x,y) - \tfrac{1}{2}kH^2, \quad (h \geq H), && \text{(B.9)} \\ \Phi(x,y) &= \tfrac{1}{2}kh^2(x,y), \quad (h \leq H). && \text{(B.10)} \end{aligned}$$

In the case of uniform flow field in the x-direction with discharge Q_0, it holds

$$\Phi = -Q_0 x + C.$$

In the case of a well, the **pumping rate** Q_w (=discharge rate) must be given. In polar coordinates (r,θ), assuming circular symmetry of the well, (B.7) gives

$$\frac{d\Phi}{dr} = \frac{Q_w}{2\pi r},$$

where the right hand side follows from the continuity of the flow across a circle of radius r around the well. Integration gives

$$\Phi(r) = \frac{Q_w}{2\pi}\ln r + C, \tag{B.11}$$

where the integration constant C must be chosen so that Φ satisfies the boundary condition. Using (B.9)-(B.10) to obtain the prescribed flow potential Φ_0 at a distance R, we can write

$$\Phi(r) = \frac{Q_w}{2\pi}\ln\frac{r}{R} + \Phi_0.$$

An important principle for flow potentials, is the principle of superposition. As an example, the flow potential of a well Q_w at the origin, placed in a uniform flow field Q_0 in the x-direction, is given by

$$\Phi = -Q_0 x + \frac{Q_w}{2\pi}\ln r + C.$$

B.6 Contaminant transport

Contaminants in the groundwater will move around with the water. This mass transport will cause changes in solute concentration. The primary causes are:

1. **Advection**. The solute flows with its carrier, the solvent.

2. **Hydrodynamic Dispersion**. The combined effects of mechanical dispersion and molecular diffusion spreads the contaminant out. Mechanical dispersion is a mechanical process: because of the stochastic nature of the pore space distribution in porous media and the nonhomogeneity of the microscopic velocity distribution, the tracer particle groups are being separated continously during the flow process. This causes the tracer to spread out more than what is expected from just the mean flow velocity. Molecular diffusion is caused by the nonhomogeneous distribution of the tracer particles in a fluid. The tracer molecules in high concentration will move towards the low concentration areas. Normally, mechanical dispersion plays the major role, but when the flow velocity is extremely low, molecular diffusion may become more prominent. Dispersion along the mean flow direction is called longitudinal dispersion; dispersion perpendicular to it is called transversal dispersion.

3. **Sources and sinks**. A well can pump contaminant in or out, a buried tank can leak contaminant, etc.

4. **Adsorption and ion exchange**. Adsorption and ion exchange occur at the interface between the solid and liquid phases. The solute in the liquid may be adsorbed by the solid. The mass in the solid may also get into the liquid by dissolution or by ion exchange.

5. **Chemical reaction and biological processes**. Chemical reactions can change the solute. Biological processes such as the reproduction of bacteria will also change the concentration of certain solutes.

6. **Radioactive decay**. Radioactive components within the fluid will decrease in concentration as a result of decay.

All these factors should be taken into consideration. However, the importance of each factor may differ strongly from case to case. The convection-diffusion equation for a contaminant in an isotropic medium (see [8, 69]) reads

$$\partial_t(\theta_0 C) = \nabla \cdot (\theta_0 \mathbf{D} \nabla C) - \nabla \cdot (\theta_0 \boldsymbol{v}_s C) + I. \tag{B.12}$$

Here, C is the concentration of the contaminant in the groundwater, I denotes the sources or sinks, the velocity $\boldsymbol{v}_s = (v_1, v_2, v_3)$ is given by (B.8), and $\mathbf{D}$ is the dispersivity tensor, given by

$$D_{ij} = \{(D_0 + \alpha_T |v_s|)\delta_{ij} + \frac{v_i v_j}{|v_s|}(\alpha_L - \alpha_T)\}, \quad i,j = 1,2,3, \tag{B.13}$$

where D_0 is the molecular diffusion, α_L the longitudinal dispersivity, α_T the transversal dispersivity, and δ_{ij} the Kronecker symbol. If we inject water with tracer concentration C_0 into an aquifer, and the water injected per unit time per unit porous media is W_I, we have $I = W_I C_0$. In the case of extraction, with W_E the water extracted per unit time per unit porous media, the sink term becomes $I = -W_E C$.

Remark B.6.1. *The expression (B.13) follows from a specific model, see [69]. Different models are possible. Therefore, care should be taken to verify the validity of the approach to a specific set-up. Certain properties must be checked, e.g., the aquifer must be isotropic. Note also that mechanical dispersion cannot cause a contaminant to move against the direction of flow; only molecular diffusion can have this effect. If, however, in (B.13) the velocity is very high, the diffusion is high too and a net flow of contaminant against the groundwater flow could be observed, breaking the validity of this specific dispersivity model. One could then, as a first modification of (B.13), make α_L and α_T dependant on the mean velocity $\boldsymbol{v}_s$, reducing their value when the mean velocity becomes large.*

We now rewrite (B.12) for the case of the Dupuit-Forchheimer approximation. In a Dupuit-Forchheimer approximation, we reduce the problem to a 2D setting by making the mass balance in the xy-plane. Here, diffusive and advective fluxes are multiplied by the effective height of the watercolumn, h_{eff}. Thus, the advective flux through a point (x, y) is given by $h_{\text{eff}}\theta_0 \boldsymbol{v}_s C(x, y)$. Hence, (B.12) reduces to

$$\partial_t(\theta_0 h_{\text{eff}} C) = \nabla \cdot (\theta_0 h_{\text{eff}} \mathbf{D} \nabla C) - \nabla \cdot (\theta_0 h_{\text{eff}} \boldsymbol{v}_s C) + h_{\text{eff}} I, \tag{B.14}$$

where $\boldsymbol{v}_s$ and $\mathbf{D}$ are now given by their two-dimensional analogon.

B.7 Adsorption

The effect of adsorption can be assigned to a source/sink term. Consider simultaneously the mass balance within the solid phase (the soil matrix) and the fluid phase. In the equilibrium assumption, the quantities of the tracer in solid and groundwater are continuously in equilibrium. Thus, a change in one phase immediately causes a change in the other. Denoting by S the concentration of the tracer in the solid phase, i.e., the tracer mass in a unit volume of the solid, the tracer mass conservation gives

$$\partial_t(\varrho S) = f. \tag{B.15}$$

Here, we have introduced the unknown function f for the mass of tracer transferred from liquid to solid per unit time and per unit volume of porous media. ϱ is the density of the porous media where adsorption takes place. In certain applications it is valid to put $\varrho = 1 - \theta_0$, i.e. the fraction of the soil matrix in the total volume. In the liquid phase we have

$$\partial_t(\theta_0 C) = \nabla \cdot (\theta_0 \mathbf{D} \nabla C) - \nabla \cdot (\theta_0 \boldsymbol{v}_s C) - f. \tag{B.16}$$

Eliminating the unknown function f from (B.15)-(B.16) yields

$$\partial_t(\theta_0 C) = \nabla \cdot (\theta_0 \mathbf{D} \nabla C) - \nabla \cdot (\theta_0 \boldsymbol{v}_s C) - \partial_t(\varrho S). \tag{B.17}$$

Here, S will be some function of C, $S = \Psi(C)$, which we call the sorption isotherm[2].

Many isotherms are considered in the geo-hydrological literature, depending on the particular problem considered. The selection of the appropriate isotherm is based on the study of the interacting components and on experiments. The following sorption isotherm are most common, [8]

1. The linear isotherm
$$\Psi(C) = aC + b. \tag{B.18}$$

2. Langmuir isotherm
$$\Psi(C) = \frac{aC}{1 + bC}. \tag{B.19}$$

3. Freundlich isotherm
$$\Psi(C) = aC^p. \tag{B.20}$$

4. Lindstrom-Van Genuchten isotherm
$$\Psi(C) = aCe^{-2b\Psi(C)}. \tag{B.21}$$

If the porosity θ_0 and the density ϱ are constants, we can write (B.17) concisely as

$$\partial_t(C) = \nabla \cdot (\mathbf{D} \nabla C) - \nabla \cdot (\boldsymbol{v}_s C) - \partial_t(\Psi(C)), \tag{B.22}$$

where the fraction $\frac{\varrho}{\theta_0}$ is included into the constants that appear in the sorption isotherm $\Psi(C)$.

[2] In some textbook S is the tracer mass per unit *mass* of solid. Then, in the formulas, $S\rho_s$ will be the tracer mass per unit volume of the solid, where ρ_s is the density of the solid, [8].

In non-equilibrium, the amount of adsorption will deviate from the equilibrium adsorption. Therefore, the concentration of the tracer in the solid phase obeys

$$\partial_t S = \kappa(\Psi(C) - S), \tag{B.23}$$

where κ is the rate constant of adsorption. This equation is then coupled with (B.17). Equilibrium is reached for $\kappa \to \infty$.

B.8 Radioactive decay

For radioactive decay, the term I in (B.12) is replaced by

$$I = -\lambda\theta_0 C,$$

where λ is the decay constant. If we consider both decay and adsorption, the tracer mass conservation in the solid phase leads to

$$\partial_t((1-\theta_0)S) = f - \lambda(1-\theta_0)S, \tag{B.24}$$

where again f is the mass of tracer transferred from liquid to solid per unit time and per unit volume of porous media. In the liquid phase we have

$$\partial_t(\theta_0 C) = \nabla \cdot (\theta_0 \mathbf{D}\nabla C) - \nabla \cdot (\theta_0 \boldsymbol{v}_s C) - f - \lambda\theta_0 C. \tag{B.25}$$

Eliminating the unknown function f from (B.24)-(B.25), we obtain

$$\partial_t(\theta_0 C) = \nabla\cdot(\theta_0 \mathbf{D}\nabla C) - \nabla\cdot(\theta_0 \boldsymbol{v}_s C) - \partial_t((1-\theta_0)S) - \lambda(1-\theta_0)S - \theta_0\lambda C, \tag{B.26}$$

which, under constant porosity θ_0, can be written as

$$\partial_t(C) = \nabla \cdot (\mathbf{D}\nabla C) - \nabla \cdot (\boldsymbol{v}_s C) - \partial_t(\Psi_1(C)) - \lambda\Psi_1(C) - \lambda C, \tag{B.27}$$

where $\frac{1-\theta_0}{\theta_0}$ is included in Ψ_1.

In the non-equilibrium case, (B.26) needs to be coupled with

$$\partial_t S = \kappa(\Psi_2(C) - S) - \lambda S. \tag{B.28}$$

Appendix C

Basic facts on numerical methods for inverse problems

C.1 Introduction

An inverse problem aims at determining the parameters of a model so that the solution itself satisfies certain given conditions. Typically, the conditions are experimental values at a certain time point t. However, other types of conditions can be required, such as for example smoothness of a prescribed function.

Inverse problems can be considered to be at least as important, if not more, than direct problems. Consider for example CFD: huge effort has been done to efficiently calculate the pressure and turbulent flow around airplane wings, giving the impression that little will be gained from future developments. However, the industry aims at creating airplanes with optimal drag and lift, not at calculating pressure curves around the wings, although that information is needed to achieve the goal. The real quest, which will require huge research effort in the coming years, is that for models that not only solve the direct problem, but also determine the optimal wing within certain design bounds.

Inverse problems are known to be often *ill-posed.* This means that small changes in the conditions which the solution of the inverse problem must satisfy, can lead to large changes in the parameters searched. This is the so-called instability of the inverse method. Fortunately, it is often possible to impose additional constraints that bias the solution. This is called *regularization.* Regularization is often essential to obtain reliable solutions to ill-posed or ill-conditioned inverse problems.

In the case of diffusion problems like the ones studied in Part I, the ill-posedness comes from the physical background. No matter what the diffusion coefficient is, if the boundary conditions considered are no-flow boundaries, the result at large time will be the same constant value over the entire domain. From this constant value, no inverse method is capable of determining parameters of the model. This emphasizes the importance of the construction of meaningfull experiments, like e.g. the dual-well.

Two additional reasons, apart from the ill-posedness, make inverse problems *hard*. The first is the problem of *existence of a solution*: the existence of a model fitting the given required conditions is uncertain. The reason can be the approximation in the physical model or noise in the data. Linked to this is the error margin of the measurements which will result in ranges of parameter values instead of a well defined value. Secondly, *uniqueness* is an issue: if exact solutions exist, they may not be unique. The classic example is the external gravitational field from a spherically symmetric mass distribution, which depends only on the mass, not on the radial density distribution.

In this chapter we focus on the mathematical tools (inverse problem for PDE, and the theory of adjoint methods) needed in the later chapters. We refer to [1] for a general introduction to inverse problems, as well as to [74], where the emphasis is on computational methods. For optimization techniques we refer the reader to [64].

C.2 The penalty function

If an analytical solution is not known, a precise numerical method is needed to solve the direct problem. The inverse problem is then solved by minimising a *penalty function* $\mathcal{F}$, also called *cost functional.* The penalty function measures the deviation of the experiment from the numerical solution. Suppose that N experimental values, ϕ_i, taken at time t_i and/or at position x_i, $i = 1, \ldots, N$, are given. Let σ_i be their deviation. Then, a typical choice for the penalty function is the least squares fit of the data:

$$\mathcal{F}(\boldsymbol{p}) = \frac{1}{2} \sum_{i=1}^{N} \left(\frac{\phi_{\boldsymbol{p}}(t_i, x_i) - \phi_i}{\sigma_i} \right)^2 , \tag{C.1}$$

where $\phi_{\boldsymbol{p}}$ is the numerical solution obtained with the parameter set $\boldsymbol{p}$.

The penalty function is often regularized to overcome the ill-posedness of the inverse problem. The most common regularization is the so-called Tikhonov filter, see [1, 74]. We will not need any regularization for the inverse problems considered here, because

- the results obtained without regularization are physically acceptable and reasonable;
- regularization creates a bias in the solution of the inverse problem. This is acceptable if it has a physical basis, which is not obvious for the problems we will consider. Indeed, the parameters are not related with each other, which makes Tikhonov type of regularization unsuitable.

It must be mentioned that not applying some type of regularization is not standard. However, if good results are obtained, it is acceptable. Applying a regularization adds extra terms to the penalty function (C.1), see again [1, 74] for details. All inverse methods given in the next sections can be adapted consequently.

C.3 Abstract framework

We present the different methods in an abstract framework. Consider the initial value problems of the form

$$\begin{aligned} \dot{u}(t) &= A(\boldsymbol{q})u(t) + f(t), \quad 0 < t < t_F, && \text{(C.2)} \\ u(0) &= 0. && \text{(C.3)} \end{aligned}$$

Here $A(\boldsymbol{q})$ is a bounded linear operator on a Hilbert space H, depending on parameters $\boldsymbol{q}$; the inner product on H is denoted by $\langle \cdot, \cdot \rangle_H$ and $\dot{u}(t)$ indicates differentiation of $u(t)$ with respect to t. We assume that $\boldsymbol{q}$ lies in a set $\mathcal{Q}_{AD}$ of admissible parameters contained in a normed linear parameter space $\mathcal{Q}$. The map $\boldsymbol{q} \longmapsto A(\boldsymbol{q})$ is supposed to be Gateaux differentiable in the operator norm, and the derivative is denoted by $\frac{dA}{d\boldsymbol{q}}$.

In general it is assumed that the solution u belongs to the state space

$$\mathcal{H} = L^2(0, t_F; H),$$

which is a Hilbert space with inner product

$$\langle f, g \rangle_{\mathcal{H}} = \int_0^{t_F} \langle f, g \rangle_H \, \mathrm{d}t.$$

Moreover, we assume the existence of an observation space $\mathcal{Z}$, which is a Hilbert space with inner product $\langle \cdot, \cdot \rangle_{\mathcal{Z}}$, and an *observation operator*

$$\mathcal{C} : \mathcal{H} \to \mathcal{Z}.$$

Given an observation z of u, the goal is to determine the parameter $\boldsymbol{q}$. The abstract translation of (C.1), neglecting the deviation, is then: determine $\boldsymbol{q} \in \mathcal{Q}_{AD}$ that minimizes the functional

$$\mathcal{J}(\boldsymbol{q}) := \tfrac{1}{2} \|\mathcal{C}u(\boldsymbol{q}) - z\|_{\mathcal{Z}}^2, \tag{C.4}$$

where $u(\boldsymbol{q})$ is the solution of (C.2).

The gradient of $\mathcal{J}$ defined by

$$[\operatorname{grad} \mathcal{J}(\boldsymbol{q})]_i := \frac{d}{dh} \mathcal{J}(\boldsymbol{q} + h\boldsymbol{e_i})\Big|_{h=0} = \frac{\partial \mathcal{J}(\boldsymbol{q})}{\partial q_i}, \tag{C.5}$$

is often used to obtain $\min_{\boldsymbol{q} \in \mathcal{Q}_{AD}} \mathcal{J}(\boldsymbol{q})$. Here, $\boldsymbol{e}_i$ is the standard ith unit vector of $\mathcal{Q}$. For further use we also define the Hessian

$$[\operatorname{Hess} \mathcal{J}(\boldsymbol{q})]_{ij} := \frac{\partial^2 \mathcal{J}(\boldsymbol{q})}{\partial q_i \partial q_j}. \tag{C.6}$$

C.4 Optimization: Gradient based methods

The goal of this section is to provide the tools to analyze and compute minimizers for the cost functional. We focus on gradient based methods, also including those based on higher order derivatives like the Hessian that can effectively be estimated by the gradient. These methods are the most common ones, but other methods, like genetic algorithms, exist.

Consider a functional $\mathcal{J} : \mathbb{R}^n \to \mathbb{R}$. Assume throughout that $\mathcal{J}$ is sufficiently smooth, i.e. it has derivatives of sufficiently high order to implement the envisaged methods.

C.4.1 Steepest descent method

We consider the following algorithm.

Steepest Descent Method

```
    $n := 1$;
    $\boldsymbol{p}_0 :=$ initial guess;
```

```
begin steepest descent iterations
    $\boldsymbol{g}_n := -\mathrm{grad}\,\mathcal{J}(\boldsymbol{p}_0)$;        #negative gradient
    $z_n := \arg\min_{z>0} \mathcal{J}(\boldsymbol{p}_0 + z\boldsymbol{g}_n)$; #line search
    $\boldsymbol{p}_{n+1} := \boldsymbol{p}_n + z_n\boldsymbol{g}_n$;
    $n := n + 1$;
end steepest descent iterations
```

One of the key components is the line search. A common way to implement it is to suppose that the functional $\mathcal{J}$ is quadratic, and estimate the initial value of z in the line search. Optimizing this search is relevant since each evaluation of $\mathcal{J}$ involves the computationally difficult task of solving a direct problem. Inexact line search algorithms have been developed. See [74].

The steepest gradient method exhibits slow convergence in the case of ill-conditioned systems, i.e. $\mathcal{J}$ has an ill-conditioned Hessian. A rapidly convergent alternative is then the conjugate gradient method.

C.4.2 Conjugate gradient method

The conjugate gradient method converges faster than the steepest descent method. Indeed, it provides the means to gain an optimal search direction pointing to the minimum of the functional instead of to the steepest descent direction. This method is mathematically based on the observation that finding the minimum of $\|Gh - f\|$ corresponds to solving the equation $Gh = f$, where G is a $n \times n$ nonsingular matrix. This equation is readily solved by expressing the problem in terms of a G^TG-conjugate basis.

We now describe the version of the CG method for positive quadratic functionals.

CG method for quadratic minimization
We minimize $\mathcal{J}(\boldsymbol{p}) = c + \langle \boldsymbol{b}, \boldsymbol{p} \rangle + \frac{1}{2}\langle H\boldsymbol{p}, \boldsymbol{p} \rangle$, where H is symmetric, positive definite.

```
$n := 1$;
$\boldsymbol{p}_0 :=$ initial guess;
$\boldsymbol{g}_0 := H\boldsymbol{p}_0 + \boldsymbol{b}$; #initial gradient
$\boldsymbol{h}_0 := -\boldsymbol{g}_0$; #initial search direction
begin CG iterations
    $z_n := \arg\min_{z>0} \mathcal{J}(\boldsymbol{p}_n + z\boldsymbol{h}_n)$; #line search
    $\boldsymbol{p}_{n+1} := \boldsymbol{p}_n + z_n\boldsymbol{h}_n$;
    $\boldsymbol{g}_{n+1} := H\boldsymbol{p}_{n+1} + \boldsymbol{b}$;
    $\gamma_n := \frac{\langle H\boldsymbol{h}_n, \boldsymbol{g}_{n+1}\rangle}{\langle \boldsymbol{h}_n, H\boldsymbol{h}_n\rangle}$;
    $\boldsymbol{h}_{n+1} := -\boldsymbol{g}_{n+1} + \gamma_n \boldsymbol{h}_n$;        # so that $\langle \boldsymbol{h}_{n+1}, H\boldsymbol{h}_n\rangle = 0$
    $n := n + 1$;
end CG iterations
```

In the above algorithm it is possible to replace the line search with a calculated step based on the quadratic form. However, this is not the case in general problems. In fact, for those one needs a nonlinear CG method such as the Fletcher-Reeves conjugate gradient method, [64, 74].

C.4.3 Newton methods and the practical Gauss-Newton method

Newton's algorithm is one of the oldest and best methods for solving root finding problems. In its simplest form it converges only if the initial guess is sufficiently close to a solution.

In the Newton method we consider the quadratic approximation to $\mathcal{J}(\boldsymbol{q}+\boldsymbol{s})$,

$$Q_n(\boldsymbol{s}) = \mathcal{J}(\boldsymbol{q}) + \langle \mathrm{grad}\,\mathcal{J}(\boldsymbol{q}), \boldsymbol{s}\rangle + \tfrac{1}{2}\langle \mathrm{Hess}\mathcal{J}(\boldsymbol{q})\boldsymbol{s}, \boldsymbol{s}\rangle. \tag{C.7}$$

If the Hessian is positive definite, then $Q_n(\boldsymbol{s})$ has a unique minimizer which satisfies

$$\mathrm{grad}\mathcal{J}(\boldsymbol{q}) + \mathrm{Hess}\mathcal{J}(\boldsymbol{q})\boldsymbol{s} = 0. \tag{C.8}$$

Taking $\boldsymbol{q} + \boldsymbol{s}$ as a new estimate for the minimizer of $\mathcal{J}$ starting from $\boldsymbol{q}$, we obtain the Newton iteration

$$\boldsymbol{q}_{n+1} = \boldsymbol{q}_n - [\mathrm{Hess}\mathcal{J}(\boldsymbol{q})]^{-1}\mathrm{grad}\mathcal{J}(\boldsymbol{q}), \quad n = 1, 2, \ldots.$$

The advantage is that the method is straightforward, and no line search is needed.

However, the disadvantage is twofold. First, convergence is guaranteed only when the initial estimate is sufficiently close to a local minimizer, with rapid convergence only obtained near the minimizer. Secondly, it may be quite expensive to compute the Hessian and solve the linear system (C.8).

The first disadvantage can be overcome by applying a trust region globalization, as explained in the next subsection. The second disadvantage can be overcome by replacing the true Hessian by an approximation derived from current and previous gradients of $\mathcal{J}$. The most popular scheme for this is the BFGS method, see [64, 74]. This method, however, does not guarantee positive definiteness of the approximated Hessian in all cases. In this thesis, we choose the Gauss-Newton method, [1].

Gauss-Newton method The Hessian of $\mathcal{J}$ is approximated by

$$\text{Hess}\mathcal{J}(\boldsymbol{q}) \approx \frac{\text{grad}\,\mathcal{J}(\boldsymbol{q})[\text{grad}\,\mathcal{J}(\boldsymbol{q})]^T}{2\mathcal{J}(\boldsymbol{q})}. \tag{C.9}$$

This approach is validated by (C.4). From that equation we have $[\text{grad}\,\mathcal{J}(\boldsymbol{q})]_k = \|\mathcal{C}u(\boldsymbol{q}) - z\|_{\mathcal{Z}}\frac{\partial}{\partial p_k}\|\mathcal{C}u(\boldsymbol{q}) - z\|_{\mathcal{Z}}$, and the (j,k) element of the Hessian reads

$$\begin{aligned}
[\text{Hess}\mathcal{J}(\boldsymbol{q})]_{jk} &= \frac{\partial}{\partial q_j}\left(\|\mathcal{C}u(\boldsymbol{q}) - z\|_{\mathcal{Z}}\frac{\partial}{\partial q_k}\|\mathcal{C}u(\boldsymbol{q}) - z\|_{\mathcal{Z}}\right) \\
&= \frac{\partial}{\partial q_j}\|\mathcal{C}u(\boldsymbol{q}) - z\|_{\mathcal{Z}}\frac{\partial}{\partial q_k}\|\mathcal{C}u(\boldsymbol{q}) - z\|_{\mathcal{Z}} \\
&\quad +\|\mathcal{C}u(\boldsymbol{q}) - z\|_{\mathcal{Z}}\frac{\partial^2}{\partial q_j \partial q_k}\|\mathcal{C}u(\boldsymbol{q}) - z\|_{\mathcal{Z}}.
\end{aligned}$$

From the above it follows that in the Gauss-Newton approach, the last term is neglected, which is valid if $\|\mathcal{C}u(\boldsymbol{q}) - z\|_{\mathcal{Z}}$ is small.

In practice, this method is applied to (C.1) as follows (see also [6]). Assume $\boldsymbol{p}_n$ to be the parameter vector at the n-th iteration of the inverse algorithm, having dimension M. The new parameter vector $\boldsymbol{p}_{n+1}$ can be found from

$$J_n^T J_n\left(\boldsymbol{p}_{n+1} - \boldsymbol{p}_n\right) + J_n^T F_n = 0, \tag{C.10}$$

where J_n^T is the transposed of the Jacobian matrix J_n, defined by

$$(J_n)_{i,j} = \partial_{p_j}\frac{\phi_{\boldsymbol{p}}(t_i, x_i)}{\sigma_i}, \quad i = 1, \ldots, N, \quad j = 1, \ldots, M, \tag{C.11}$$

and F_n is the column matrix defined by

$$(F_n)_i = \frac{\phi_{\boldsymbol{p}}(t_i, x_i) - \phi_i}{\sigma_i}, \quad i = 1, \ldots, N. \tag{C.12}$$

From (C.10) it follows that

$$\boldsymbol{p}_{n+1} = \boldsymbol{p}_n - (J_n^T J_n)^{-1} J_n^T F_n. \tag{C.13}$$

The practical Gauss-Newton method given here is somewhat more elaborated than the standard scheme. Indeed, it uses a Jacobian matrix calculated in all experiment points (x_i, t_i), instead of the gradient of the cost functional. This supports the practical implemantation of a finite difference based deduction of the gradient which relies on storing the change in the measurement value in every experimental point, see Section C.5.

Although the Newton-Gauss method only uses first order derivatives, it converges quadratically if $\boldsymbol{p}_n$ is close to the optimum value $\boldsymbol{p}^*$ which in turn has a residual tending to zero, i.e. $\|\mathcal{C}u(\boldsymbol{p}^*) - z\|_{\mathcal{Z}} \to 0$. If the residual is non-zero the rate may become linear, and if it is too large, or if the initial value is far from the optimum, the method will diverge.

C.4.4 Trust region globalization and the practical Levenberg-Marquardt method

In this approach, the implementation requires the solution $\boldsymbol{s}$ of a quadratic constrained minimization problem,

$$\min_s \mathcal{Q}_n(\boldsymbol{s}) \quad \text{subject to} \quad \|\boldsymbol{s}\| \leq \Delta_n. \tag{C.14}$$

Here $\mathcal{Q}_n$ is the quadratic approximation (C.7) of $\mathcal{J}$, while Δ_n is a positive scalar, called the trust region radius, which is varied as the iteration proceeds. We obtain iterations of the form

$$\boldsymbol{q}_{n+1} = \boldsymbol{q}_n - [\text{Hess}\mathcal{J}(\boldsymbol{q}) + \lambda_n I]^{-1}\text{grad}\mathcal{J}(\boldsymbol{q}), \quad n = 1, 2, \ldots,$$

where λ_n is zero if the constraint is inactive at iteration n (Newton's method is obtained), and a positive Lagrange multiplier otherwise. Trust region methods tend to be more robust for ill-conditioned problems than line search techniques. However, their implementation can be problematic, as the exact solution of (C.14) can be computationally very expensive. Therefore, several approximate solution schemes have been developed, see [64] section 1.2.4. One of earliest applications is the Levenberg-Marquardt method.

Levenberg-Marquardt method The Levenberg-Marquardt method is the simplest extension of the Gauss-Newton method. It minimizes (C.1) under the constraint that the step taken from $\boldsymbol{p}_n$ to $\boldsymbol{p}_{n+1}$ is on a hypersphere of radius Δ_n. With the Lagrange multiplicator method, this gives, see [6], the update formula

$$\boldsymbol{p}_{n+1} = \boldsymbol{p}_n - (J_n^T J_n + \lambda_n I)^{-1} J_n^T F_n, \tag{C.15}$$

where λ is the Lagrange multiplicator, I the unity matrix of order n, and J_n, F_n are given by (C.11), (C.12). A starting value for the extra parameter λ_n must be chosen. A good initial value is $\lambda_1 = \mathrm{Tr}(J_1^T J_1)/\mathrm{Tr}(I)$, [70]. If $\mathcal{F}(\boldsymbol{p}_{n+1}) < \mathcal{F}(\boldsymbol{p}_n)$, it is retained and we take a smaller multiplicator $\lambda_{n+1} < \lambda_n$ in the next step. Otherwise, the parameter set $\boldsymbol{p}_{n+1}$ is discarded and another one is sought with a larger λ_n value, e.g. $\lambda_n = 2\lambda_n$, replacing λ by twice its value. Note that if $\lambda_n = 0$, the Gauss-Newton method is obtained, whereas if λ_n is large we are close to a steepest descent method.

The advantage of the Levenberg-Marquardt method is that, by reducing λ, a quadratic behaviour of the method near the optimum is achieved. On the other hand, by increasing λ further away from the optimum, the method does not diverge as easily, and might escape from local minima of the penalty function.

If the parameters are all positive, and have values several orders of magnitude apart, it is worthwile to apply the Levenberg-Marquardt method to the logarithm of the parameters. The reason is that the method produces equal step sizes for all parameters, which is of no use if the parameters are not of the same order. Minimizing the logarithm overcomes this drawback. Thus, the parameter vector in (C.15) is $\widetilde{\boldsymbol{p}}_n = [\ln p_1^{(n)}, \ldots, \ln p_M^{(n)}]^T$, where $p_j^{(n)}$ is the jth component of parameter $\boldsymbol{p}$ in the nth iteration. Hence, the jth component of $\widetilde{\boldsymbol{p}}$ is found from

$$\ln p_j^{(n+1)} = \left[-(\widetilde{J}_n^T \widetilde{J}_n + \lambda_n I)^{-1} \widetilde{J}_n^T F_n\right]_j + \ln p_j^{(n)}, \tag{C.16}$$

where F_n is similar as in (C.12), and $\widetilde{J}_n$ is given by

$$[\widetilde{J}_n]_{i,j} = \partial_{\ln(p_j)} \phi_{\boldsymbol{p}_n}(t_i, x_i) = p_j \partial_{p_j} \phi_{\boldsymbol{p}_n}(t_i, x_i) = p_j [J_n]_{i,j},$$

$(i = 1, \ldots, N,\ j = 1, \ldots, M)$. Here, J_n is as in (C.11). From (C.16) we get

$$p_j^{(n+1)} = p_j^{(n)} e^{\left[-(\widetilde{J}_n^T \widetilde{J}_n + \lambda I)^{-1} \widetilde{J}_n^T F_n\right]_j}, \tag{C.17}$$

which can be approximated by

$$p_j^{(n+1)} = p_j^{(n)} + \left[(\widetilde{J}_n^T \widetilde{J}_n + \lambda I)^{-1} \widetilde{J}_n^T F_n\right]_j p_j^{(n)}. \tag{C.18}$$

C.5 Gradient determination methods

Having outlined several methods that are based on the gradient of a least squares fit in the observation space, we now discuss diffrent ways to obtain this gradient.

C.5.1 Finite differences

The most straithforward way of determining the gradient is by finite differences (FD), as this implies that a solution method must be developed only for the direct problem.

The formulas used for the gradient are

$$\frac{\partial \mathcal{J}(\boldsymbol{q})}{\partial p_i} \approx \frac{1}{h}[\mathcal{J}(\boldsymbol{q} + h\boldsymbol{e}_i) - \mathcal{J}(\boldsymbol{q})],$$

which is of first order in h, and

$$\frac{\partial \mathcal{J}(\boldsymbol{q})}{\partial p_i} \approx \frac{1}{2h}[\mathcal{J}(\boldsymbol{q} + h\boldsymbol{e}_i) - \mathcal{J}(\boldsymbol{q} - h\boldsymbol{e}_i)],$$

which is of second order in h.

Determining $\mathcal{J}(\boldsymbol{q})$ needs solving the direct problem in order to obtain $\mathcal{C}u(\boldsymbol{q})$. Thus, the first order FD needs one extra solution of the direct problem for every parameter p_i (M in total), and the second order FD needs two extra direct problems to be solved ($2M$ in total). When dealing with many parameters this is not efficient.

Note further that the step h cannot be arbitrarily small because of limits in the approximation of the direct problem. Here, roundoff errors also play a role.

C.5.2 Adjoint or costate methods

Adjoint methods can greatly reduce the cost of gradient evaluations. Here, the gradient is found by solving a PDE related to the direct problem, and performing several inner products. As an illustration, we give a formal presentation of the method for the steady state case of (C.2).

In the steady state case, we have that

$$A(\boldsymbol{q})u = f. \tag{C.19}$$

For example, in a diffusion setting where the diffusion coefficient is a constant parameter, we could consider $A(D) = -D\Delta(\cdot)$.

For simplicity assume that A is linear, invertible, and Fréchet differentiable. Differentiation of the identity

$$A(\boldsymbol{q})A(\boldsymbol{q})^{-1} = I,$$

yields

$$\frac{d}{d\boldsymbol{q}} A(\boldsymbol{q})^{-1} = -A(\boldsymbol{q})^{-1} \frac{dA}{d\boldsymbol{q}} A(\boldsymbol{q})^{-1}.$$

Assume further that the forward problem is well-posed, and denote its solution by

$$u = A(\boldsymbol{q})^{-1} f.$$

Introduce the *parameter-to-observation* map $F : \mathcal{Q} \to \mathcal{Z}$,

$$F(\boldsymbol{q}) = \mathcal{C} A(\boldsymbol{q})^{-1} f.$$

Setting $r(\boldsymbol{q}) = F(\boldsymbol{q}) - z$, (C.5) leads to

$$\begin{aligned}
[\operatorname{grad} \mathcal{J}(\boldsymbol{q})]_i &:= \frac{d}{dh} \mathcal{J}(\boldsymbol{q} + h\boldsymbol{e_i})\Big|_{h=0} \\
&= \left\langle \frac{d}{dh} F(\boldsymbol{q} + h\boldsymbol{e}_i)\Big|_{h=0}, r(\boldsymbol{q}) \right\rangle_{\mathcal{Z}} \\
&= -\left\langle \mathcal{C} A(\boldsymbol{q})^{-1} \left(\frac{dA}{d\boldsymbol{q}} \boldsymbol{e}_i \right) A(\boldsymbol{q})^{-1} f, r(\boldsymbol{q}) \right\rangle_{\mathcal{Z}} \\
&= -\left\langle \left(\frac{dA}{d\boldsymbol{q}} \boldsymbol{e}_i \right) A(\boldsymbol{q})^{-1} f, A^*(\boldsymbol{q})^{-1} \mathcal{C}^* r(\boldsymbol{q}) \right\rangle_H .
\end{aligned}$$

The last equality follows by taking Hilbert space adjoints.
Denote by v the solution to the costate or adjoint equation

$$A^*(\boldsymbol{q}) v = -\mathcal{C}^* r(\boldsymbol{q}), \tag{C.20}$$

then

$$[\operatorname{grad} \mathcal{J}(\boldsymbol{q})]_i = \left\langle \left(\frac{dA}{d\boldsymbol{q}} \boldsymbol{e}_i \right) u, v \right\rangle_H , \quad i = 1, \ldots, M. \tag{C.21}$$

Thus, the gradient is obtained by solving the adjoint equation (C.20) and performing M inner products (C.21). For further details and an example we refer to [74].

This deduction can also be developed for time-dependent problems such as (C.2), see [75]. This leads to a formulation of the adjoint equation with a final condition instead of an initial condition, as we encounter also in the Part II. The adjoint equation for the problems discussed in Part I are deduced in Part II by a more constructive approach, which is more suitable for specific applications than the formal deduction given here, see Section 4.1 and Section 5.3.

The application of the costate method to time dependent problems is in full development, and several authors obtained disappointing results. Quoting from [2] "... it was extremely difficult to obtain accurate search directions with gradients computed in this manner". Questions have thus been raised concerning the accuracy and convergence of costate approximations, even when the numerical methods being used are known to converge rapidly on the direct problem. In [75] it is shown that high order accuracy time-marching schemes on the forward problem do not necessarily lead to high order accurate costate approximations. Moreover, in some cases these approximations don't converge at all. However, these authors show that under certain circumstances, rapidly converging gradient approximations do follow. It is also shown that the numerical quadrature scheme, used for the inner products, plays an important role in the final accuracy of the gradient. Many of these problems come from the observation operator $\mathcal{C}$ involving pointwise evaluations in time.

In Part II, we deduce the adjoint equation method for the problems discussed in Part I and show that in these settings the method works well. We also prove convergence of the approximation to the adjoint equation in the case of the annealing problem.

List of Symbols

Abbreviations

a.e.	almost everywhere
rhs, lhs	right hand side, left hand side

Various symbols

$\mathbb{R}$	space of real values
f, u, V, F	scalar functions, denoted with normal font
$\boldsymbol{p}, \boldsymbol{q}$	vectors, denoted in bold
$\boldsymbol{D}$	diffusion matrix
$\boldsymbol{v}$	velocity field
$\mathcal{T}$ and $\mathcal{D}$	solution operators of a transport equation and a diffusion equation, respectively
$\mathcal{T}_\eta$ and $\mathcal{D}_\eta$	discrete solution operators, where η represents the discretization parameters
$[a, b]$	closed interval in $\mathbb{R}$
(a, b)	open interval in $\mathbb{R}$
Ω	open bounded domain with Lipshitz continuous boundary in $\mathbb{R}^n, n = 1, 2$ or 3
$\partial\Omega$	boundary of domain Ω

Γ	part of the boundary $\partial\Omega$
$\overline{\Gamma}$	closure of Γ
I	the time interval $(0,T)$, $T > 0$ fixed
Q_T	$= \Omega \times (0,T)$
$\boldsymbol{a} \cdot \boldsymbol{b}$	scalar product of the vectors $\boldsymbol{a}$ and $\boldsymbol{b}$
$(\cdot,\cdot)$ or $< \cdot,\cdot >$	standard inner product in the considered innerproduct space , e.g. $(\boldsymbol{a},\boldsymbol{b}) = \int_\Omega \boldsymbol{a} \cdot \boldsymbol{b}$ in $(L_2(\Omega))^n$
$\lvert \cdot \rvert$	$=$ the modulus in $\mathbb{R}$
$\lVert \cdot \rVert_X$	$=$ norm in the space X
$\lVert \cdot \rVert_p$	$=$ norm in the space $L_p(\Omega)$
$\lVert \cdot \rVert$	$=$ norm in the space $L_2(\Omega)$
$\partial_t f$	derivative of function f with respect to t
$\partial_x f$	partial derivative of function f with respect to x
f'	derivative of a function f on an interval with respect to its independent variable
$\dot{s}$	time derivative of a function $s(t)$
Δx	size of a grid cell along X-axis
$\boldsymbol{\nu}$ or $\boldsymbol{n}$	outward unit normal vector
C	generic positive constant
ε	arbitrary small positive constant
δ_{ij}	Kronecker delta function
$\rightarrow$	strong convergence in a normed space
$\rightharpoonup$	weak convergence in a normed space

Functions and operators

∇f Gradient of f, e.g. $(\partial_{x_1} f, \partial_{x_2} f, \partial_{x_3} f)$ in 3-D

$\nabla_p f$ Gradient of f with respect to $\boldsymbol{p} = (p_1, \dots, p_N)$, i.e. $(\partial_{p_1} f, \dots, \partial_{p_N} f)$

$\nabla \cdot \boldsymbol{v}$ Divergence of a vector function $\boldsymbol{v} = (v_1, v_2, v_3)$, i.e. $\nabla{\cdot}\boldsymbol{v}(x_1, x_2, x_3) = \partial_{x_1} v_1 + \partial_{x_2} v_2 + \partial_{x_3} v_3$

$\boldsymbol{u} \times \boldsymbol{v}$ vector product

Δf Laplace operator, $\nabla \cdot (\nabla f)$

π and $\widetilde{\pi}$ projection operators

TV total variation of a function

Function spaces

As above $\Omega \subset \mathbb{R}^n$, $(n = 1, 2$ or $3)$, is an open, bounded domain with Lipschitz-continuous boundary.

$C(\Omega)$ space of continuous functions on Ω

$C^k(\Omega)$ space of p times continuously differentiable functions on Ω, $k \geq 0$ integer

$C^\infty(\Omega)$ space of infinitely many times continuously differentiable functions on Ω.

$L_p(\Omega)$ space of the measurable p-th power Lesbesgue integrable functions $(p > 0$ finite$)$

$L_\infty(\Omega)$ space of measurable bounded functions on Ω (i.e. there exists a constant C such that $|f(x)| \leq C$ a.e. over Ω.)

$X_{\mathrm{loc}}(\Omega)$ The subspace of functions from a function space $X(\Omega)$ of which the support is contained in Ω (i.e. functions from $X(\Omega)$ which are zero in a neighbourhood of $\partial\Omega$.)

List of Figures

2.1 (i) Convex flux function $f(u)$ (ii) Two shocks with corresponding speeds . . . 25

2.2 Determination of the solution of the Riemann problem. (i) Flux function f and it's lower convex envelope $f_{\smile}$. (ii) The derivatives, where $\mathrm{v_s} = \frac{f(u_1)-f(u_2)}{u_1-u_2}$. (iii) The solution $(f'_{\smile})^{-1}(x/t)$, consisting of a shock and two rarefaction waves. . . . 29

2.3 Control Volume for the finite volume method in 2 dimensions. . . 31

2.4 Aquifer with dual-well. One recharge well and one pumping well. 33

2.5 Flow potential and flow lines for a dual well, modeled by a point source and sink. . . . 36

2.6 Boundary of the domain Ω in the (x, y) plane, and the domain $\widetilde{\Omega}$ in (u, v)-coordinates after bipolar transfomation. A is the injection well, B the extraction well. . . . 37

2.7 Flow equipotential lines for a dual well in uv-coordinates. Left: no regional flow. Right: regional flow aligned to dual-well flow, recirculation between the dual wells: flow from top to bottom. . 39

2.8 Flow equipotential lines for a dual well in uv-coordinates, regional flow against dual-well flow. Left: no recirculation between the dual wells, flow from top to infinity, and from infinity to bottom. Right: partial recirculation, part of top goes to bottom, and part goes to infinity. . . . 40

2.9 The (u, v) domain divided in strips and cells. . . . 47

2.10 Computational grid in Cartesian coordinates for equally spaced, 40×40, (u, v) nodal points. . . . 96

2.11 Importance of moving grid points in BTC obtained with the benchmark solution. Cut-out of BTC for $\alpha_L = 0.1$m, $\alpha_T = 0.0$m, pulse of 1 day. Top curve: 200 grid, no moving points; bottom curve: 800 grid with 20 moving gridpoints per shock. 100

2.12 BTC with step input for $\alpha_L/D = 0.2$, 0.1, 0.05, 0.02, 0.01, 0.005 and 0.002m scaled to match [76]. 101

2.13 BTC with pulse input for $\alpha_L/D = 0.2$, 0.1, 0.05, 0.02, 0.01, 0.005 and 0.002m scaled to match [76]. 102

2.14 BTC for 4 different head values. The top curve corresponds to confined flow and $\Delta h = 5$. The other curves correspond to partially confined-unconfined flow with $h_2 = 13$m, $h_1 = 8$m, and unconfined flow with $h_2 = 10$m, $h_1 = 5$m and $h_2 = 6$m, $h_1 = 1$m, respectively 102

2.15 Concentration levels for relative values $0.05, 0.10, \ldots, 0.95$ for the standard example with $\alpha_L = 0.5$m and $\alpha_T = 0$m, after 1, 3, 6 and 9 days of operation. 104

2.16 Concentration levels for relative values $0.05, 0.10, \ldots, 0.95$ for the standard example with $\alpha_L = 0.5$m and $\alpha_T = 0.01$m, after 1, 3, 6 and 9 days of operation. 105

2.17 Concentration levels for relative values $0.05, 0.10, \ldots, 0.95$ for the standard example with $\alpha_L = 0.05$m and $\alpha_T = 0.01$m, after 1, 3, 6 and 9 days of operation. 106

2.18 Concentration levels for relative values $0.05, 0.10, \ldots, 0.95$ for the standard scheme with $\alpha_L = 0.0$m and $\alpha_T = 0.0$m, after 1, 3, 6 and 9 days of operation. 107

2.19 BTCs of the tests in experiment 5, Fig. 2.15-2.18. 108

2.20 BTC with pulse input for $\alpha_L = 0.1$m, and $\alpha_T = 0$ (top curve), 0.02, 0.05, 0.1 and 0.2m (lowest curve). 109

2.21 Concentration levels for relative values $0.05, 0.10, \ldots, 0.95$ with method of lines after 1, 3, 6 and 9 days of operation. 110

2.22 BTC with step input for $\alpha_L/L = 0.2$, 0.1, 0.05, 0.02, 0.01, 0.005 and 0.002m. Left for Dirichlet inflow BC with operator splitting method. Right for advective flux inflow BC under OS. 113

2.23 BTC with pulse input of 1day for $\alpha_L/L = 0.2$, 0.1, 0.05, 0.02, 0.01, 0.005 and 0.002m. Left for Dirichlet inflow BC with operator splitting method. Right for advective flux inflow BC under OS. 114

2.24 Solution of equilibrium mode problem for $D = 0.001$, $K_0 = 1$, $p = 0.75$, with $\frac{\tau}{\Delta y} = 40$ fixed and with τ respectively being 5, 1, 0.1, 0.05, at final time $T = 15$. 115

2.25 Solution of an equilibrium mode problem with $K_0 = 1$, $p = 0.75$, at time $T = 12$, for 3 values of the diffusion: $D = 0.1$, 0.01 and 0.0001, each with 3 different discretizations: $\Delta y = 0.05$, $\tau = 1.5$ (dashed line) ; $\Delta y = 0.02$, $\tau = 0.1$ (dotted line); and $\Delta y = 0.002$, $\tau = 0.01$ (solid line). Also the analytical solution for $D = 0$ is given. 116

2.26 Dependence of the BTC for pulse input on the value of ρ with fixed $p = 0.9$ and $\alpha_L = 0.02$m. From left to right we have ρ equal to 0 (linear case), 0.01, 0.1 and 0.5. 118

2.27 Dependence of the BTC for pulse input on the value of p with fixed $\rho = 0.1$ and $\alpha_L = 0.02$m. From top to bottom we have p equal to 0.9, 0.75, 0.5 and 0.25. 118

3.1 Coating for FeSi 3wt%, fastly cooled. 129

3.2 Coating for FeSi 2.4wt%, slowly cooled. 129

3.3 Composition through the thickness of the coating for FeSi 3wt%, dipped during 5s. 130

3.4 Ternary phase diagram for Fe-Al-Si at 800°C, [53], and diffusion path at 800°C during 5s dipping, fastly cooled FeSi 3wt% substrate.131

3.5 Ternary phase diagram for Fe-Al-Si at 1100°C, [53], and diffusion path at 1100°C during 5min, FeSi 3wt% substrate. 132

3.6 Diffusion profiles for samples annealed at 1000°C during 30 min (top) and at 1100°C during 60 min(bottom). 133

3.7 Diffusion profiles of Si (at%) into steel (μm). A typical Si-profile after hot dipping is depicted in red. The lines in green, blue and purple show the diffusion after 1, 7 and 30 minutes, respectively for $D(C) = 0.4C^{1.59}$. An experimental Si-profile taken after 30 min of diffusion annealing is depicted in light blue. 134

3.8 Diffusion profiles of the Barenblatt-Pattle solution with $D(C) = 2.5C^{1.5}$. The initial profile is shown in red. The other curves depict profiles after 10, 60 and 120 sec, respectively, as modeled with a nonequidistant moving grid. 135

3.9 Diffusion profiles of the Barenblatt-Pattle solution with $D(C) = 2.5C^{1.5}$. The initial profile is shown in red. The other curves depict the concentration profiles after 10, 60 and 120 sec, respectively, obtained with an equidistant grid over the interval $(0, 13.5)$. For comparison, we also show the exact solution at $t = 120$sec (compare with Fig. 3.8). 136

3.10 Apparent diffusion coefficient at 1100°C (full) and Fe-Si interdiffusion coefficient (dashed), in m^2/s. 137

3.11 Experimental and modeled concentration profiles for 5 min, 30 min, 1h and 3h, respectively, at 1100°C. 138

4.1 a) Domain decomposition in two parts: $(0, L) = (0, s(t)) \cup (s(t), L)$. b) Mapping to the fixed uniform y interval. 154

4.2 Numerical approximation on the space grid. 155

4.3 The region around the edge $x = s(t)$ is left out in order to obtain a priori error estimates of the energy type. 164

4.4 Motivation for estimate (4.87) 167

4.5 Function $\chi^2(\boldsymbol{p})$ for $\hat{\boldsymbol{p}} = (2.5, 1.5)$. Left is the p_2-axis, right the p_1-axis. The contour lines are for the values 0.5, 1, 5, 10, 20 and 30. 179

5.1 BTC of pulse input for $\alpha_L = 0.1$m and the result of artificial noise with $\sigma = 0.001$ and $\sigma = 0.01$. 195

6.1 Area in Central Alberta with parallelism of creeks. 203

6.2 The domain. 205

6.3 The domain 207

6.4 The error on the top boundary in case of a Neumann BC for different N-values, with data $a/L = 0.05$, $V = 50$, $d = 0$, $L = 20000$, $T = 2000$. 215

6.5 Equipotential lines in case of a Neumann BC. Data in left part: $a/L = 0.05$, $V = 50$, $d = 0$, $n = 4$, $L = 20000$, $T = 2000$. Data in right part: $a/L = 0.02$, $V = 50$, $d = 0$, $n = 4$, $L = 20000$, $T = 1400$. 216

6.6 Equipotential lines in case of a Neumann BC with $a/L = 0.02$, $V = 50$, $n = 4$, $L = 20000$, $T = 1400$, $d = 0.00235$ (left) and $d = 0.0235$(right). 217

6.7 Equipotential lines in case of a Dirichlet BC at the base for $a/L = 0.02$, $V = 50$, $n = 4$, $L = 20000$, $T = 1400$. Left: $d = 0.00235$, $u = -50/L$, $v = -250$. Right: $d = 0.0$, $u = -50/L$, $v = -400$. . . 217
6.8 Topology elevation between two rivers. Top of elevation is at $z = 0$, $L = 5500m$, underlying aquifer at $z = -400m$. Equipotential lines for $h_{x=0} = -290m$, $h_{x=L} = -300m$, $N = 0.05/L$ and $d = 0.008$. 220
6.9 The mapping of the infinite rectangular elements. 221
6.10 The nodes on $\overline{\Gamma}_3$, and the connected semi-infinite rectangles. . . 222
6.11 Equipotential lines. Left: FEM in case of Neumann BC at the base, with data $L = 8000$, $a/L = 0.1$, $V = 80$, $d = 0.00235$, $T = 3000$. Right: IEM with the same data and $\alpha = -12000$. . . . 224

B.1 Darcy's experiment 234

List of Tables

2.1 RMS error of the benchmark solution, depending on the number of strips, the number of fixed y-divisions, the number of moving grid points per shock, and the operator splitting time step. . . . 98

2.2 RMS error of the standard scheme, depending on the number of strips, the number of cells and on the operator splitting time step. 99

2.3 RMS difference with increasing α_L of the benchmark scheme and of the standard scheme when compared with the benchmark scheme for $\alpha_L = 0$m. 99

2.4 Total mass inflow and outflow, total mass present, and mass balance, in the case of $\alpha_L = 2$ with step input. Left for Dirichlet inflow BC with operator splitting method. Right for advective flux inflow BC under operator splitting (OS). 113

2.5 Total mass inflow and outflow, total mass present, and mass balance, in the case of $\alpha_L = 0.1$ with step input. Left for Dirichlet inflow BC with operator splitting method. Right for advective flux inflow BC under OS. 113

2.6 Total mass inflow and outflow, total mass present, and mass balance, in the case of $\alpha_L = 2$ with pulse input of 1 day. Left for Dirichlet inflow BC with operator splitting method. Right for advective flux inflow BC under OS. 114

3.1 Theoretical and measured composition of the intermetallic compounds. 130

4.1 Values of $-\nabla_p \mathcal{F}$ and root-mean square error for several parametersets. 180

5.1 Successive estimated parameter values for the pulse input injection case, using only the extraction well averages, use of standard scheme. 193
5.2 Successive estimated parameter values for the pulse input injection case, using only the extraction well averages, use of benchmark scheme. 194
5.3 Successive estimated parameter values for the pulse input injection case, using the perturbed BTC ($\sigma = 0.01$), use of benchmark scheme. 195
5.4 Successive estimated parameter values for the pulse input injection case, using the perturbed extraction well averages ($\sigma = 0.001$) and the perturbed concentration values at an extra point (-5.6m, 0m) ($\sigma = 0.0002$). 196
5.5 Gradient by 3 different techniques 197
5.6 Levenberg-Marquardt iterations starting from 2 parameter sets . 198

B.1 Representative values of porosity θ_0. 232
B.2 Representative values of hydraulic conductivity k. 235

Bibliography

[1] Aster, R., Borchers, B., and Thurber, C. *Parameter Estimation and Inverse Problems.* Elsevier, Amsterdam, 2005.

[2] Banks, H., and Rosen, I. Numerical schemes for the estimation of functional parameters in distributed models for mixing mechanisms in lake and sea sediment cores. *Inverse Problems 3* (1987), 1–23.

[3] Barros, J., Malengier, B., Van Keer, R., and Houbaert, Y. Modeling silicon and aluminum diffusion in electrical steel. *Journal of phase equilibria and diffusion 26*, 5 (2005), 417–422.

[4] Barros, J., Ros-Yanez, T., Vandenbossche, L., Dupré, L., Melkebeek, J., and Y., H. The effect of Si and Al concentration gradients on the magnetic and mechanical properties of electrical steel. *Journal of Magnetism and Magnetic Materials 290-291* (2005), 1457–1460.

[5] Barros, J., Ros-Yanez, T., and Y., H. Chemical and physical interactions of Si rich steel substrates with a molten Al-25 wt%Si bath. *Defect and Diffusion Forum 237-240* (2005), 1115–1120.

[6] Batens, N. *Bijdrage tot numerieke benaderingsmethoden en existentietheorie voor vrije grensproblemen uit de chemische kinetiek.* PhD thesis, Ghent University, 2002.

[7] Bear, J. *Dynamics of Fluid in Porous Media.* Elsevier, New York, 1972.

[8] Bear, J., and Bachmat, Y. *Introduction to Modeling of Transport Phenomena.* Kluwer, Dordrecht, 1991.

[9] Bedient, P. B., Rifai, H. S., and Newell, C. *Ground Water Contamination: Transort and Remediation.* Prentice Hall, New Jersey, 1994.

[10] Bettess, P. *Infinite Elements.* Penshaw Press, Sunderland, 1992.

[11] Brézis, H. *Analyse fonctonnelle. Théorie et applications.* Dunod, Paris, 1999.

[12] Bürger, R., Frid, H., and Karlsen, K. On the well-posedness of entropy solutions to conservation laws with a zero-flux boundary condition. *J. Math. Anal. Appl.* (2006). [to appear].

[13] Constales, D. Solution of the dual-well by an upwind method. Private communication, 2002.

[14] Constales, D., and Kačur, J. Determination of soil parameters via the solution of inverse problems in infiltration. *Computational Geosciences 5* (2004), 25–46.

[15] Constales, D., Kačur, J., and Malengier, B. A precise numerical scheme for contaminant transport in dual-well flow. *Water Resources Research 39*, 30 (2003), 1303–1315.

[16] Crandall, M. G., and Majda, A. The method of fractional steps for conservation laws. *Numer. Math. 34* (1980), 285–314.

[17] Curtain, R., and Pritchard, A. *Functional analysis in modern applied mathematics*, vol. 132 of *Mathematics in science and engineering.* Academic Press, London, 1977.

[18] Dagan, G. *Flow and Transport in Porous Formations.* Springer-Verlag, Berlin, 1989.

[19] Darcy, H. *Les Fountaines Publiques de la Ville de Dijon.* Victor Dalmont, Paris, 1856.

[20] Gelhar, L. W., and Collins, M. A. General analysis of longitudinal dispersion in nonuniform flow. *Water Resour. Res. 7*, 6 (1971), 1511–1521.

[21] Glicksman, M. *Diffusion in Solids, Field theory, Solid-state principles, and Applications.* Wiley, New York, 2000.

[22] Gradshteyn, I. S., and Ryzhik, I. *Table of Integrals, Series, and Products.* Academic Press, New York, 1980.

[23] Grove, D. B. An analysis of the flow field of a discharging recharging pair of wells. *U.S. Geol. Surv. Rep. (USGS) 474* (1971).

[24] Haitjema, H. M. *Analytic Element Modeling of Groundwater Flow.* Academic Press, San Diego, 1995.

[25] Holden, H., Karlsen, K., and Lie, K.-A. Operator splitting methods for degenerate convection-diffusion equations i: Convergence and entropy estimates. In *Stochastic processes, physics and geometry: new interplays, II, CMS Conf. Proc., (Leipzig, 1999)* (RI, 2000), F. Gesztesy, H. Holden, J. Jost, S. Paycha, M. Röckner, and S. Scarlatti, Eds., vol. 29, Providence, pp. 293–316.

[26] Holden, H., and Risebro, N. H. A method of fractional steps for scalar conservation laws without the cfl condition. *Mathematics of Computation 60*, 201 (1993), 221–232.

[27] Holden, H., and Risebro, N. H. *Front tracking for hyperbolic conservation laws.* Springer-Verlag, New York, 2002.

[28] Hoopes, J. A., and Harleman, D. R. F. Dispersion in radial flow from a recharge well. *J. Geophys. Res. 72*, 14 (1967), 3595–3607.

[29] Hundsdorfer, W., and Verwer, J. *Numerical solution of time-dependent advection-diffusion-reaction equations.* Springer Series in Computational Mathematics. Springer-Verlag, Berlin, 2003.

[30] Jäger, W., and Kačur, J. Solution of doubly nonlinear and degenerate parabolic problems by relaxation scheme. *M2AN 29* (1995), 605–627.

[31] Karlsen, K. H., and Lie, K.-A. An unconditionally stable splitting for a class of nonlinear parabolic equations. *IMA J. Numer. Anal. 19*, 4 (1999), 609–635.

[32] Karlsen, K. H., Lie, K.-A., and Risebro, N. A front tracking method for conservation laws with boundary conditions. In *Hyperbolic problems: theory, numerics, applications (Seventh international conference in Zurich, 1998)* (1999), M. Fey and R. Jeltsch, Eds., vol. 192 of *Int. Series of Numerical Mathematics*, Birkhaüser, pp. 493–502.

[33] Karlsen, K. H., and Risebro, N. An operator splitting method for nonlinear convection-diffusion equations. *Numer. Math. 77*, 3 (1997), 365–382.

[34] Karlsen, K. H., and Risebro, N. On the uniqueness and stability of entropy solutions of nonlinear degenerate parabolic equations with rough coefficients. *IMA J. Numer. Anal. 19*, 4 (1999), 609–635.

[35] Kačur, J. Solution to strongly nonlinear parabolic problems by a linear approximation scheme. *IMA J. Num. Anal. 19* (1999), 119–154.

[36] Kačur, J., and Babušíková, J. Determination of model parameters for contaminant transport in dual-well setting. *IASME Transactions 2*, 6 (2005).

[37] Kačur, J., and Frolkovič, P. Semi-analytical solutions for contaminant transport with nonlinear sorption in 1D. *University of Heidelberg Preprint 24*, SFB 359 (2002), 1–20.

[38] Kačur, J., and Luckhaus, S. Approximation of degenerate parabolic systems by nondegenerate elliptic and parabolic systems. *Appl. Numer. Math. 26* (1998), 307–326.

[39] Kačur, J., Malengier, B., and Remešíková, M. Solution of contaminant transport with equilibrium and non-equilibrium adsorption. *Computer Methods in Applied Mechanics and Engineering 194*, 2-5 (2005), 479–489.

[40] Kačur, J., Malengier, B., and Van Keer, R. Determination of the diffusion annealing process of Si into Fe. In *European Congress on Computational Methods in Applied Sciences and Engineering, ECCOMAS 2004* (2004), P. Neittaanmäki, T. Rossi, K. Majava, and O. Pironneau, Eds.

[41] Kačur, J., Remešíková, M., and Malengier, B. Contaminant transport with adsorption and their inverse problems. *Computing and visualization in Science* (2006). [to appear].

[42] Kačur, J., and Van Keer, R. Solution of contaminant transport with adsorption in porous media by the method of characteristics. *M2AN 35*, 5 (2001), 981–1006.

[43] Knabner, P., and Angermann, L. *Numerical methods for elliptic and parabolic partial differential equations.* Texts in applied mathematics. Springer-Verlag, New York, 2003.

[44] Knabner, P., and Otto, F. Solute tansport in porous media with equilibrium and nonequilibrium multiple-site adsorption. *Nonlinear Analysis 42*, 3 (2000), 381–403.

[45] Knabner, P., and van Duijn, C. Solute tansport in porous media with equilibrium and non-equilibrium multiple-site adsorption: Traveling waves. *Jounal für die reine und angewandte Mathematic 415* (1991), 1–49.

[46] Kružkov, S. First order quasi linear equations in several independent variables. *Math. USSR Sbornik 10*, 2 (1970), 217–243.

[47] Le Veque, R. J. *Numerical methods for conservation laws.* Birkhäuser, Basel, 1992.

[48] Le Veque, R. J. *Finite volume methods for hyperbolic problems.* Cambridge texts in applied mathematics. Cambridge University Press, Cambridge, 2002.

[49] Lee, T.-C. *Applied Mathematics in Hydrogeology.* Lewis Publishers, Boca Raton (US), 1998.

[50] Li, R., Chen, Z., and Wu, W. *Generalized difference methods for differential equations: numerical analysis of finite volume methods.* Pure and applied mathematics. Marcel Dekker, Inc., New York, 2000.

[51] Lie, K.-A. Front tracking for one-dimensional quasilinear hyperbolic equations with variable coefficients. *Numerical Algorithms 24*, 3 (2000), 275–298.

[52] Lockwood, E. H. *Bipolar Coordinates.* Cambridge University Press, Cambridge, England, 1967, ch. 25, pp. 186–190.

[53] Maitra, T., and Gupta, S. Intermetallic compound formation in fe-al-si ternary system. *Materials Characterization 49*, 4 (2002), 269–311.

[54] Malengier, B. Aspecten van de numerieke modellering van stromingsvraagstukken in poreuze media. Master's thesis, Vrije Universiteit Brussel, 2001. [Thesis for the Master Degree in Applied Informatics].

[55] Malengier, B. Parameter identification in stationary groundwater flow problems in drainage basins. *Journal of Computational and Applied Mathematics (JCAM) 168* (2004), 299–307.

[56] Malengier, B. Parameter estimation of Si diffusion in Fe substrates after hot dipping and diffusion annealing. In *Numerical Analysis and Its Applications 2004* (2005), Z. Li, L. Vulkov, and J. Wasniewski, Eds., vol. 3401 of *Lecture Notes in Computer Science*, Springer, pp. 399–407.

[57] Malengier, B., and Van Keer, R. An analysis of groundwater flow in an finite region with a sinusoidal top. *Numer. Funct. Anal. and Optimiz. 23*, 5&6 (2002), 589–607.

[58] Malengier, B., and Van Keer, R. Parameter estimation in convection dominated nonlinear convection-diffusion problems by the relaxation method and the adjoint equation. *Journal of Computational and Applied Mathematics (JCAM)* (2006). [accepted].

[59] Murakami, K., Nishida, N., Osamura, K., and Tomota, Y. Aluminization of high purity iron by powder liquid coating. *Acta Materialia 52* (2004), 1271–1281.

[60] Murakami, K., Nishida, N., Osamura, K., Tomota, Y., and Suzuki, T. Aluminization of high purity iron and stainless steel by powder liquid coating. *Acta Materialia 52* (2004), 2173–2184.

[61] Oleĭnik, O., and Kružkov, S. Quasi-linea second-order parabolic equations with many independent variables. *Russian Math. Surveys 16*, 5 (1961), 105–146.

[62] Patankar, S. V. *Numerical Heat Transfer and Fluid Flow.* Taylor & Francis, 1980.

[63] Pickens, J. F., and Grisak, G. E. Scale-dependent dispersion in a stratified granular aquifer. *Water Resour. Res. 17*, 4 (1981), 1191–1211.

[64] Polak, E. *Optimization: algorithms and consistent approximations.* Springer, 1997.

[65] Polyanin, A., and Zaitsev, V. *Handbook of Nonlinear partial differential equations.* Chapman&Hall/CRC, 2004.

[66] Rabkin, E., and et al. The influence of an ordering transition on the interdiffusion in fe-si alloys. *Acta metall. mater 43*, 8 (1995), 3075–3083.

[67] Ruiz, D., Ros-Yáñez, T., Vandenberghe, R., and Y., H. On the influence of order on the soft magnetic properties of Fe-Si alloys. *Steel Research International 76*, 6 (2005), 429–435.

[68] Shivakumar, P. N., Williams, J. J., Ye, Q., and Chuanxiang, J. An analysis of groundwater flow in an infinite region with sinusoidal top. *Numer. Funct. Anal. and Optimiz. 21*, 1&2 (2000), 263–271.

[69] SUN, N.-Z. *Mathematical modeling of groundwater pollution.* Springer-Verlag, New York, 1996.

[70] TEUKOLSKY, S. A., AND ET AL. *Numerical recipes in C: the art of scientific computing.* Springer-Verlag, Cambridge, 1992.

[71] THOMAS, J. W. *Numerical partial differential equations: Conservation laws and elliptic equations.* Texts in Applied Mathematics 33. Springer, New York, 1999.

[72] TÒTH, J. A theory of groundwater motion in small drainage basins in central alberta, canada. *J. Geophys. Res. 67* (1962), 4375–4381.

[73] TÒTH, J. A theoretical analysis of groundwater flow in small drainage basins. *J. Geophys. Res. 67* (1963), 4795–4812.

[74] VOGEL, C. R. *Computational methods for inverse problems.* Frontiers in applied mathematics. Siam, USA, 2002.

[75] VOGEL, C. R., AND WADE, J. G. Analysis of costate discretizations in parameter identification for linear evolution equations. *SIAM Journal on Control and Optimization 33* (1995), 227–254.

[76] WELLY, C., AND GELHAR, L. W. Evaluation of longitudinal dispersivity from nonuniform flow tracer tests. *Journal of Hydrology 153* (1971), 71–102.

[77] YOSIDA, K. *Functional Analysis.* Springer-Verlag, 1968.

[78] ZIENKIEWICZ, O. C., AND MORGAN, K. *Finite Elements and Approximation.* John Wiley & Sons, New York, 1983.