Some Aspects of
Mixed Finite Element Methods
for
Semiconductor Simulation

Some Aspects of Mixed Finite Element Methods for Semiconductor Simulation

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam,
op gezag van de Rector Magnificus
prof. dr. P. W. M. de Meijer
in het openbaar te verdedigen in de Aula der Universiteit
(Oude Lutherse Kerk, ingang Singel 411, hoek Spui),
op vrijdag 8 mei 1992 te 13.00 uur
door

Ronald Robert Paul van Nooyen

Geboren te Rotterdam

Faculteit der Wiskunde en Informatica.
Promotor: Prof. dr. P. W. Hemker.

Het hier beschreven onderzoek werd mede mogelijk gemaakt door steun van het innovatiegericht onderzoekprogramma IC-Technologie.

From G.J. and R.R.P.
to the memory of A.H.

Acknowledgements

I thank Prof. Dr. P. W. H. Hemker for leaving me considerable freedom to do research, but still giving me the benefit of his considerable experience when it came to presenting my results. I am grateful to Dr. Ir. H. C. de Graaff for his excellent help in formulating chapter one. Both Dr. Ir. de Graaff and Prof. Dr. Memelink should be thanked for their willingness to explain their insights into the workings of semiconductors to our group at the CWI and their efforts on behalf of the IOP-IC technology projects in general. I am grateful to my colleagues Hans Molenaar and Paul de Zeeuw for the many fruitful discussions we had and the pleasant working environment. I thank Walter Lioen and Dik Winter for their help in coming to grips with UNIX and ditroff. To all other members of the department of Numerical Mathematics at the CWI, whose doors were always open for a ″promovendus″ searching for answers, my thanks. I thank the CWI computer support group and the CWI printing shop for their help in preparing and printing this thesis. Last, but not least I thank my family and friends for their support and their tolerance of my preoccupation with this thesis.

Contents

Chapter 1: Introduction 11

1.1 Semiconductors, numerical mathematics and society 11

1.2 A short history of semi-conductor devices 12

1.3 The importance of semi-conductor modelling 12

1.4 The special importance of semiconductors for numerical mathematics 13

1.5 A short introduction to semiconductors 14

Homogeneous material 14

One junction: the diode 16

Two junctions: the transistor 16

The role of silicon oxide: the MOSFET 17

1.6 Description of the semi-conductor model to be used 17

Scaling the equations 18

The choice of variables 19

The Scharfetter-Gummel discretisation and the Scharfetter-Gummel iteration 20

1.7 Contents of this thesis 21

References for chapter 1 24

Chapter 2: An Improved Accuracy Version of the Mixed Finite Element Method for a Second Order Elliptic Equation 28

2.1 Introduction 28

2.2 The equation 29

2.3 The discretisation 29

2.3.1 The partitioning of the domain 29

2.3.2 The approximation spaces 30

2.3.3 The discretisation scheme 31

2.3.4 The definition of α_h 32

2.4 Other quadrature rules 34

2.4.1 Exact evaluation of the form α on test and trial functions 34

2.4.2 Use of the trapezoidal rule 34
2.5 An error estimate 35
2.5.1 Error estimates for integration formulas 35
2.5.2 A special norm on V_h 39
2.5.3 The error estimate for the modified method 41
2.5.4 A proof of condition(C3) on a uniform mesh with constant a 45
2.6 The effect of a non-zero c 47
2.7 Numerical experiments 49
2.8 An a-posteriori error estimate 50
2.9 Conclusions 51
References for chapter 2 52
Chapter 3: The One Dimensional Convection-Diffusion Equation 54
3.1 Introduction 54
3.2 Regularity of the problem 55
3.2.1 Facts on integration and differentiation of Lebesgue integrable functions 56
3.2.2 The derivation of expressions for the solution and the flux 59
3.2.3 Upper bounds on the norms of the solution and the flux 60
3.3 The adjoint problem 61
3.4 The discretisation 63
3.4.1 The projections onto the trial spaces for the solution and its flux 63
3.4.2 An example of a set of trial spaces 64
3.4.3 The discrete scheme 64
3.5 Properties of the projections 65
3.6 A priori error estimates 67
3.7 Conclusions 70
References for chapter 3 70
Chapter 4: A Finite Volume Discretisation Scheme with a-Posteriori Error Estimates for the Symmetrised Continuity Equation 72
4.1 Introduction 72
4.2 The model equation 73
4.3 The discretisation spaces 74
4.3.1 The partition 75
4.3.2 Local coordinates 77
4.3.3 Some local projections 77

4.3.3 Some global projections and the trial spaces 78
4.4 Discretisation of the system 79
4.5 Existence and uniqueness of the solution 82
4.6 The quadrature rule 86
4.7 Consistency 87
4.7.1 Effects of piecewise bilinear interpolation for ψ 88
4.7.2 The discretisation error 89
4.8 An a posteriori error estimator 92
4.8.1 A derivation of a deferred correction scheme 92
4.8.2 Numerical results 98
4.9 Conclusions 100
References for chapter 4 100
Chapter 5: A Petrov-Galerkin Mixed Finite Element Method with Exponential Fitting 102
5.1 Introduction 102
5.2 The equation 102
5.3 Preparations 104
5.3.1 The partitioning of the domain 104
5.3.2 The adjoint problem 104
5.3.3 Some projections 105
5.3.4 Error estimates for projections 106
5.4 The discretisation 109
5.4.1 The quadrature rule 109
5.4.2 The discrete system 110
5.5 Several technical results 111
5.5.1 The properties of $\overline{a}_q$ 112
5.5.2 The difference between $\overline{a}$ and $\overline{a}_q$ 114
5.6 The error estimates 115
5.6.1 Errors due to approximation of the bilinear forms 115
5.6.2 An a priori error estimate 116
5.6.3 A different approach 121
5.7 A verification of the local maximum principle 123
5.8 An a-posteriori estimator 123
5.8.1 The derivation of the quadrature rule 124
5.8.2 An estimator for the local discretisation error and a lower bound for the global error 124

5.9 Numerical results 125
5.10 Conclusions 126
References for chapter 5 126
Samenvatting (abstract in Dutch) 128
List of Symbols 129
Name Index 130
Subject Index 132

1. Introduction

1.1 Semiconductors, numerical mathematics and society.

The role of the semiconductor device in modern society is fundamental. This view is supported by the following quote from the proceedings of the IEEE, "These products [semiconductor devices], by leveraging man's mental capabilities, have the potential for causing an impact on society exceeding that of the Industrial Revolution, which leveraged man's physical capabilities." [1] This view was expressed somewhat mcre prosaically by Steve Jobs, one of the founders of Apple Computer Inc., who, after reading an article in the Scientific American that showed how the bicycle gave man the most energy efficient mode of transportation in the animal kingdom [2], likened computers to "bicycles for the mind." [3] It is of course natural that man wants bigger "levers" and faster "bicycles". To achieve this it is necessary to improve not only our understanding of semiconductors, but also to predict the behaviour of semiconductor devices not yet built. For this we need a model of a semiconductor and a method to obtain predictions of semiconductor behaviour from this model. Numerical approximation techniques are a way to obtain such a method. The simplest semiconductor model is the stationary drift diffusion model. This thesis deals with the search for a-posteriori error estimates for discrete equations contained in a numerical approximation to this model. It is our intention in this first chapter to place in perspective the problem of obtaining a-priori and a-posteriori error estimates for the equations found in the drift-diffusion model. Section 1.2 gives a brief history of the development of semiconductor devices. In Section 1.3 and 1.4 we discuss the importance of numerical modelling and the influence of semiconductor devices, as building blocks for computers, on numerical mathematics. Section 1.5 is the only part of this thesis where physical aspects of semiconductor devices appear. It is included to give the reader an impression of the phenomena the mathematical model attempts to describe. In section 1.6 we describe the drift-diffusion model and discuss how these equations may be scaled. Section 1.6 also includes several choices for variable transformations that result in different forms for the model equations. Finally this section describes the one-dimensional Scharfetter-Gummel discretisation for the continuity equation. Section 1.7 gives a motivation for the research in this thesis and an outline of its contents.

1.2 A short history of semiconductor devices.

The development of semiconductor devices can be divided into three periods. In the first period researchers discovered the special properties of semiconductors and primitive applications appear. The second period starts with the creation of the first transistor and sees rapid development and commercial application of new devices, each equivalent to one circuit element and separately packaged. The third period - in which we are now - is the period of the integrated circuit. The devices developed earlier and new devices specifically developed for the IC environment are now constructed and linked together in one piece of silicon called a chip. According to J. M. Early [4], the first period of semiconductor development lies between 1800 and 1947. As a first consumer application from that era Petritz [5] mentions the crystal rectifier. He describes this as an arrangement of a piece of semiconductor crystal clamped or soldered in a receptacle and a flexible wire, a "cat's whisker", held in light contact with the crystal. This was used as the detector in early radio receivers. According to the 1962 paper by J. M. Early, the second period starts with the discovery of the transistor in 1947. The 1980 paper by S. M. Sze [6] places the start of the integrated circuit era in the year 1959. However, the cautious statement on integrated circuits at the end of the 1962 paper by Petritz [5], "The last category [i.e. the full development of the integrated circuit concept] will place the most stringent demands of all upon the materials technologist" suggests that in 1962 the concept of integrated circuits was still new and experimental.

The era of the integrated circuit deserves a closer examination. The paper by Sze gives graphs of technological development that show exponential decrease of the size of device components and exponential increase of the number of components per chip in the years between 1960 and 1980. He predicts a slowdown for size decrease and component number increase but expects the feature size to drop below one micrometer and the number of components per chip to exceed one million by 1990. These predictions are confirmed by the 1986 paper by Meyers et al. [7] and the 1990 paper by Fair [8]. The commercial effects of this growth are discussed by Jones et al. [1] This last paper shows the enormous increase in computational power available for a given price.

1.3 The importance of semiconductor modelling.

As shown in section 1.2, the use of semiconductors such as (doped) silicon for the construction of electronic devices is a fairly recent development and saw rapid evolution from single transistors through the first integrated circuits containing a few devices to the Very Large Scale Integrated circuits in use today. The development is rapid, a chip that is a masterpiece of chip technology today may seem commonplace in four or five years time. Modern Computer Aided Design and Manufacturing systems built from VLSI-components play an important role in this development. More and more the capacity for computation that VLSI-chips give to modern CAD-CAM systems acts as a catalyst for the further development of these chips. Due to the expense in both time and money of the fabrication of a prototype circuit, fast and accurate models of semiconductor devices are essential. Moreover very large scale integration makes two-dimensional models essential and three-dimensional models highly desirable. Designers of numerical simulations face the following difficult problem. From certain aspects of the devices, e.g. the presence of extremely sharp transitions in impurity concentrations, it follows that, at least locally, a very fine mesh may be needed to achieve any accuracy at all. Faced with these demands, numerical analysts have two tasks. On the one hand they have to create algorithms that solve the model equations supplied by the physicist as efficiently as possible, e.g. by local mesh refinement, and on the other hand to give a measure for the accuracy of the solution obtained by such an algorithm, either by a-priori bounds or by a-posteriori bounds. Good control on the accuracy is essential not only because the nature of the equations makes the accuracy extremely dependent on the method of discretisation but also because good control on accuracy helps us to avoid unnecessary computations. In this thesis we restrict ourselves to the study of one of these two aspects, namely error estimation. For more information on the fascinating field of semiconductor physics and engineering, I refer to the books and papers mentioned below.

Good references for semiconductor and device physics are the books by Blakemore [9], Ashcroft and Mermin [10], and Wang [11]. For information on the mathematical modelling of semiconductors, one may consult the books by Markowich [12] and Selberherr [13]. Review papers on mathematical modelling of semiconductors can be separated into physically oriented, e.g. Engl et al. [14] and Polak et al. [15] and more mathematically oriented papers such as the papers by Bank et al. [16, 17]. There are many research papers dealing with this subject [18-30]. The general trend for such papers is to attempt to give a useful mathematical derivation for the Scharfetter-Gummel discretisation in two or three dimensions. The derivations use almost all possible techniques from finite volumes to hybrid mixed finite element schemes.

1.4 The special importance of semiconductors for numerical mathematics.

We have already mentioned the need for practical numerical mathematics in the semiconductor industry. At this point it is educational to look at the inverse relationship. The explosive growth of both computer memory and computational power available for a given price make possible the utilisation of algorithms that seemed otherwise doomed to an existence as mathematical curiosities. The semiconductor industry gave numerical mathematicians new and hitherto undreamed of tools for research. However, the development is not all positive. There is a tendency to use simple algorithms and make up for their deficiencies by the use of large amounts of computing power. In principle there is nothing wrong with taking the simplest possible approach and letting the computer do the hard work. But in practice the need to run more and more complicated simulations inevitably outstrips the increases in speed provided by hard and software development. So the availability of relatively cheap computing resources may slow down the development of efficient algorithms needed for more complicated models and may result in a waste of computing resources.

1.5 A short introduction to semiconductors.

First, we give a description of homogeneous undoped and doped semiconductors, then of a junction diode and a junction transistor, and finally of a MOSFET, a Metal Oxide Semiconductor Field Effect Transistor. We start by explaining the name semiconductor. With regard to conduction of electrical current, a substance that forms an ordered crystal lattice can fall into two obvious categories. Either it conducts or it does not conduct electrical current. In the first case we call it a metal and in the second case an insulator. However, when we perform experiments, it turns out that instead of just these two categories, there exists a whole spectrum of substances with different conductivity behaviour. We are concerned with the substances whose behaviour earned them the name semiconductors. At low temperature, they are fairly good insulators, but as the temperature increases, their conductivity improves. Note that the boundary between the categories is not very sharply defined. This section gives a - much simplified - version of the explanations given in the books by Blakemore [9], Ashcroft and Mermin [10], and Wang [11]. All errors or omissions are my own.

Homogeneous material.

In this section we give a sketch of the band model for solids. A model of a physical system is judged by its capability to correctly predict the qualitative and quantitative behaviour of the system it models. It can never be proved to be correct in the sense that a mathematical theorem can be proved to hold. Please keep this in mind when reading this section. The above mentioned band model is very successful in explaining the behaviour of metals and crystalline semiconductors and insulators. The model is based on the Pauli exclusion principle and the structure of the energy spectrum of the Schrödinger operator for a periodic potential. We discuss this below.

As electrons are fermions, i.e. particles with quantum-mechanical spin ½, they are subject to the Pauli exclusion principle, i.e. no two otherwise indistinguishable electrons may be in the same quantum state. This principle is equivalent to the rule that exchanging two fermions in a system should change the phase of the wave function of the system by π. Moreover, for the Schrödinger wave equation in a finite volume with periodic potential and boundary conditions, the set of real, positive (i.e. physically possible) energy eigenvalues is discrete and displays a band structure, between bands of closely spaced allowed energy levels there may be intervals without allowed energies, these forbidden zones are called bandgaps. From quantum mechanics it follows that for each energy level in the spectrum, there is a finite number of eigensolutions, i.e. electron quantum states, of the Schr̈odinger wave equation.

The exclusion principle implies that the electrons cannot all be in the same quantum state. The three above cases can now be described as follows. If we are at an absolute temperature of zero degrees Kelvin and fill the states associated with the energy levels from the lowest level upwards, as indicated by minimisation of total energy, we find a highest occupied level and an associated energy. We call this the Fermi-energy.

If the Fermi-energy does not correspond with the start of a band gap, then the substance in question is a metal. If the Fermi-energy does correspond with the start of a band gap, then the substance in question is a semiconductor or an insulator. In that case, we call the the highest filled band the valence band and the next empty band the conduction band. We see that in a metal relatively small changes in velocity, hence energy, are possible, because there are empty energy levels close to the Fermi-energy. These changes make electrical conduction possible. Note that exchanging two electrons in a filled band does not lead to conduction because the new state differs from the old only in phase. In an insulator or a semiconductor the valence band and the conduction band are separated by a gap of forbidden energies. To conduct electricity, we need electrons in the conduction band. Thermal excitation can supply such electrons, but the availability depends exponentially on the ratio between the available thermal energy $k_B T$ and the width of the band gap. Note that k_B is the Boltzmann energy and T is the absolute temperature in Kelvin. So the larger the bandgap in relation to the thermal energy $k_B T$,the better the material insulates. In a semiconductor the gap is of the order of $10k_B T$ to $100k_B T$, this allows the material to be a - very bad - conductor in stead of an insulator. In a semiconductor the thermal excitations that displace some electrons to the conduction band where they can take part in electrical conduction create free states in the valence band that behave as particles with positive charge - called holes - that can also take part in conduction.

The conductivity can be changed by the addition of donor or acceptor centres. Donor centres are sites in the crystal where there is an extra energy level in the forbidden gap close to the conduction band that in the case of electrical neutrality contains an electron. Due to its closeness to the conduction band the electron can easily enter that band, i.e. the donor site can "donate" that electron to the conduction band. So, in a piece of material with added donors, there will be more electrons than holes available for conduction. Acceptor centres are sites in the crystal where there is an extra energy level in the intrinsic gap close to the valence band that is vacant when the site is electrically neutral. Due to its closeness to the valence band an electron can easily enter the site, i.e. the donor site can "accept" that electron from the valence band and so create a hole that can take part in conduction. So, in a piece of material with added acceptors, there will be more holes than electrons available for conduction. The dominant charge carrier in a material is called the majority carrier.

One junction: the diode.

To create the simplest silicon semiconductor device, a diode, which allows significant current flow in one direction only, we use a junction between a region of silicon with an additive that acts to provide acceptors (p-type) and a region with an additive that acts to provide donors (n-type). A crude description of its mode of operation is the following. If we connect the diode to a voltage source and we apply the higher voltage to the (n-type) material - this is called reverse bias - then this depletes the region around the junction, i.e. it pulls all free charge carriers away from the junction and so turns the region around the junction into an insulator, only a very small constant leakage current flows through the device. Note that beyond a certain voltage this no longer holds and electrical breakdown may occur. If we reverse the connections we get a forward bias, the behaviour is now more complicated, it turns out that the current increases exponentially with increasing voltage. Again above a certain level this no longer holds and damage to the device may occur because of internal heating.

Two junctions: the transistor.

The next simplest device is the transistor, which can act as a switch or as an amplifier. The simplest way to construct one is to use a succession of two junctions between different dopants, n-p-n or p-n-p. Here we take an (n-type)-(p-type) junction followed closely by a (p-type)-(n-type) junction, i.e. an n-p-n transistor. Such a transistor resembles two diodes in series with opposite reverse bias directions. The strong point of the transistor is that a very small current applied to the (p-type) material can switch a large current between the (n-type) regions on and off. We make three connections to this structure, one to each of the n-regions and one to the p-region. The p-region is called the base. The names of the n-regions are based on the intended direction of the current flow in normal operation. The direction of the flow of current determines the direction of flow of majority carriers in the n-regions - i.e. electrons -, the n-regions are named in such a way that this carrier flow goes from the emitter to the collector. Usually there is an asymmetry in the doping profile corresponding to the intended direction of current flow.

We give a crude description of its operation in its normal mode, i.e. the range of base, emitter and collector voltages where the collector voltage is larger than the emitter voltage and at least one of the two diodes is reverse biased. As long as we apply a voltage to the base, that is lower than either one of the voltages applied to the collector and emitter, we have two diodes in reverse bias with opposite bias directions, so hardly any current can flow between collector and emitter. If we increase the base voltage until it is just above the emitter voltage, we get two current flows, a base current from base to emitter and a much larger flow from collector to emitter. Both currents depend exponentially on the voltage difference between base and emitter. For more details I refer to part 6 of the paper by W. Shockley [31] or Section 1 of Chapter 9 of the book by Wang [11].

The role of silicon oxide: the MOSFET.

Of course the method used in the n-p-n junction transistor is not the only way to create a transistor. When we think of capacitors another way to create a switch comes to mind, namely the method used in the MOSFET, the Metal Oxide Semiconductor Field Effect Transistor. The MOSFET is especially useful in integrated circuits. It consists of a block of n-type material with two separate, embedded regions of p-type material adjacent to a surface. On the surface between the p-type areas, we create a layer of an insulator, for silicon this is often silicon oxide. On the two p-regions, the oxide layer and the opposite surface of the n-type block, we make metal contacts. The connection on the isolating layer is called the gate, the connection on the n-type material supplies the bulk bias and the connectors on the p-type material are called source and drain. Holes flow from source to drain in normal operation. When the gate is at zero volt there are too few holes to allow an appreciable current flow from source to drain. If we apply a suitable negative voltage to the gate a layer of holes is pulled to the surface. This improves the conductivity of the channel - i.e. the region beneath the gate - and the source-drain current increases. For a description of a MOSFET I refer to section 10 of chapter 9 of the book by Wang. [11]

1.6 Description of the semiconductor model used.

A derivation of the drift-diffusion model is given by e.g. Selberherr [13]. We discuss possible choices of variables and give a scaling. A more elaborate derivation of alternative forms for the model equations can be found in e.g. Polak, den Heijer, Schilders and Markowich [15]. We use the term doping to indicate donor or acceptor centres in the material. The distribution of these impurities is given by position-dependent functions N_d for the donor impurities and N_a for acceptor impurities. The value of the function gives the density of the impurities. In addition to the charge carriers introduced by the doping, there are the conduction electron-hole pairs already present in the undoped material. The density n_i of these pairs in the undoped material is called the intrinsic density. Even in a device made of just one material, the intrinsic density is position-dependent, because it depends on the temperature. An important aspect of semiconductor devices, the creation and recombination of electron-hole pairs, has not yet been mentioned explicitly. The model represents generation and recombination by a position-dependent term R. We use the following version of the semiconductor device equations: equations (1) and (2) give the standard relations between the potential ψ, the electric field $\boldsymbol{E}$ and the total charge density. The total charge density is made up of the hole density p, the electron density n and the charges captured by the impurities when fully ionised,

$$\operatorname{div}(\epsilon\, \boldsymbol{E}) = q(p-n+N_d-N_a)\ , \tag{1.1}$$

$$\boldsymbol{E} = -\,\mathbf{grad}\,\psi\ , \tag{1.2}$$

equations (3) and (5), the so-called continuity equations, indicate charge conservation. Equations (4) and (6), the carrier transport equations, relate the current densities J to the charge carrier densities p and n.

$$q\ \partial p / \partial t = -\operatorname{div} \boldsymbol{J}_p - qR\ , \tag{1.3}$$

$$\boldsymbol{J}_p = -q\mu_p \left[\frac{D_p}{\mu_p} \,\mathbf{grad}\, p - p[\boldsymbol{E} + U_T \,\mathbf{grad}\log n_i] \right]\ , \tag{1.4}$$

$$q\ \partial n / \partial t = \operatorname{div} \boldsymbol{J}_n - qR\ , \tag{1.5}$$

$$\boldsymbol{J}_n = q\mu_n \left[\frac{D_n}{\mu_n} \,\mathbf{grad}\, n + n[\boldsymbol{E} - U_T \,\mathbf{grad}\log n_i] \right]\ , \tag{1.6}$$

where D_p and D_n are diffusion constants, μ_p and μ_n are mobilities, ϵ is the permittivity and q is the absolute value of the charge of an electron. From this point on, we assume that the Einstein relations hold for the electron and hole diffusion coefficients, i.e. $D_n = U_T\mu_n$ and $D_p = U_T\mu_p$. The thermal voltage U_T, used above, is defined as,

$$U_T = \frac{k_B T}{q}\ . \tag{1.7}$$

We can further simplify the equations by only considering the stationary case.

Scaling the equations.

We go from the physical equations to a mathematical model. We scale the equations to make all relevant quantities dimensionless. This can be done in various ways [12, 13, 15]. We shall use the same symbols for the scaled and unscaled quantities with the exception of the permittivity, where we replace ϵ by λ^2. All references give the same general approach to scaling. The differences in the scalings are caused by different choices of values for the scaling parameters. We start by scaling all densities by a chosen reference density. We then make the resulting quantities dimensionless by scaling with an appropriate combination of a reference length l, a reference mobility $\tilde{\mu}$, the elementary charge q and the thermal voltage U_T. Note that this still leaves open the choice of the scaling factors l, N and $\tilde{\mu}$. The permittivity is replaced by a quantity λ^2, where λ is defined as

$$\lambda = \lambda_D / l \tag{1.8}$$

with

$$\lambda_D = \left[\epsilon \frac{U_T}{qN} \right]^{1/2}\ , \tag{1.9}$$

where λ_D is a Debye length of the device. By choosing appropriate N, $\tilde{\mu}$ and l, we can emphasise different aspects of the problem. We should rewrite the boundary conditions in terms of the scaled variables.

Two types of scaling are in general use, $N = n_i$ and $N = \max(N_d,N_a)$, in both cases l is taken to be of the order of the device length and $\tilde{\mu}$ is taken to be of the order of μ_p and μ_n.

The choice of variables.

The scaled equations can be expressed in several different sets of variables [12, 13, 15]. All sets of variables contain J_p,J_n, E and ψ, but they differ in the choice of variable for the scaled version of the equations (4) and (6). We find the following scaled equations for ψ, $\mathbf{E}$, $\mathbf{J}_p$ and $\mathbf{J}_n$,

$$\mathrm{div}\,\lambda^2\mathbf{E} = n - p - N_d + N_a \,, \tag{1.10}$$

is the scaled version of (1) and

$$\mathbf{E} = \mathbf{grad}\,\psi \,, \tag{1.11}$$

is the scaled version of (2). Instead of (3) and (5), we use

$$\mathrm{div}\boldsymbol{J}_p = -\frac{\partial p}{\partial t} - R \,, \tag{1.12}$$

$$\mathrm{div}\boldsymbol{J}_n = \frac{\partial n}{\partial t} + R \,. \tag{1.13}$$

As we consider only the stationary case, we may ignore the time derivatives. If we use

$$\overline{p} := \frac{p}{n_i} \,, \overline{n} := \frac{n}{n_i} \,, \tag{1.14}$$

as variables associated with the densities, we find the following scaled versions for (4) and (6),

$$\boldsymbol{J}_p = -\mu_p n_i(\,\mathbf{grad}\overline{p} + \overline{p}\,\mathbf{grad}\,\psi) \,, \tag{1.15}$$

$$\boldsymbol{J}_n = \mu_n n_i(\,\mathbf{grad}\,\overline{n} + \overline{n}\,\mathbf{grad}\,\psi) \,. \tag{1.16}$$

This set of variables has the disadvantage that the range of magnitudes for $\overline{n}$ and $\overline{p}$ is large. Moreover $\mathbf{grad}\,\psi$ may be very large around junctions, leading to a locally singularly perturbed problem. An advantage of this form is the linearity in the densities. The quasi Fermi potentials ϕ_p,ϕ_n have a more favourable range and avoid convection terms, but the equations are strongly nonlinear. The variables are defined by,

$$\varphi_p := \ln(\overline{p}\,) + \psi \,, \, \varphi_n := \psi - \ln(\,\overline{n}\,) \,. \tag{1.17}$$

The equations (4) and (6) take the form,

$$\boldsymbol{J}_p = -\mu_p n_i \exp(\varphi_p - \psi)\nabla\varphi_p \,, \tag{1.18}$$

$$\boldsymbol{J}_n = -\mu_n n_i \exp(\psi - \varphi_n)\nabla\varphi_n \,. \tag{1.19}$$

Finally we mention the Slotboom variables Φ_p,Φ_n [32, 33], they give us a formulation that is well suited for theoretical study, but the range of the Φ is very large, so these variables are not really suited for numerical work. The variables are defined as follows.

$$\Phi_p = \exp(\varphi_p) \ , \ \Phi_n = \exp(-\varphi_n) \ . \tag{1.20}$$

For the Slotboom variables equations (4) and (6) have the form,

$$\boldsymbol{J}_p = -\mu_p n_i \exp(-\psi) \, \mathbf{grad} \, \Phi_p \ , \tag{1.21}$$

$$\boldsymbol{J}_n = \mu_n n_i \exp(\psi) \, \mathbf{grad} \, \Phi_n \ . \tag{1.22}$$

We now have several mathematical models in the form of sets of equations in dimensionless parameters and variables.

The Scharfetter-Gummel discretisation.

None of the sets of model equations derived above are integrable for arbitrary doping and recombination terms. So, if we want an approximation of the solution, we need to use numerical analysis. The most popular method is an exponentially fitted difference scheme, called the Scharfetter-Gummel scheme. It was first formulated for the one-dimensional case. It can be found e.g. in a paper on the Read diode by Scharfetter and Gummel [34]. As equations to be discretised they take the set corresponding to the variables p and n and a constant intrinsic density. Note that they do not use a scaling that renders the equations dimensionless. They start of with the following set of equations,

$$\frac{\partial p}{\partial t} = -R - \frac{1}{q}\frac{\partial J_p}{\partial x} \ , \tag{1.23}$$

$$\frac{\partial n}{\partial t} = -R + \frac{1}{q}\frac{\partial J_n}{\partial x} \ , \tag{1.24}$$

$$\frac{\partial E}{\partial x} = \frac{q}{\epsilon}(p - n + N_d - N_a) \ , \tag{1.25}$$

$$J_p = q\mu_p pE - k_B T \mu_p \frac{\partial p}{\partial x} \ , \tag{1.26}$$

$$J_n = q\mu_n nE + k_B T \mu_n \frac{\partial n}{\partial x} \ . \tag{1.27}$$

They discretise this as follows. The device is partitioned into mesh cells. The variables p and n are determined on cell edges and the variables J_p, J_n and E are determined in cell centres. The equation for E and the continuity equations are replaced by the obvious finite difference equivalents. The carrier transport equations are treated differently. There they proceed as follows. They assume that E, J_p, J_n, μ_p and μ_n are constant between mesh points. They then solve the resulting ordinary differential equations for p and n. This results in a local version of the Il'in scheme [35]. As we do not consider the time-dependent problem in this thesis, we do not give their time discretisation.

1.7 Contents of this thesis.

When consulting the literature on numerical semiconductor device modelling, it is surprising that whereas so many papers use a finite volume, box or Mixed Finite Element Method for the semiconductor equations, papers that discuss a-priori error estimates for box or finite volume methods such as [36, 37] are rare. Articles on a-posteriori error estimates for MFEM methods are available only in fluid dynamics, e.g. the paper on a-posteriori error estimates for a mixed finite element discretisation for the Navier-Stokes equations by Verfürth [38]. The fundamental differences between the MFEM for Navier-Stokes and MFEM for our problem preclude the use of that method here. Even for the simple case of mixed finite elements for the Poisson equation we did not find any papers. This curious lack of information, combined with the ongoing research on the use of local refinement in multi-grid methods for the semiconductor problem at the institute [39], was the stimulus to write this thesis. Chapters 2, 4 and 5 deal with the search for a-posteriori error estimates for discretisations of the equations for the electric potential and charge transport that make up the stationary drift-diffusion model.

To discuss the contents of the remaining chapters it is necessary to keep in mind the general structure of the mixed finite element method. We recall the form of the mixed finite element method for a simple problem with constant coefficients,

$$\operatorname{div}\left\{\frac{1}{\alpha}[\,\mathbf{grad}\,u+\boldsymbol{\beta}u]\right\} = f \quad \text{on} \quad \Omega\,, \tag{1.28}$$

$$u = g \quad \text{on} \quad \partial\Omega\,. \tag{1.29}$$

The mixed finite element method is a discretised version of a variational formulation in terms of the solution of (28) with boundary conditions (29) and flux $\boldsymbol{\sigma}$, given by (30),

$$\boldsymbol{\sigma} = -\frac{1}{\alpha}[\,\mathbf{grad}\,u+\boldsymbol{\beta}u] \quad \text{on} \quad \Omega\,. \tag{1.30}$$

It is easily seen, that any solution of (28), (29) and (30) satisfies,

$$(\alpha\boldsymbol{\sigma},\boldsymbol{\tau}) - (\operatorname{div}\boldsymbol{\tau},u) + (u\boldsymbol{\beta},\boldsymbol{\tau}) = -[(\operatorname{div}\boldsymbol{\tau},u) + (\boldsymbol{\tau},\mathbf{grad}\,u)] = \tag{1.31a}$$

$$- < g,\boldsymbol{\tau}_h{\cdot}\mathbf{n}_{\partial\Omega} > \quad \forall\ \boldsymbol{\tau} \in \mathrm{H}(\mathrm{div};\Omega)\,,$$

$$(\operatorname{div}\boldsymbol{\sigma},t) = (f,t) \quad \forall\ t \in \mathrm{L}^2(\Omega)\,. \tag{1.31b}$$

A general Petrov-Galerkin discretisation now looks for a solution $(\boldsymbol{\sigma}_h,u_h)$ in a trial space $V_h \times W_h$, for the variational problem,

$$(\alpha\boldsymbol{\sigma}_h,\boldsymbol{\tau}_h) - (\operatorname{div}\boldsymbol{\tau}_h,u_h) + (u_h\boldsymbol{\beta},\boldsymbol{\tau}_h) = \ < g,\boldsymbol{\tau}_h{\cdot}\mathbf{n}_{\partial\Omega} > \quad \forall\ \boldsymbol{\tau}_h \in X_h\,,$$

$$(\operatorname{div}\boldsymbol{\sigma}_h,t_h) = (f,t_h) \quad \forall\ t_h \in Y_h\,,$$

where $X_h \times Y_h$ is a test space. We use a projection operator $\Pi_h \times \mathrm{P}_h$ [40, 41], where P_h is self-adjoint in the Hilbert space $\mathrm{L}^2(\Omega)$ and $\operatorname{div}\circ\Pi_h = \mathrm{P}_h\circ\operatorname{div}$, to define the discretisation error. We use a theorem proved first by Brezzi [42] and generalised by Nicolaides [43] to obtain estimates of the global discretisation error from the local discretisation error. The local discretisation error $(G,F) \in X_h^* \times Y_h^*$, has the form $G(\boldsymbol{\tau}_h) = (\alpha(\boldsymbol{\sigma}-\Pi_h\boldsymbol{\sigma}),\tau_h)$ and $F=0$, where we used the properties of the projection operator. If we use a quadrature rule to evaluate $(\alpha\boldsymbol{\sigma}_h,\boldsymbol{\tau}_h)$ and we denote the bilinear form that takes into account the quadrature by $a_h(\boldsymbol{\sigma}_h,\boldsymbol{\tau}_h)$, then G changes to $G(\boldsymbol{\tau}_h) = (\alpha\boldsymbol{\sigma},\tau_h) - a_h(\Pi_h\boldsymbol{\sigma},\tau_h)$. We note that to minimise the local discretisation error, we need weights in the quadrature rule that minimise this G for all relevant $\boldsymbol{\sigma}$.

Here, we introduce the point of view that forms the basis for the research contained in this thesis. The point of view is the following. If all elements of the flux space V are smoother than the elements of V_h, then minimising G for $\boldsymbol{\sigma} \in V$ is not equivalent to minimising $(\alpha\boldsymbol{\sigma}_h,\tau_h) - a_h(\boldsymbol{\sigma}_h,\tau_h)$ for $\boldsymbol{\sigma}_h \in V_h$. In Chapter two we use this viewpoint to find two alternatives to standard mixed finite elements for a symmetric second order elliptic operator. We use the lowest order Raviart-Thomas [41] space as test and trial spaces. Two quadrature rules are considered, a one-point rule and a three-point rule. We compare the results with exact quadrature of $(\alpha\boldsymbol{\sigma}_h,\boldsymbol{\tau}_h)$. The one-point rule gives the known lumped scheme. The scheme obtained by using the three-point rule is new. We show that the one-point rule is the most efficient. We also show that solving with the three-point rule gives us a more accurate solution than the one-point rule or even exact quadrature of $(\alpha\boldsymbol{\sigma}_h,\boldsymbol{\tau}_h)$. If the local discretisation error for the one-point rule is of order k in the mesh-width, then it is of order $k+2$ for the three-point rule. We give numerical results to illustrate the behaviour of the discretisations based on the one-point and the three-point rule. Other papers use the properties of Π_h to determine an estimate for the global error [44-46], but they do not consider lumping or other quadrature rules.

An important problem in numerical simulation of semiconductors is the existence of locally large electric fields around junctions between differently doped materials. We look for discretisation schemes whose accuracy outside areas of large electric field does not depend on the electric field in those areas. In chapter three we consider Galerkin mixed finite elements for the symmetrised continuity equation in one dimension. We use an abstract discrete space $V_h \times W_h$ as test and trial space and an abstract projection $\Pi_h \times \mathrm{P}_h$ with the properties mentioned earlier. We rederive the scheme as a Petrov-Galerkin mixed finite element method for the original continuity equation. We assume that the original equation has the form

$$-\frac{d}{dx}\left[a(x)\frac{d}{dx}u(x)+b(x)u(x)\right] = f(x) \quad \text{on} \quad \Omega\ ,$$

and that we have homogeneous boundary conditions. We also assume that $1/a$ and b are integrable and of fixed sign. We show that the abstract discretisations have the desired property by giving new error estimates for these discretisations. We find a uniform $\mathrm{L}^\infty(\Omega)$ error for the flux. The error is completely determined by the accuracy with which the right hand side f is approximated by $P_h f$. For the charge density, we obtain a cell-wise upper bound on a problem dependent discretisation error. The coefficient in the estimate depends only on the existence of a positive lower bound on the absolute value of the convection. The discretisation error is close to the $L^\infty(\Omega)$ discretisation error for cells with normal convection terms. We see that the discretisation error caused by small areas with large convection-diffusion ratios is restricted to those areas.

Chapter four deals with the two-dimensional Scharfetter-Gummel discretisation as described by Bank et al. [17]. We use a special a_h to write this finite volume scheme in the mixed finite element form given earlier. We then formulate an expression for the local discretisation error and use the theorem by Nicolaides to get a global error estimate on sufficiently fine meshes. We use the expression for the local discretisation error to determine the effects of non-uniform meshes and large electric fields on the accuracy of this discretisation. Moreover, we use the local discretisation error to construct a deferred correction process. We prove that the deferred correction process increases the accuracy of our solution. We show numerical results for the deferred correction process.

In Chapter five we discuss a new Petrov-Galerkin mixed finite element discretisation. We use the techniques of chapter two, i.e. we introduce a one-point and a three-point quadrature rule. The scheme that uses the one-point rule is stable and consistent. Moreover, if the local discretisation error for the one-point rule is of order k in the mesh-width, then it is of order $k+2$ for the three-point rule. Unlike the discretisation for the symmetrised equation, this discretisation can deal with convection terms that are not generated by a gradient, so it is applicable to more general convection-diffusion problems. Our error estimate for the discretisation based on the one-point rule takes into account the use of cell-wise averages for the coefficients. This error estimate degenerates for singular perturbed problems. However, we can show that the coefficients of the scheme approach a two-dimensional upwind scheme in the limit of vanishing diffusion. Moreover, if the convection term is the gradient of a linear function ψ, the solution $\exp(-\psi)$ is recovered exactly, just as in the finite difference scheme by Il'in [35].

Finally, it is necessary to explain why this thesis seemingly ignores the projection error. This can be explained as follows. The main reason is that the projections used give average densities over cells and average flux densities through cell edges. In most cases these are the quantities of interest. Moreover, if more information is needed, the assumption that the quantities of interest are differentiable makes an analysis of the order behaviour of the projection error almost trivial. An a-posteriori estimate of the projection error is easily obtained in the case where we we have second order behaviour of our global discretisation error, because that implies we can approximate first order derivatives of the unknowns by first order divided differences of the discrete solution. A variation on this idea is the basis for the deferred correction scheme in chapter 4, we refer to that chapter for additional information on the approximation of derivatives mentioned above.

References

1. Morton E. Jones, William C. Holton, and Robert Stratton, "Semiconductors: The Key to Computational Plenty," *Proceedings of the IEEE*, vol. 70, no. 12, pp. 1380-1409, December 1982.
2. S. S. Wilson, "Bicycle Technology," *Scientific American*, vol. 228, no. 3, pp. 81-91, 1973.
3. Hugh Kenner, "Print Queue: Bicycles for the Mind," *Byte*, vol. 16, no. 8, p. 334, McGraw-Hill, 1991.
4. J. M. Early, "Semiconductor Devices," *Proceedings of the IRE*, pp. 1006-1011, May 1962.
5. Richard L. Petritz, "Contributions of Materials Technology to Semiconductor Devices," *Proceedings of the IRE*, pp. 1006-1011, May 1962.
6. S. M. Sze, "Semiconductor Device Development in the 1970's and 1980's-A Perspective," *Proceedings of the IEEE*, vol. 69, no. 9, pp. 1121-1131, September 1981.
7. Glenford J. Meyers, Albert Y. C. Yu, and David L. Hause, "Micro Processor Technology Trends," *Proceedings of the IEEE*, vol. 74, no. 12, pp. 1605-1622, December 1986.
8. Richard B. Fair, "Challenges to Manufacturing Ultra Large Scale Integrated Circuits," *Proceedings of the IEEE*, vol. 78, no. 11, pp. 1687-1706, November 1989.
9. J. S. Blakemore, *Semiconductor statistics,* Dover, New York, 1987.
10. Neil W. Ashcroft and N. David Mermin, *Solid State Physics,* Holt-Sounders, 1981.
11. Shyh Wang, *Fundamentals of Semiconductor Theory and Device Physics,* Prentice-Hall, Englewood Cliffs, New Jersey, 1989.
12. Peter A. Markowich, *The Stationary Semiconductor Device Equations,* Springer-Verlag, Wien New York, 1986.
13. Siegfried Selberherr, *Analysis and simulation of semiconductor devices,* Springer-verlag, Wien New York, 1984.
14. Walter L. Engl, Heinz K. Dirks, and Bernd Meinerzhagen, "Device Modeling," *Proceedings of the IEEE*, vol. 71, no. 1, pp. 10-33, January 1983.
15. S. J. Polak, C. den Heijer, H. A. Schilders, and P. Markowich, "Semiconductor device modelling from the numerical point of view," *International Journal for Numerical Methods in Engineering*, vol. 24, pp. 763-838, 1987.

16. Wolfgang Fichtner, Donald J. Rose, and Randolph E. Bank, "Semiconductor device simulation," *SIAM journal of scientific and statistical computing*, vol. 4, no. 3, pp. 391-415, 1983.

17. Wolfgang Fichtner, Donald J. Rose, and Randolph E. Bank, "Numerical methods for semiconductor device simulation," *SIAM journal of scientific and statistical computing*, vol. 4, no. 3, pp. 416-435, 1983.

18. R. K. Smith, W. H. Coughran, Jr., W. Fichtner, D. J. Rose, and R. E. Bank, "Some aspects of semiconductor device simulation," in *Computing Methods in Applied Sciences and Engineering*, ed. J. L. Lions, vol. VII, pp. 3-12, North-Holland, 1986.

19. Randolph E. Bank, Joseph W. Jerome, and Donald J. Rose, "Analytical and numerical aspects of semiconductor device modelling," in *Computing Methods in Applied Sciences and Engineering*, ed. J. L. Lions, vol. V, pp. 593-597, North-Holland, 1982.

20. R. E. Bank, W. Fichtner, D. J. Rose, and R. K. Smith, "Algorithms for semiconductor device simulation," in *Large Scale Scientific Computation*, ed. B. Engquist, Progress in Scientific Computing, Birkhäuser, 1987.

21. M. S. Mock, "On Equations Describing Steady-State Carrier Distributions in a Semiconductor Device," *Communications on Pure and Applied Mathematics*, vol. XXV, pp. 781-792, 1972.

22. M. S. Mock, "Analysis of a discretisation algorithm for stationary continuity equations in semiconductor device models," *COMPEL*, vol. 2, pp. 117-139, 1983.

23. M. S. Mock, "Some recent results and open questions in numerical simulation of semiconductor devices," in *Computing Methods in Applied Sciences and Engineering*, ed. J. L. Lions, vol. VI, pp. 713-728, North-Holand, 1984.

24. M. S. Mock, "Analysis of a discretisation algorithm for stationary continuity equations in semiconductor device models II," *COMPEL*, vol. 3, pp. 137-149, 1984.

25. F. Brezzi, L. D. Marini, and P. Pietra, "Methodes d'elements finis mixtes et schema de Scharfetter-Gummel," *C. R. Acad. Sci. ser. I*, vol. 305, pp. 599-605, 1987.

26. M. Zlamál, "Finite Element Solution of the Fundamental Equations of Semiconductor Devices. I," *Mathematics of Computation*, vol. 46, no. 173, pp. 27-43, 1986.

27. J. J. H. Miller, S. Wang, and C. H. Wu, "A mixed finite element method for the stationary semiconductor continuity equations," *Eng. Comput.*, vol. 5, pp. 285-288, 1988.

28. Song Wang and Changhui Wu, "Mixed finite element approximation of the stationary semiconductor equations," in *Simulation of semiconductor devices and processes*, ed. M. Rudan, vol. 3, pp. 475-484, Tecnoprint,

Bologna, 1988.

29. Naoyuki Shigyo, Tetsunori Wada, and Seiji Yasuda, "Discretisation Problem for Multidimensional Current Flow," *IEEE Transactions on Computer-Aided Design*, vol. 8, no. 10, pp. 1046-1050, 1989.

30. Gen-Lin Tan, Xiao-Li Yuan, Qi-Ming Zhang, Walter H. Ku, and An-Jui Shey, "Two-Dimensional Semiconductor Device Analysis Based on New Finite-Element Discretization Employing the S-G Scheme.," *IEEE Transactions on Computer-Aided Deisign*, vol. 8, no. 5, pp. 468-478, 1989.

31. W. Shockley, "The Theory of *p-n* Junctions in Semiconductors and *p-n* Junction Transistors," *Bell System Technical Journal*, pp. 435-489, July 1949.

32. J. W. Slotboom, "Interactive scheme for 1- and 2-dimensional d.c. transistor simulation," *Electronics Letters*, vol. 5, no. 26, pp. 677-678, 1969.

33. J. W. Slotboom, "Computer-aided two-dimensional analysis of bipolar transistors," *IEEE Transactions on Electron Devices*, vol. ED-20, no. 8, pp. 669-679, 1973.

34. D. L. Scharfetter and H. K. Gummel, "Large-Signal Analysis of a Silicon Read Diode Oscillator," *IEEE Transactions on Electron Devices*, vol. ED-16, no. 1, pp. 64-77, 1969.

35. A. M. Il'in, "Differencing scheme for a differential equation with a small parameter affecting the highest derivative," *Mathematical Notes of the Academy of Sciences of the USSR*, vol. 6, no. 1-2, pp. 596-602, 1969.

36. W. Hackbusch, "On First and Second Order box Schemes," *Computing*, vol. 41, pp. 277-296, 1989.

37. Randolph E. Bank and Donald J. Rose, "Some Error Estimates for the Box Method.," *SIAM J. Numer. Anal.*, vol. 24, pp. 777-787, 1987.

38. R. Verfürth, "A Posteriori Error Estimators for the Stokes Equations," *Numerische Mathematik*, vol. 55, pp. 309-325, 1989.

39. J. Molenaar and P. Hemker, "A multigrid approach for the solution of the 2D semiconductor equation," *IMPACT of computing in Science and Engineering*, vol. 2, no. 3, pp. 219-243, 1990.

40. M. Fortin, "An analysis of the convergence of mixed finite element methods," *RAIRO Numerical Analysis*, vol. 11, no. 4, pp. 341-354, 1977.

41. P. A. Raviart and J. M. Thomas, "A mixed finite element method for 2-nd order elliptic problems," in *Mathematical aspects of the finite element method*, Lecture Notes in Mathematics, vol. 606, pp. 292-315, Springer, 1977.

42. F. Brezzi, "On the existence, Uniqueness and Approximation of saddle-point problems arising from Lagrangian multipliers," *RAIRO Num. Anal.*, vol. 8-R2, pp. 129-151, 1974.

43. R. A. Nicolaides, "Existence, uniqueness and approximation for generalized saddle point problems," *SIAM J. Numer. Anal.*, vol. 19, no. 2, pp. 349-357, 1982.

44. Junping Wang, "Superconvergence and extrapolation for mixed finite element methods on rectangular domains.," *Math. Comp.*, vol. 56, pp. 477-503, 1991.

45. J. Douglas, Jr. and J. Wang, "Superconvergence of mixed finite element methods on rectangular domains," *Calcolo*, vol. 26, pp. 121-133, 1989.

46. Mie Nakata, Alan Weiser, and Mary Fanett Wheeler, "Some superconvergence results for mixed finite element methods for elliptic problems on rectangular domains," in *The Mathematics of Finite Elements and Applications*, ed. J. R. Whiteman, vol. 5, pp. 367-389, 1985.

2. An Improved Accuracy Version of the Mixed Finite Element Method for a Second Order Elliptic Equation

2.1 Introduction.

In this chapter, we describe a modification of the mixed finite element method for a second order elliptic equation. The modified method is based on standard mixed finite elements with lowest order Raviart-Thomas elements on rectangles [1]. To give a brief description of our method, we recall, that a mixed finite element formulation of

$$-\operatorname{div} a\,\mathbf{grad}\,u + cu = f \ \text{ on } \ \Omega ,$$
$$u\,|_{\partial\Omega} = g ,$$

can be written as

$$(\sigma_h, \tau_h / a) - (\operatorname{div} \tau_h, u_h) = - < g, \tau_h{\cdot}\mathbf{n}_{\partial\Omega} > \quad \forall \ \tau_h \in V_h ,$$
$$(\operatorname{div} \sigma_h, t_h) + (cu_h, t_h) = (f, t_h) \quad \forall \ t_h \in W_h ,$$

where u_h is a discrete approximation of u and σ_h is a discrete approximation of $-a$ **grad** u. In this chapter, we show that, if we use a special quadrature rule for the inner product $(\sigma_h, \tau_h / a)$ and if the coefficient a is piecewise constant, then the difference between a suitable projection of the continuous solution and the discrete approximation is of order $\mathcal{O}(h^3)$. We give numerical evidence that confirms the theoretical result for smooth σ. In section 2.2, we formulate the boundary value problem to which we apply the modified mixed finite element method. Section 2.3 describes our mixed finite element discretisation and the quadrature rule for the inner product. There we also give a motivation for the use of the special quadrature rule. We give two other choices for the quadrature rule in section 2.4. One choice results in the usual scheme for lowest order Raviart-Thomas elements, the other choice corresponds to the use of the trapezoidal rule. We derive an error estimate for the modified version in section 2.5. A different but related approach to such error estimates can be found in [2-4]. In section 2.6, we use a one dimensional example to illustrate the importance of the ratio ch^2 / a for the usual scheme and our modified scheme. For these methods, the value of this ratio determines whether or not u_h satisfies a local maximum principle (cf. Polak, Schilders and Couperus [5]) For the scheme based on the trapezoidal rule, u_h satisfies a local maximum principle for all $c \geqslant 0$. In Section 2.7, we show numerical results. Section 2.8 gives an a-posteriori error estimator for the method based on the trapezoidal rule. In the last section, we summarise our results.

2.2 The equation.

We consider a second order elliptic equation with Dirichlet boundary conditions, as given in equation (1),

$$- \operatorname{div} a \, \mathbf{grad} \, u + cu = f \text{ on } \Omega \,, \tag{2.1a}$$

$$u = g \text{ on } \partial\Omega \,. \tag{2.1b}$$

on a rectangle $\Omega =]0,L_1[\times]0,L_2[$. We introduce a special notation for $-a$ **grad** u,

$$\sigma := -a \, \mathbf{grad} \, u \,. \tag{2.1c}$$

We assume, that there is a finite set of rectangles, the union of which covers Ω, such that a , c are constant on each separate rectangle. We assume that $a > 0$ and $c \geqslant 0$. We also assume, that a , c , f and g are such, that (1) has a unique solution $u \in C(\Omega)$, with a σ that is sufficiently smooth for our purposes.

2.3 The discretisation.

In this section, we give a description of our discretisation. We divide Ω into rectangular subdomains $\Omega_{i+½,j+½}$, we introduce some notation and we define our test function spaces V_h and W_h. We then introduce two projections P_h and Π_h. Such projections were suggested by Fortin [6] and are used by Raviart and Thomas [1] and Douglas and Roberts [7]. Next, we give the discretisation and discuss the special quadrature rule.

2.3.1. The partitioning of the domain.

We restrict ourselves to subdivisions of the rectangle Ω, that can be generated by the Cartesian product of subdivisions of its sides. Let

$$D_1 = \{ \, 0 = x_{1,0} < x_{1,1} < \cdots < x_{1,N_1} = L_1 \, \}$$

and

$$D_2 = \{ \, 0 = x_{2,0} < x_{2,1} < \cdots < x_{2,N_2} = L_2 \, \}$$

be partitions such, that a and c are constant on the interior of each separate rectangle of the subdivision $D_1 \times D_2$ of Ω. We set

$$h_{1,i+½} = x_{1,i+1} - x_{1,i} \,, \tag{2.2a}$$

$$h_{2,j+½} = x_{2,j+1} - x_{2,j} \,, \tag{2.2b}$$

and

$$\Omega_{i+½,j+½} = \{ \, (x_1,x_2) \mid x_{1,i} < x_1 < x_{1,i+1} \,, x_{2,j} < x_2 < x_{2,j+1} \, \} , \tag{2.3}$$

$$\Gamma_{i,j+½} = \{ \, (x_1,x_2) \mid x_1 = x_{i,1} \,, x_{2,j} < x_2 < x_{2,j+1} \, \} \,, \tag{2.4a}$$

$$\Gamma_{i+\frac{1}{2},j} = \{ (x_1,x_2) \mid x_{1,i} < x_1 < x_{1,i+1} , x_2 = x_{2,j} \} . \tag{2.4b}$$

2.3.2. The approximation spaces.

We define our approximation spaces for σ and u, by giving a basis for each space. We then introduce two projections onto the discrete spaces.

For each cell, $\Omega_{i+\frac{1}{2},j+\frac{1}{2}}$, we use the characteristic function $\chi_{i+\frac{1}{2},j+\frac{1}{2}}$,

$$\chi_{i+\frac{1}{2},j+\frac{1}{2}} = \delta_{ik}\delta_{jl} \quad \text{on} \quad \Omega_{k+\frac{1}{2},l+\frac{1}{2}} , \tag{2.5}$$

as an element in the set of basis functions for W_h. For V_h, we introduce the basis functions $\boldsymbol{\eta}_{i,j+\frac{1}{2}}$ and $\boldsymbol{\eta}_{i+\frac{1}{2},j}$, where $\boldsymbol{\eta}_{i,j+\frac{1}{2}}$ is linear in x_1 and constant in x_2 on each cell with

$$\boldsymbol{\eta}_{i,j+\frac{1}{2}} = \delta_{ik}\delta_{jl}\mathbf{e}_1 \quad \text{on} \quad \Gamma_{k,l+\frac{1}{2}} , \tag{2.6}$$

for $i,k=0,1,\ldots,N_1$, $j,l=0,1,\ldots,N_2-1$ and $\boldsymbol{\eta}_{i+\frac{1}{2},j}$ is linear in x_2 and constant in x_1 on each cell with

$$\boldsymbol{\eta}_{i+\frac{1}{2},j} = \delta_{ik}\delta_{jl}\mathbf{e}_2 \quad \text{on} \quad \Gamma_{k+\frac{1}{2},l} , \tag{2.7}$$

for $i,k=0,1,\ldots,N_1-1$, $j,l=0,1,\ldots,N_2$. Here $\mathbf{e}_1$ and $\mathbf{e}_2$ are unit vectors in the x_1- and x_2-direction respectively.

With these basis functions, we construct V_h and W_h,

$$V_h = Span(\{ \boldsymbol{\eta}_{i,j+\frac{1}{2}} \mid i=0,1,\ldots,N_1 , j=0,1,\ldots,N_2-1 \} \cup \tag{2.8}$$
$$\{ \boldsymbol{\eta}_{i+\frac{1}{2},j} \mid i=0,1,\ldots,N_1-1 , j=0,1,\ldots,N_2 \}) ,$$
$$W_h = Span(\{ \chi_{i+\frac{1}{2},j+\frac{1}{2}} \mid i=0,1,\ldots,N_1-1 , j=0,1,\ldots,N_2-1 \}). \tag{2.9}$$

The product space $V_h \times W_h$ is the space of lowest order Raviart-Thomas elements. To prepare for the definition of the two projections onto the discrete spaces, we introduce averages over cells and cell boundaries for $f \in C(\bar{\Omega})$,

$$P[\Omega_{i+\frac{1}{2},j+\frac{1}{2}}](f) = \frac{1}{\mu(\Omega_{i+\frac{1}{2},j+\frac{1}{2}})} \int_{\Omega_{i+\frac{1}{2},j+\frac{1}{2}}} f \, d\mu , \tag{2.10}$$

$$P[\Gamma_{i,j+\frac{1}{2}}](f) = \frac{1}{\lambda(\Gamma_{i,j+\frac{1}{2}})} \int_{\Gamma_{i,j+\frac{1}{2}}} f \, d\lambda , \tag{2.11}$$

$$P[\Gamma_{i+\frac{1}{2},j}](f) = \frac{1}{\lambda(\Gamma_{i+\frac{1}{2},j})} \int_{\Gamma_{i+\frac{1}{2},j}} f \, d\lambda . \tag{2.12}$$

In the above definitions, λ is the Lebesgue measure on $\mathbb{R}$ and μ is the Lebesgue measure on $\mathbb{R}^2$. We define $\mathrm{P}_h\colon \mathrm{L}^2(\Omega) \rightarrow W_h$ for all $u \in \mathrm{L}^2(\Omega)$,

$$\mathrm{P}_h u = P[\Omega_{i+\frac{1}{2},j+\frac{1}{2}}](u) \quad \text{on} \quad \Omega_{i+\frac{1}{2},j+\frac{1}{2}} \quad \forall \ i,j , \tag{2.13}$$

and we define $\Pi_h\colon \mathrm{H}^1(\Omega)^2 \rightarrow V_h$ for all $\sigma \in \mathrm{H}^1(\Omega)^2$,

$$(\Pi_h\sigma)_1 = P[\Gamma_{i,j+\frac{1}{2}}](\sigma_1) \quad \text{on} \quad \Gamma_{i,j+\frac{1}{2}} , \tag{2.14a}$$

$$(\Pi_h\sigma)_2 = P[\Gamma_{i+\frac{1}{2},j}](\sigma_2) \quad \text{on} \quad \Gamma_{i+\frac{1}{2},j} \quad \forall \ i,j . \tag{2.14b}$$

The spaces V_h and W_h and the projection Π_h were introduced by Raviart and Thomas [1, 6]. The projections have the following special properties.

Lemma 2.1.

$$\forall\ u \in \mathrm{L}^2(\Omega)\ ,\ t_h \in W_h : (u,t_h) = (\mathrm{P}_h u,t_h)\ , \tag{2.15a}$$

$$\forall\ \sigma \in \mathrm{H}^1(\Omega)^2\ ,\ t_h \in W_h : (\ \mathrm{div}\ \sigma,t_h) = (\ \mathrm{div}\ \Pi_h\sigma,t_h)\ . \tag{2.15b}$$

Proof.
Equation (15a) follows immediately from the definition of P_h. Green's formula,

$$\int_{\Omega_{i+\frac{1}{2},j+\frac{1}{2}}} \mathrm{div}\ \sigma\ d\mu = \int_{\partial\Omega_{i+\frac{1}{2},j+\frac{1}{2}}} \sigma\cdot\mathbf{n}_{\partial\Omega_{i+\frac{1}{2},j+\frac{1}{2}}}\ d\lambda\ ,$$

proves equation (15b) □

2.3.3. The discretisation scheme.

We first give the discretisation without specifying the quadrature rule. The choice of a quadrature rule is discussed in section 2.3.4.

We introduce the space

$$V = \mathrm{H}(\mathrm{div},\Omega) := \{\ \tau \in \mathrm{L}^2(\Omega)^2\ |\ \mathrm{div}\ \tau \in \mathrm{L}^2(\Omega)\ \}\ , \tag{2.16}$$

with inner product,

$$(\sigma,\tau)_V = (\sigma,\tau)_{\mathrm{L}^2(\Omega)^2} + (\ \mathrm{div}\ \sigma,\ \mathrm{div}\ \tau)_{\mathrm{L}^2(\Omega)}\ . \tag{2.17}$$

This space is discussed by Roberts and Thomas [8]. We also introduce

$$W = \mathrm{L}^2(\Omega)\ . \tag{2.18}$$

Note, that Π_h is only defined on $\mathrm{H}^1(\Omega)^2 \subset \mathrm{H}(\mathrm{div},\Omega)$. In this chapter, when we apply Π_h to the σ defined in (1c), the assumption that this σ lies in $\mathrm{H}^1(\Omega)^2$ is included in the condition "σ is smooth enough."

We can now write problem (1) in the form:

$$(\sigma,u) \in V\times W\ ,$$

$$\alpha(\sigma,\tau) - (\ \mathrm{div}\ \tau,u) = -\ <\ g,\tau\cdot\mathbf{n}_{\partial\Omega}\ > \quad \forall\ \tau \in V\ , \tag{2.19a}$$

$$(\ \mathrm{div}\ \sigma,t) + (cu,t) = (f,t) \quad \forall\ t \in W\ , \tag{2.19b}$$

where

$$\alpha(\sigma,\tau) := (\sigma,\tau/a) \quad \forall\ \sigma,\tau \in V\ . \tag{2.20}$$

For our discrete problem, we take

$$(\sigma_h,u_h) \in V_h\times W_h\ ,$$

$$\alpha_h(\sigma_h,\tau_h) - (\ \mathrm{div}\ \tau_h,u_h) = -\ <\ g,\tau_h\cdot\mathbf{n}_{\partial\Omega}\ > \quad \forall\ \tau_h \in V_h\ , \tag{2.21a}$$

$$(\ \mathrm{div}\ \sigma_h,t_h) + (cu_h,t_h) = (f,t_h) \quad \forall\ t_h \in W_h\ , \tag{2.21b}$$

where α_h is a bilinear form on $V \times V_h$, that approximates α and that satisfies

$$\alpha_h(\sigma,\tau_h) = \alpha_h(\Pi_h\sigma,\tau_h) \ . \tag{2.22}$$

2.3.4. The definition of α_h.

The bilinear form α_h describes the quadrature rule used to evaluate α. The idea behind the introduction of a special quadrature rule is the following, if we combine (19), (21) and (22) with the results of lemma 1, we find,

$$\alpha_h(\Pi_h\sigma - \sigma_h,\tau_h) - (\text{ div } \tau_h, \mathrm{P}_h u - u_h) = \alpha_h(\Pi_h\sigma,\tau_h) - \alpha(\sigma,\tau_h) \ , \tag{2.23a}$$

$$(\text{ div } (\Pi_h\sigma - \sigma_h), t_h) + (ct_h, \mathrm{P}_h u - u_h) = 0 \ . \tag{2.23b}$$

We see, that the only term on the right hand side of this equation is,

$$\alpha_h(\Pi_h\sigma,\tau_h) - \alpha(\sigma,\tau_h) \ . \tag{2.24}$$

If the discrete problem is uniquely solvable, then it is invertible. In that case this term is a measure for the difference between $(\Pi_h\sigma, \mathrm{P}_h u)$ and (σ_h, u_h). We now seek to minimise (24). To do this, we construct a special quadrature rule for the evaluation of $\alpha(\sigma,\tau_h)$ by defining this rule for $\alpha(\sigma,\boldsymbol{\eta}_1)$ and $\alpha(\sigma,\boldsymbol{\eta}_2)$, for each $\boldsymbol{\eta}_1$, $\boldsymbol{\eta}_2$ given by (6) and (7). We first introduce the obvious notations,

$$\begin{aligned} a_{i+½,j+½} &= P[\Omega_{i+½,j+½}](a) \ , \\ \sigma_{1,i,j+½} &= P[\Gamma_{i,j+½}](\sigma_1) \ , \\ \sigma_{2,i+½,j} &= P[\Gamma_{i+½,j}](\sigma_2) \ . \end{aligned}$$

Our two-dimensional integration rule corresponds to the use of a one-dimensional three-point rule in one direction and exact integration in the other. To simplify the definition of the quadrature rule, we introduce the following functions,

$$A(h,\tilde{h},L,R) = \frac{hL}{12} + \frac{h\tilde{h}L}{12(h+\tilde{h})} - \frac{\tilde{h}^3 R}{12h(h+\tilde{h})} \ , \tag{2.25a}$$

$$B(h,\tilde{h},L,R) = \frac{\tilde{h}(\tilde{h}+4h)R}{12h} + \frac{h(h+4\tilde{h})L}{12\tilde{h}} \ , \tag{2.25b}$$

$$C(h,\tilde{h},L,R) = A(\tilde{h},h,R,L) \ , \tag{2.25c}$$

$$D(h,\tilde{h},L,R) = \frac{3hL}{12} + \frac{h\tilde{h}L}{12(h+\tilde{h})} - \frac{\tilde{h}^3 R}{12h(h+\tilde{h})} \ , \tag{2.25d}$$

$$E(h,\tilde{h},L,R) = \frac{2\tilde{h}R}{12} + \frac{\tilde{h}^2 R}{12h} + \frac{2hL}{12} + \frac{h^2 L}{12\tilde{h}} \ , \tag{2.25e}$$

$$F(h,\tilde{h},L,R) = D(\tilde{h},h,R,L) \ . \tag{2.25f}$$

Where (25a-c) are used in rules for basis functions with their maximum in the interior of Ω and (25d-f) are used in rules for basis functions with a their maximum on the boundary of Ω.

Now, we define α_h for all basis functions. We start by defining its action for the $\mathbf{e}_1$ component of σ. We have to distinguish between basis functions with their maximum on the left boundary of Ω, (26a), in the interior, (26b), or on the right boundary of Ω (26c).

$$\alpha_h(\sigma,\boldsymbol{\eta}_{0,j+\frac{1}{2}})/h_{2,j+\frac{1}{2}} := \qquad (2.26a)$$

$$D(h_{1,\frac{1}{2}},h_{1,1+\frac{1}{2}},1/a_{\frac{1}{2},j+\frac{1}{2}},0)\sigma_{1,0,j+\frac{1}{2}} +$$

$$E(h_{1,\frac{1}{2}},h_{1,1+\frac{1}{2}}1/a_{\frac{1}{2},j+\frac{1}{2}},0)\sigma_{1,1,j+\frac{1}{2}} +$$

$$F(h_{1,\frac{1}{2}},h_{1,1+\frac{1}{2}}1/a_{\frac{1}{2},j+\frac{1}{2}},0)\sigma_{1,2,j+\frac{1}{2}} ,$$

$$\alpha_h(\sigma,\boldsymbol{\eta}_{i,j+\frac{1}{2}})/h_{2,j+\frac{1}{2}} := \qquad (2.26b)$$

$$A(h_{1,i-\frac{1}{2}},h_{1,i+\frac{1}{2}},1/a_{i-\frac{1}{2},j+\frac{1}{2}},1/a_{i+\frac{1}{2},j+\frac{1}{2}})\sigma_{1,i-1,j+\frac{1}{2}} +$$

$$B(h_{1,i-\frac{1}{2}},h_{1,i+\frac{1}{2}},1/a_{i-\frac{1}{2},j+\frac{1}{2}},1/a_{i+\frac{1}{2},j+\frac{1}{2}})\sigma_{1,i,j+\frac{1}{2}} +$$

$$C(h_{1,i-\frac{1}{2}},h_{1,i+\frac{1}{2}},1/a_{i-\frac{1}{2},j+\frac{1}{2}},1/a_{i+\frac{1}{2},j+\frac{1}{2}})\sigma_{1,i+1,j+\frac{1}{2}} ,$$

$$\alpha_h(\sigma,\boldsymbol{\eta}_{N_1,j+\frac{1}{2}})/h_{2,j+\frac{1}{2}} := \qquad (2.26c)$$

$$D(h_{1,N_1-1-\frac{1}{2}},h_{1,N_1-\frac{1}{2}},0,1/a_{N_1-\frac{1}{2},j+\frac{1}{2}})\sigma_{1,N_1-2,j+\frac{1}{2}} +$$

$$E(h_{1,N_1-1-\frac{1}{2}},h_{1,N_1-\frac{1}{2}},0,1/a_{N_1-\frac{1}{2},j+\frac{1}{2}})\sigma_{1,N_1-1,j+\frac{1}{2}} +$$

$$F(h_{1,N_1-1-\frac{1}{2}},h_{1,N_1-\frac{1}{2}},0,1/a_{N_1-\frac{1}{2},j+\frac{1}{2}})\sigma_{1,N_1,j+\frac{1}{2}} ,$$

$$\text{for } \ i=1,2,\ldots,N_1-1 \ , \ j=0,1,\ldots,N_2-1 \ .$$

Next, we define the rule for basis elements for the $\mathbf{e}_2$ component. Again, we have to distinguish between basis functions with their maximum on the boundary of Ω, (26d, 26f), and basis functions with their maximum in the interior (26e).

$$\alpha_h(\sigma,\boldsymbol{\eta}_{i+\frac{1}{2},0})/h_{1,i+\frac{1}{2}} := \qquad (2.26d)$$

$$D(h_{2,\frac{1}{2}},h_{2,1+\frac{1}{2}},1/a_{i+\frac{1}{2},\frac{1}{2}},0)\sigma_{2,i+\frac{1}{2},0} +$$

$$E(h_{2,\frac{1}{2}},h_{2,1+\frac{1}{2}},1/a_{i+\frac{1}{2},\frac{1}{2}},0)\sigma_{2,i+\frac{1}{2},1} +$$

$$F(h_{2,\frac{1}{2}},h_{2,1+\frac{1}{2}},1/a_{i+\frac{1}{2},\frac{1}{2}},0)\sigma_{2,i+\frac{1}{2},2} ,$$

$$\alpha_h(\sigma,\boldsymbol{\eta}_{i+\frac{1}{2},j})/h_{1,i+\frac{1}{2}} := \qquad (2.26e)$$

$$A(h_{2,j-\frac{1}{2}},h_{2,j+\frac{1}{2}},1/a_{i+\frac{1}{2},j-\frac{1}{2}},1/a_{i+\frac{1}{2},j+\frac{1}{2}})\sigma_{2,i+\frac{1}{2},j-1} +$$

$$B(h_{2,j-\frac{1}{2}},h_{2,j+\frac{1}{2}},1/a_{i+\frac{1}{2},j-\frac{1}{2}},1/a_{i+\frac{1}{2},j+\frac{1}{2}})\sigma_{2,i+\frac{1}{2},j} +$$

$$C(h_{2,j-\frac{1}{2}},h_{2,j+\frac{1}{2}},1/a_{i+\frac{1}{2},j-\frac{1}{2}},1/a_{i+\frac{1}{2},j+\frac{1}{2}})\sigma_{2,i+\frac{1}{2},j+1} ,$$

$$\alpha_h(\sigma,\boldsymbol{\eta}_{i+\frac{1}{2},N_2})/h_{1,i+\frac{1}{2}} := \qquad (2.26f)$$

$$D(h_{2,N_2-1-\frac{1}{2}},h_{2,N_2-\frac{1}{2}},0,1/a_{i+\frac{1}{2},N_2-\frac{1}{2}})\sigma_{2,i+\frac{1}{2},N_2-2} +$$

$$E(h_{2,N_2-1-\frac{1}{2}},h_{2,N_2-\frac{1}{2}},0,1/a_{i+\frac{1}{2},N_2-\frac{1}{2}})\sigma_{2,i+\frac{1}{2},N_2-1} +$$

$$F(h_{2,N_2-1-\frac{1}{2}},h_{2,N_2-\frac{1}{2}},0,1/a_{i+\frac{1}{2},N_2-\frac{1}{2}})\sigma_{2,i+\frac{1}{2},N_2} \ ,$$

$$\text{for } i=0,1,\ldots,N_1-1 \ , \quad \text{for } j=1,2,\ldots,N_2-1 \ , \ .$$

In section 2.5.1, we show, that for the above choice of coefficients, (24) is $O(h^3)$. The use of a three point integration rule means, that we cannot obtain a higher order than this for (24) unless the mesh is uniform and the coefficients are constant on Ω, in which case we gain a factor of h due to symmetry.

2.4 Other quadrature rules.

If we take different coefficients in our quadrature rule α_h, we find other variations on the mixed finite element method for lowest order Raviart-Thomas elements.

2.4.1. Exact evaluation of the form α on test and trial functions.

If we assume piecewise constant coefficients and we use exact integration for the product of test and trial functions, we obtain,

$$A(h,\tilde{h},L,R) = \frac{hL}{6} \ , \tag{2.27a}$$

$$B(h,\tilde{h},L,R) = \frac{hL}{3} + \frac{\tilde{h}R}{3} \ , \tag{2.27b}$$

$$C(h,\tilde{h},L,R) = A(\tilde{h},h,R,L) \ , \tag{2.27c}$$

$$D(h,\tilde{h},L,R) = \frac{hL}{3} \ , \tag{2.27d}$$

$$E(h,\tilde{h},L,R) = \frac{hL}{6} + \frac{\tilde{h}R}{6} \tag{2.27e}$$

$$F(h,\tilde{h},L,R) = D(\tilde{h},h,R,L) \ , \tag{2.27f}$$

this choice results in the usual mixed finite element scheme for this choice of test and the trial function spaces.

2.4.2. Use of the trapezoidal rule.

The use of the trapezoidal rule corresponds to the choice,

$$A(h,\tilde{h},L,R) = 0 \ , \tag{2.28a}$$

$$B(h,\tilde{h},L,R) = \frac{hL}{2} + \frac{\tilde{h}R}{2} \ , \tag{2.28b}$$

$$C(h,\tilde{h},L,R) = A(\tilde{h},h,R,L) \ , \tag{2.28c}$$

$$D(h,\tilde{h},L,R) = \frac{hL}{2} \ , \tag{2.28d}$$

$$E(h,\tilde{h},L,R) = 0 \tag{2.28e}$$

$$F(h,\tilde{h},L,R) = D(\tilde{h},h,R,L) \ . \tag{2.28f}$$

For this scheme, elimination of σ_h by static condensation is trivial. For $c \geqslant 0$, the resulting matrix is an M-matrix. This implies, that u_h satisfies a local maximum principle for $c \geqslant 0$. If $a \equiv 1$ and $c \equiv 0$ then the matrix after static condensation corresponds to the classical five point finite difference stencil for the Laplace operator.

2.5 An error estimate.

We derive estimates for $\| \Pi_h \sigma - \sigma_h \|_{L^2(\Omega)}$ and $\| P_h u - u_h \|_{L^2(\Omega)}$ under the conditions,

$$c \geqslant 0 \text{ on } \Omega \,, \tag{C1}$$

$$\sigma \text{ is smooth enough,} \tag{C2}$$

and

$$A_0(\tau_h,\tau_h) \leqslant \alpha_h(\tau_h,\tau_h) \leqslant A_1(\tau_h,\tau_h) \,, \tag{C3}$$

where A_0 and A_1 are positive real numbers, independent of the mesh. To derive error estimates, we need an estimate of the quadrature error, given in section 2.5.1, and a special norm on V_h, given in section 2.5.2. Section 2.5.3 contains the proof of the error estimate. In section 2.5.4 we show that condition (C3) is satisfied for a special case.

2.5.1. Error estimates for integration formulas.

We derive an error estimate for our special two dimensional quadrature rule. This rule is based on the interpretation of the values of $\Pi_h \sigma$ on the edges of cells as averages over those edges. Combined with a piecewise constant a and essentially one-dimensional weight functions, this allows a simple extension of one dimensional integration rules to two dimensions.

To prove this, we combine a special case of Theorem 2 of Bramble and Hilbert [9] with Fubini's theorem [10, 11] and a Sobolev embedding theorem [12]. In lemma 5 we combine these results to give an error estimate. In lemmas 6 and 7 we show that the coefficients given in section 2.3.4 satisfy the conditions of lemma 5. In lemmas 2, 3 and 4 we formulate the theorems used.

In this thesis we shall often use Sobolev spaces. The general Sobolev space $W^{k,p}(\Omega)$ is the space of functions for which the generalised k^{th} order derivative to the power p is integrable. The usual norm on this space is defined as,

$$\| u \|_{W^{k,p}(\Omega)} = \left[\sum_{j=0}^{k} \| d^j u / dx^j \|^p_{L^p(\Omega)} \right]^{1/p} ,$$

with

$$\| u \|_{L^p(\Omega)} = \left[\int_\Omega u^p \, d\mu \right]^{1/p} .$$

Cf. [13-15]. The usual notation for $W^{k,2}(\Omega)$ is $H^k(\Omega)$. In section 3.2.1 of chapter 3 we define the term generalised derivative in one dimension.

Lemma 2.2.
Let Ω be an interval of length $\rho < \infty$ and let $1 \leqslant p < \infty$. If F is a linear functional on the Sobolev space $W^{k,p}(\Omega)$, which satisfies

$$\exists \; C > 0 : |F(u)| \leqslant C|| \; ||u||_{k,p,\Omega} \quad \forall \; u \in W^{k,p}(\Omega) \, , \tag{i}$$

with C independent of ρ,

$$F(v) \equiv 0 \quad \forall \; v \in \{ \, 1,x, \, . \, . \, . \, ,x^{k-1} \, \} \, , \tag{ii}$$

then

$$\exists \; \tilde{C} > 0 : |F(u)| \leqslant \tilde{C}\rho^k \, || d^k u \, / \, dx^k \, ||_{p,\Omega} \, ,$$

with $\tilde{C}$ independent of ρ. Where

$$||u||_{p,\Omega} = \left[\frac{1}{\rho} \int_\Omega u^p \; d\mu \right]^{1/p} \, ,$$

$$||u||_{k,p,\Omega} = \left[\sum_{j=0}^{k} \rho^{kp} \, || d^j u \, / \, dx^j \, ||_{p,\Omega}^p \right]^{1/p} \, .$$

Proof.
This is a special case of Theorem 2 from the paper by Bramble and Hilbert [9].

□

Note that the norms used in this lemma are those used in the paper by Bramble and Hilbert and differ from the norms used in this thesis.

Lemma 2.3.
Let Ω_1, Ω_2 be bounded intervals in $\mathbb{R}$. For $x \in \Omega_1$, let $f[x]$ be the function on Ω_2 given by $f[x](y) = f(x,y) \quad \forall \; y \in \Omega_2$. If f is integrable on $\Omega_1 \times \Omega_2$, then $f[x]$ is integrable on Ω_2 for almost all $x \in \Omega_1$, $F(x) := \int_{\Omega_2} f[x] \; d\lambda$ is integrable on Ω_1 and

$$\int_{\Omega_1 \times \Omega_2} f \; d\mu = \int_{\Omega_1} F \; d\lambda \, .$$

Proof.
This is a special case of the theorem of Fubini. [10, 11]

□

We use the Sobolev embedding theorem to give a relation between the maximum norm and the norms on $W^{2,1}(\Omega)$ and $W^{2,2}(\Omega)$ if Ω is a bounded interval.

Lemma 2.4.
If Ω is a bounded interval in $\mathbb{R}$, then there are $C, \tilde{C} > 0$, such that

$$\|u\|_{L^\infty(\Omega)} \leqslant C\|u\|_{H^1(\Omega)} \quad \forall\ u \in H^1(\Omega)\ ,$$
$$\|u\|_{L^\infty(\Omega)} \leqslant \tilde{C}\|u\|_{W^{2,1}(\Omega)} \quad \forall\ u \in W^{2,1}(\Omega)\ .$$

Proof.
The Sobolev embedding theorem implies these are linear continuous maps, see e.g. Gilbarg and Trudinger, theorem 7.10 and Corollary 7.11 [13]. □

The next lemma gives an error estimate for our special two dimensional quadrature rule. To obtain this estimate, we use that our weight functions (i.e. the basis functions $\boldsymbol{\eta}$) are essentially one dimensional. We also use that the values for σ_h can be interpreted as averages over cell edges and that we can define these averages for σ, if σ is smooth enough.

Lemma 2.5.
Let Ω_1, Ω_2 and Ω_3 be bounded intervals in $\mathbb{R}$, with $\Omega_1 \subset \Omega_2$ and $\rho = \lambda(\Omega_2) < \infty$, the length of Ω_2. Furthermore, let $x_1, x_2, \ldots, x_{2k+1} \in \Omega_2$, let $w \in L^\infty(\Omega_1)$, $w_1, w_2, \ldots, w_{2k+1} \in \mathbb{R}$, with $\sum_{i=1}^{2k+1} |w_i| \leqslant K\rho$, K independent of ρ, and let $n \geqslant 0$. Set

$$G(u) := \int_{\Omega_1} wu\ d\lambda - \sum_{j=1}^{2k+1} w_j u(x_j) \quad \forall\ u \in W^{n+1,1}\ .$$

If

$$G(u) \equiv 0 \quad \forall\ u \in \{\ 1, x, \ldots, x^n\ \}\ ,$$

$f \in C(\overline{\Omega_2 \times \Omega_3})$, $f[y] \in W^{n+1,1}$ $\forall\ y \in \Omega_3$, where we have $f[y](x) := f(x,y)$ $\forall\ x \in \Omega_2$, and $\partial^{n+1}f/\partial x^{n+1} \in L^1(\Omega_2 \times \Omega_3)$ then

$$\left|\int_{\Omega_3} G(f)\ d\mu\right| \leqslant C\rho^{n+3/2}\ \|\partial^{n+1}f/\partial x^{n+1}\|_{L^2(\Omega_2 \times \Omega_3)}\ ,$$

with C independent of ρ.
Proof.
The only difficulty here is the need for a ρ-independent bound on the values of u in the nodes. We address this problem as follows. On the unit interval, the Sobolev embedding theorem implies the following inequality,

$$\|u\|_{L^\infty((0,1))} \leqslant \tilde{C}\|u\|_{H^1((0,1))} = \tilde{C}\|u\|_{1,2,(0,1)}\ ,$$

for a fixed $\tilde{C} \in \mathbb{R}$. Moreover, if $u \in H^1(\Omega)$ with $\Omega = (a,b)$ with $b - a = \rho$, then $Mu := u(\xi\rho + a)$ is an element of $H^1((0,1))$. What we need now is a relation between

$$\|Mu\|_{1,2,(0,1)}\ ,$$

and

$$\|u\|_{1,2,\Omega} \,.$$

We see immediately that,

$$\|Mu\|_{1,2,(0,1)} = \|u\|_{1,2,\Omega} \,.$$

So,

$$\|u\|_{L^\infty(\Omega)} \leqslant \tilde{C}\|u\|_{1,2,\Omega} \,,$$

with $\tilde{C}$ independent of ρ. This implies,

$$|\rho^{-1}G(u)| \leqslant$$

$$\rho^{-1}\|u\|_{L^2(\Omega)}\|w\|_{L^2(\Omega)} + \rho^{-1}\tilde{C}\sum_{i=1}^{2k+1}|w_i|\,\|u\|_{1,2,\Omega} \quad \forall \ u \in H^1(\Omega_2) \,,$$

or in the norm used by Bramble and Hilbert,

$$|\rho^{-1}G(u)| \leqslant (\|w\|_{L^\infty(\Omega)}\|u\|_{2,\Omega} + K\tilde{C}\|u\|_{1,2,\Omega}) \,.$$

So,

$$|\rho^{-1}G(u)| \leqslant (\|w\|_{L^\infty(\Omega)} + K\tilde{C})\|u\|_{1,2,\Omega} \leqslant C\|u\|_{k,2,\Omega} \,.$$

Fubini implies,

$$|\int_{\Omega_1\times\Omega_3} w(x)f(x,y)d\mu - \sum_{j=1}^{2k+1} w_j \int_{\Omega_3} f(x_j,y)\,dy| = |\int_{\Omega_3} G(f)\,dy| \,.$$

We combine this and find, that there exists a $C > 0$, such that

$$|\int_{\Omega_3} G(f)\,dy| \leqslant C\rho^{n+3/2}\lambda(\Omega_3)^{1/2}\,\|\partial^{n+1}f/\partial x^{n+1}\|_{L^2(\Omega_2\times\Omega_3)} \,.$$

This follows immediately from Cauchy-Schwartz and lemma 2.

□

In the above lemma, G corresponds to the error for a one dimensional integration rule. Next, we relate the condition on G to the coefficients from (24a-f).

Lemma 2.6.
If $f \in H^3([-h,\tilde{h}])$, A,B and C are given by (24a-c) and

$$G(f) := L\int_{-h}^{0} f(x)\frac{h+x}{h}\,dx + R\int_{0}^{\tilde{h}} f(x)\frac{\tilde{h}-x}{\tilde{h}}\,dx - \tag{2.29}$$

$$\left[A(h,\tilde{h},L,R)f(-h) + B(h,\tilde{h},L,R)f(0) + C(h,\tilde{h},L,R)f(\tilde{h})\right] ,$$

then

$$G(p) \equiv 0 \quad \forall \ p \in \{\, 1,x,x^2 \,\} \,.$$

Proof.

This can be proved by direct substitution of the appropriate functions in $G(p)$.

□

Lemma 2.7.
If $f \in C^3([-h,\tilde{h}])$, D,E and F are defined by (24d-f) and

$$G(f) := L \int_{-h}^{0} f(x)\frac{-x}{h}\, dx \; + \; R\int_{0}^{\tilde{h}} f(x)\frac{x}{\tilde{h}}\, dx \; - \qquad (2.30)$$
$$\left[D(h,\tilde{h},L,R)f(-h) \; + \; E(h,\tilde{h},L,R)f(0) \; + \; F(h,\tilde{h},L,R)f(\tilde{h})\right] ,$$

then

$$G(p) \equiv 0 \quad \forall \; p \in \{ \; 1,x,x^2 \; \} \; .$$

Proof.
This is proved as in the previous lemma.

□

Lemma 5, 6 and 7 show that we can find a quadrature rule for $\alpha_h(.,.)$ that is $O(h^3)$. If $\tilde{h} = h$ and $L = R$, then we gain an additional order h for the rule with coefficients A, B and C,

Lemma 2.8.
If $f \in C^4([-h,h])$,

$$G(f) := \int_{-h}^{0} f(x)\frac{h+x}{h}\, dx \; + \; \int_{0}^{h} f(x)\frac{h-x}{h}\, dx \; - \qquad (2.31)$$
$$\left[\frac{h}{12}f(-h) \; + \; \frac{10h}{12}f(0) \; + \; \frac{h}{12}f(h)\right]$$

then

$$G(p) \equiv 0 \quad \forall \; p \in \{ \; 1,x,x^2,x^3 \; \} \; .$$

Proof.
Again, this is proved by calculating $G(p)$ for the appropriate functions.

□

2.5.2. A special norm on V_h.

The space V_h is a finite dimensional vector space. Its natural norm is the Euclidean vector norm. For later use, we introduce $||.||_{V_h}$, a weighted version of the Euclidean vector norm on V_h and we prove, that this norm is equivalent with the $L^2(\Omega)$ norm. If $\sigma_h \in V_h$ and

$$\sigma_h = \sum_{i=0}^{N_1} \sum_{j=0}^{N_2-1} s_{1,i,j+½} \boldsymbol{\eta}_{i,j+½} + \sum_{i=0}^{N_1-1} \sum_{j=0}^{N_2} s_{2,i+½,j} \boldsymbol{\eta}_{i+½,j} \ , \tag{2.32}$$

then we define $\|.\|_{V_h}$ as,

$$\|\sigma_h\|^2_{V_h} = \tag{2.33}$$

$$\sum_{i=0}^{N_1-1} \sum_{j=0}^{N_2-1} ½\mu(\Omega_{i+½,j+½})(s^2_{1,i,j+½} + s^2_{1,i+1,j+½} + s^2_{2,i+½,j} + s^2_{2,i+½,j+1}) \ .$$

Lemma 2.9.
For the σ_h as given in (32),

$$\frac{\|\sigma_h\|^2_{V_h}}{3} \leqslant \|\sigma_h\|^2_{L^2(\Omega)} \leqslant \|\sigma_h\|^2_{V_h} \ , \tag{2.34a}$$

and

$$\mu(\Omega_{i+½,j+½}) \|\sigma_h\|^2_{L^\infty(\Omega_{i+½,j+½})^2} \leqslant 2\|\sigma_h\|^2_{V_h} \ . \tag{2.34b}$$

Proof.
For both norms, we have

$$\|\sigma_h\|^2 = \|(\sigma_h \cdot \mathbf{e}_1)\mathbf{e}_1\|^2 + \|(\sigma_h \cdot \mathbf{e}_2)\mathbf{e}_2\|^2 \ ,$$

so it suffices to prove the inequalities for a single component of σ_h. Furthermore, we know, that

$$\|(\sigma_h \cdot \mathbf{e}_1)\mathbf{e}_1\|^2_{L^2(\Omega)} = \sum_{i=0}^{N_1-1} \sum_{j=0}^{N_2-1} \|(\sigma_h \cdot \mathbf{e}_1)\mathbf{e}_1\|^2_{L^2(\Omega_{i+½,j+½})} \ .$$

We compare terms for corresponding cells,

$$\|(\sigma_h \cdot \mathbf{e}_1)\mathbf{e}_1\|^2_{L^2(\Omega_{i+½,j+½})} = \int_{\Omega_{i+½,j+½}} \left(s_{1,i,j+½} \left[1 - \frac{x_1 - x_{1,i}}{h_{1,i+½}}\right] + s_{1,i+1,j+½} \frac{x_1 - x_{1,i}}{h_{1,i+½}} \right]^2 \ .$$

The contribution of

$$\|(\sigma_h \cdot \mathbf{e}_1)\mathbf{e}_1\|^2_{V_h}$$

for this cell is,

$$½\mu(\Omega_{i+½,j+½})(s^2_{1,i,j+½} + s^2_{1,i+1,j+½}) \ .$$

The inequalities in (34a) now follow from,

$$\int_0^1 (a\xi + b[1-\xi])^2 \, d\xi = \frac{a^2+b^2}{3} + \frac{2ab}{6} \ ,$$

and

$$\frac{a^2+b^2}{6} \leqslant \frac{a^2+b^2}{3}+\frac{2ab}{6} \leqslant \frac{a^2+b^2}{2} \,.$$

Inequality (34b) is trivial.

□

2.5.3. The error estimate for the modified method.

In theorem 1, we give an estimate for $\| \Pi_h\sigma - \sigma_h \|_{L^2(\Omega)}$ and in theorem 2, we give an estimate for $\| P_h u - u_h \|_{L^2(\Omega)}$.

Theorem 2.1.
We define,

$$h_1 = \max_i h_{1,i+½} \,,$$
$$h_2 = \max_j h_{2,j+½} \,,$$

If we assume, that conditions C1 to C3 hold, then

$$\| \Pi_h\sigma - \sigma_h \|^2_{L^2(\Omega)} + \| \sqrt{c}(P_h u - u_h) \|^2_{L^2(\Omega)} \leqslant \tag{2.35a}$$
$$K(h_1+h_2)^3 \max(\| \frac{\partial^3\sigma}{\partial x^3} \|_{L^\infty(\Omega)} , \| \frac{\partial^3\sigma}{\partial y^3} \|_{L^\infty(\Omega)})$$
$$\left[\| \Pi_h\sigma - \sigma_h \|_{L^2(\Omega)} + (h_1+h_2) \| (\Pi_h\sigma - \sigma_h)\cdot\mathbf{n}_{\partial\Omega} \|_{L^2(\partial\Omega)} \right] \,,$$

and

$$\| \Pi_h\sigma - \sigma_h \|^2_{L^2(\Omega)} + \| \sqrt{c}(P_h u - u_h) \|^2_{L^2(\Omega)} \leqslant \tag{2.35b}$$
$$K(h_1+h_2)^3 \max \left[\| \frac{\partial^3\sigma}{\partial x^3} \|_{L^\infty(\Omega)} , \| \frac{\partial^3\sigma}{\partial y^3} \|_{L^\infty(\Omega)} \right] \| \Pi_h\sigma - \sigma_h \|_{L^2(\Omega)} \,.$$

Proof.
Condition C3 implies, that

$$A_0(\Pi_h\sigma - \sigma_h, \Pi_h\sigma - \sigma_h) \leqslant \alpha_h(\Pi_h\sigma - \sigma_h, \Pi_h\sigma - \sigma_h) \,.$$

If we set $\tau = \tau_h$ and $t = t_h$ in (19) and combine the resulting formulas with (21), we get,

$$\alpha_h(\Pi_h\sigma - \sigma_h, \tau_h) - (\text{ div } \tau_h, u - u_h) = \tag{2.36a}$$
$$\alpha_h(\Pi_h\sigma, \tau_h) - \alpha(\sigma, \tau_h) \quad \forall \quad \tau_h \in \tilde{V}_h \,,$$

$$(\text{ div } (\sigma - \sigma_h), t_h) + (c(u - u_h), t_h) = 0 \quad \forall \quad t_h \in W_h \,. \tag{2.36b}$$

If we take into account (22) and the properties of P_h and Π_h from lemma 1, then we find,

$$\alpha_h(\Pi_h\sigma - \sigma_h, \tau_h) - (\text{ div } \tau_h, P_h u - u_h) = \tag{2.37a}$$

$$\alpha_h(\sigma,\tau_h) - \alpha(\sigma,\tau_h) \quad \forall \ \tau_h \in \tilde{V}_h \ ,$$

$$(\ \mathrm{div}\ (\Pi_h\sigma-\sigma_h),t_h) + (c(\mathrm{P}_h u-u_h),t_h) = 0 \quad \forall \ t_h \in W_h \ . \qquad (2.37b)$$

If we set $\tau_h = \Pi_h\sigma-\sigma_h$, $t_h = \mathrm{P}_h u-u_h$, then we find

$$\alpha_h(\Pi_h\sigma-\sigma_h,\Pi_h\sigma-\sigma_h) + (c(\mathrm{P}_h u-u_h),\mathrm{P}_h u-u_h)$$
$$= \alpha(\sigma,\Pi_h\sigma-\sigma_h) - \alpha_h(\sigma,\Pi_h\sigma-\sigma_h) \ .$$

by adding (37b) to (37a).

We introduce

$$K_E = \{ \ (i,j-½) \mid i=1,\ldots,N_1 \ , j=1,\ldots,N_2 \ \} \ , \qquad (2.38a)$$

$$K_N = \{ \ (i-½,j) \mid i=1,\ldots,N_1 \ , j=1,\ldots,N_2 \ \} \ , \qquad (2.38b)$$

$$K_W = \{ \ (i,j-½) \mid i=0,\ldots,N_1-1 \ , j=1,\ldots,N_2 \ \} \ , \qquad (2.38c)$$

$$K_S = \{ \ (i-½,j) \mid i=1,\ldots,N_1 \ , j=0,\ldots,N_2-1 \ \} \ . \qquad (2.38d)$$

The measure of the support of $\boldsymbol{\eta}_k$ is denoted by $\mu(Supp(\boldsymbol{\eta}_k))$. We denote the length of the support in the $\mathbf{e}_\kappa$ direction by $\lambda_\kappa(Supp(\boldsymbol{\eta}_k))$. If A and B are sets, we use,

$$A\Delta B = (B-A)\bigcup(A-B) \ ,$$

(the symmetric set difference).

If we combine lemma 5 with lemma 6, lemma 7 and (C3), we find,

$$\alpha(\sigma,\Pi_h\sigma-\sigma_h) - \alpha_h(\sigma,\Pi_h\sigma-\sigma_h) \leqslant$$

$$\sum_{k \in (K_W \bigcap K_E) \bigcup (K_W \Delta K_E)} |\,P[\Gamma_k](\Pi_h\sigma-\sigma_h)\cdot\mathbf{e}_1\,|\,(\alpha(\sigma,\boldsymbol{\eta}_k)-\alpha_h(\sigma,\boldsymbol{\eta}_k)) \ +$$

$$\sum_{m \in (K_N \bigcap K_S) \bigcup (K_N \Delta K_S)} |\,P[\Gamma_m](\Pi_h\sigma-\sigma_h)\cdot\mathbf{e}_2\,|\,(\alpha(\sigma,\boldsymbol{\eta}_m)-\alpha_h(\sigma,\boldsymbol{\eta}_m)) \ \leqslant$$

$$C \sum_{\substack{k \in (K_W \bigcap K_E) \bigcup \\ (K_W \Delta K_E)}} |\,P[\Gamma_k]((\Pi_h\sigma-\sigma_h)\cdot\mathbf{e}_1)\,|\,\mu(Supp(\boldsymbol{\eta}_k))\lambda_1(Supp(\boldsymbol{\eta}_k))^3 \,\|\, \frac{\partial^3\sigma}{\partial x^3} \,\|_{\mathrm{L}^\infty(\Omega)} \ +$$

$$+ \ C \sum_{\substack{m \in (K_N \bigcap K_S) \bigcup \\ (K_N \Delta K_S)}} |\,P[\Gamma_m]((\Pi_h\sigma-\sigma_h)\cdot\mathbf{e}_2)\,|\,\mu(Supp(\boldsymbol{\eta}_m))\lambda_2(Supp(\boldsymbol{\eta}_m))^3 \,\|\, \frac{\partial^3\sigma}{\partial y^3} \,\|_{\mathrm{L}^\infty(\Omega)} \ .$$

From this formula we can derive (35a) and (35b). We start by deriving (35a),

$$\alpha(\sigma,\Pi_h\sigma-\sigma_h)-\alpha_h(\sigma,\Pi_h\sigma-\sigma_h) \leqslant$$

$$C \left\{ 2\mu(\Omega) \sum_{k \in K_W \bigcap K_E} \mu(Supp(\boldsymbol{\eta}_k))P[\Gamma_k]((\Pi_h\sigma-\sigma_h)\cdot\mathbf{e}_1)^2 \right\}^{½} 8h_1^3 \,\|\, \frac{\partial^3\sigma}{\partial x^3} \,\|_{\mathrm{L}^\infty(\Omega)} \ +$$

$$C \left\{ 2\lambda_1(\Omega) \sum_{k \in (K_W \Delta K_E)} \lambda_1(\Gamma_k)P[\Gamma_k]((\Pi_h\sigma-\sigma_h)\cdot\mathbf{e}_1)^2 \right\}^{½} 8h_1^3 h_2 \,\|\, \frac{\partial^3\sigma}{\partial x^3} \,\|_{\mathrm{L}^\infty(\Omega)}$$

$$+\ C\left\{2\mu(\Omega)\sum_{m\,\in\,K_N\bigcap K_S}\mu(Supp(\boldsymbol{\eta}_k))P[\Gamma_m]((\Pi_h\sigma-\sigma_h)\cdot\mathbf{e}_2)^2\right]^{1/2}8h_2^3\,||\,\frac{\partial^3\sigma}{\partial y^3}\,||_{\,\mathrm{L}^\infty(\Omega)}\ +$$

$$C\left\{2\lambda_2(\Omega)\sum_{m\,\in\,(K_N\Delta K_S)}\lambda_2(\Gamma_k)P[\Gamma_m]((\Pi_h\sigma-\sigma_h)\cdot\mathbf{e}_2)^2\right]^{1/2}8h_1h_2^3\,||\,\frac{\partial^3\sigma}{\partial y^3}\,||_{\,\mathrm{L}^\infty(\Omega)}\ \leqslant$$

$$\tilde{C}\,||\,(\Pi_h\sigma-\sigma_h)\,||_{\,\mathrm{L}^2(\Omega)}\,(h_1+h_2)^3\left[\,||\,\frac{\partial^3\sigma}{\partial x^3}\,||_{\,\mathrm{L}^\infty(\Omega)}\ +\ ||\,\frac{\partial^3\sigma}{\partial y^3}\,||_{\,\mathrm{L}^\infty(\Omega)}\right]\ +$$

$$\tilde{C}\,||\,(\Pi_h\sigma-\sigma_h)\cdot\mathbf{n}_{\partial\Omega}\,||_{\,\mathrm{L}^2(\partial\Omega)}(h_1+h_2)^4\left[\,||\,\frac{\partial^3\sigma}{\partial x^3}\,||_{\,\mathrm{L}^\infty(\Omega)}\ +\ ||\,\frac{\partial^3\sigma}{\partial y^3}\,||_{\,\mathrm{L}^\infty(\Omega)}\right]\ .$$

Here, we used the equivalence proved in lemma 9. Next, we derive (35b),

$$\alpha(\sigma,\Pi_h\sigma-\sigma_h)-\alpha_h(\sigma,\Pi_h\sigma-\sigma_h)\ \leqslant$$

$$C\left\{2\mu(\Omega)\sum_{k\,\in\,K_W\bigcap K_E}\mu(Supp(\boldsymbol{\eta}_k))P[\Gamma_k]((\Pi_h\sigma-\sigma_h)\cdot\mathbf{e}_1)^2\right]^{1/2}8h_1^3\,||\,\frac{\partial^3\sigma}{\partial x^3}\,||_{\,\mathrm{L}^\infty(\Omega)}\ +$$

$$C\left\{2\lambda_1(\Omega)\sum_{k\,\in\,(K_W\Delta K_E)}\mu(Supp(\boldsymbol{\eta}_k))P[\Gamma_k]((\Pi_h\sigma-\sigma_h)\cdot\mathbf{e}_1)^2\right]^{1/2}8h_1^3h_2^{1/2}\,||\,\frac{\partial^3\sigma}{\partial x^3}\,||_{\,\mathrm{L}^\infty(\Omega)}$$

$$+\ C\left\{2\mu(\Omega)\sum_{m\,\in\,K_N\bigcap K_S}\mu(Supp(\boldsymbol{\eta}_m))P[\Gamma_m]((\Pi_h\sigma-\sigma_h)\cdot\mathbf{e}_2)^2\right]^{1/2}8h_2^3\,||\,\frac{\partial^3\sigma}{\partial y^3}\,||_{\,\mathrm{L}^\infty(\Omega)}\ +$$

$$C\left\{2\lambda_2(\Omega)\sum_{m\,\in\,(K_N\Delta K_S)}\mu(Supp(\boldsymbol{\eta}_m))P[\Gamma_m]((\Pi_h\sigma-\sigma_h)\cdot\mathbf{e}_2)^2\right]^{1/2}8h_1^{1/2}h_2^3\,||\,\frac{\partial^3\sigma}{\partial x^3}\,||_{\,\mathrm{L}^\infty(\Omega)}\ \leqslant$$

$$\tilde{C}\left\{\,||\,(\Pi_h\sigma-\sigma_h)\,||_{\,\mathrm{L}^2(\Omega)}\,(h_1+h_2)^3\left[\,||\,\frac{\partial^3\sigma}{\partial x^3}\,||_{\,\mathrm{L}^\infty(\Omega)}\ +\ ||\,\frac{\partial^3\sigma}{\partial y^3}\,||_{\,\mathrm{L}^\infty(\Omega)}\right]\right]\ .$$

Again, we used the equivalence proved in lemma 9.

□

For cells in areas of constant a and uniform mesh-size, the proof of lemma 8 implies that their contribution to the global error is of order h^4. If the areas of constant a and uniform mesh-size are large enough, we treat the cells adjacent to the boundaries of such areas in the same way as the cells adjacent to the boundaries of Ω, this results in an $O(h^{3\frac{1}{2}})$ error. If furthermore,

$$||\,(\Pi_h\sigma-\sigma_h)\cdot\mathbf{n}_{\partial\Omega}\,||_{\,\mathrm{L}^2(\partial A\,\bigcup\,\partial\Omega)}\ \leqslant\ ||\,\Pi_h\sigma-\sigma_h\,||_{\,\mathrm{L}^2(\Omega)}\ ,$$

where ∂A is the union of edges between areas of constant a and uniform mesh-size, then formula (35a) gives us an $O(h^4)$ error estimate. These effects are seen in our numerical results.

Next, we express $||\,\mathrm{P}_h u-u_h\,||_{\,\mathrm{L}^2(\Omega)}$ in terms of $||\,\Pi_h\sigma-\sigma_h\,||_{\,\mathrm{L}^2(\Omega)}$.

Theorem 2.2.
Take h_1 and h_2 as in theorem 1. Under the conditions C1, C2 and C3, we have

$$\| \mathrm{P}_h u - u_h \|_{\mathrm{L}^2(\Omega)} \leqslant \tag{2.39}$$
$$K \left(\| \Pi_h \sigma - \sigma_h \|_{\mathrm{L}^2(\Omega)} + h_1^3 \left\| \frac{\partial^3 \sigma_1}{\partial x^3} \right\|_{\mathrm{L}^\infty(\Omega)} + 2h_1^4 \left\| \frac{\partial^3 \sigma_1}{\partial x^3} \right\|_{\mathrm{L}^\infty(\Omega)} \right] .$$

Proof.
To obtain this estimate, we examine $\mathrm{P}_h u - u_h$ for each subdomain separately. We use the following relation, which can be obtained from (19) and (21),

$$\alpha(\sigma,\tau_h) - \alpha_h(\sigma_h,\tau_h) - (\operatorname{div} \tau_h, \mathrm{P}_h u - u_h) = 0 \quad \forall \ \tau_h \in V_h \ .$$

When combined with (22) this implies,

$$(\operatorname{div} \tau_h, \mathrm{P}_h u - u_h) = \alpha_h(\sigma,\tau_h) - \alpha(\sigma,\tau_h) + \alpha_h(\sigma_h - \Pi_h\sigma,\tau_h) \quad \forall \ \tau_h \in V_h \ . \quad \text{(A)}$$

We concentrate for the moment on the sub-domain $\Omega_{i+½,j+½}$. We define a special τ_h,

$$\tau_{h,1} = \left\{ \begin{array}{l} 0 \ \text{ on } \ \Omega_{k+½,l+½} \ \text{ if } \ l < j \ \text{ or } \ l > j \ , \\ \quad 0 \ \text{ on } \ \Omega_{k+½,j+½} \ \text{ if } \ k < i \ , \\ \quad 1 \ \text{ on } \ \Omega_{k+½,j+½} \ \text{ if } \ k > i \\ \quad \dfrac{x_1 - x_{1,i}}{h_{1,i+½}} \ \text{ on } \ \Omega_{i+½,j+½} \end{array} \right. \tag{2.40a}$$

$$\tau_{h,2} = 0 \ \text{ on } \ \Omega \ . \tag{2.40b}$$

Substituting this for τ_h, we find,

$$h_{2,j} \ \| \mathrm{P}_h u - u_h \|_{\mathrm{L}^\infty(\Omega_{i+½,j+½})} \leqslant$$
$$C \left(h_{2,j+½} \left\| \frac{\partial^3 \sigma_1}{\partial x^3} \right\|_{\mathrm{L}^\infty(\Omega)} \left[\sum_{k=0}^{N_1-1} h_{1,k+½}^4 + h_{1,½}^4 + h_{1,N_1-½}^4 \right] + \right.$$
$$\left. \sum_{k=0}^{N_1-1} \mu(\Omega_{k+½,j+½}) \| \Pi_h \sigma - \sigma_h \|_{\mathrm{L}^\infty(\Omega_{k+½,j+½})} \right] .$$

The first term in the right hand side of this inequality corresponds with the quadrature error in (A) in the interior and on the edge respectively, the second term corresponds with the remaining term in (A). So,

$$\| \mathrm{P}_h u - u_h \|_{\mathrm{L}^\infty(\Omega_{i+½,j+½})} \leqslant$$
$$C \left((h_1^3 + 2h_1^4) \left\| \frac{\partial^3 \sigma_1}{\partial x^3} \right\|_{\mathrm{L}^\infty(\Omega)} + \frac{1}{h_{2,j}} \sum_{k=0}^{N_1-1} \mu(\Omega_{k+½,j+½}) \| \Pi_h \sigma - \sigma_h \|_{\mathrm{L}^\infty(\Omega_{k+½,j+½})} \right] ,$$

where we used that $\mathrm{P}_h u - u_h$ is constant on $\Omega_{i+½,j+½}$, Cauchy-Schwartz and

(35b).

We multiply both sides of this equation by the square root of the area of the cell,

$$\mu(\Omega_{i+½,j+½})^{½} \| \mathrm{P}_h u - u_h \|_{\mathrm{L}^\infty(\Omega_{i+½,j+½})} = \| \mathrm{P}_h u - u_h \|_{\mathrm{L}^2(\Omega_{i+½,j+½})} \leqslant$$
$$\mu(\Omega_{i+½,j+½})^{½} C \left\{ \left[h_1^3 + 2h_1^4 \right] \| \frac{\partial^3 \sigma_1}{\partial x^3} \|_{\mathrm{L}^\infty(\Omega)} + \right.$$
$$\left. \frac{1}{h_{2,j}} \sum_{k=0}^{N_1-1} \mu(\Omega_{k+½,j+½}) \| \Pi_h \sigma - \sigma_h \|_{\mathrm{L}^\infty(\Omega_{k+½,j+½})} \right\} .$$

If we square the left and right hand sides and then sum over i and j, we find,

$$\| \mathrm{P}_h u - u_h \|^2_{\mathrm{L}^2(\Omega)} \leqslant$$
$$K \left\{ h_1^6 \| \frac{\partial^3 \sigma_1}{\partial x^3} \|^2_{\mathrm{L}^\infty(\Omega)} + 2h_1^8 \| \frac{\partial^3 \sigma_1}{\partial x^3} \|^2_{\mathrm{L}^\infty(\Omega)} + \| \Pi_h \sigma - \sigma_h \|^2_{\mathrm{L}^2(\Omega)} \right\} .$$

□

Again, if the conditions following the proof of theorem 1 hold, then we gain an additional order of h, because in that case $\| \Pi_h \sigma - \sigma_h \|_{\mathrm{L}^2(\Omega)}$ is $O(h^4)$ and we can replace the term $h_1^3 \| \partial^3 \sigma_1 / \partial x_1^3 \|_{\mathrm{L}^\infty(\Omega)}$, that represents the quadrature error, by $h_1^4 \| \partial^4 \sigma_1 / \partial x_1^4 \|_{\mathrm{L}^\infty(\Omega)}$.

If, in the above proofs, we replace the explicit expression for the local quadrature error by a more general form, we see, that the order of the error is equal to the order of the quadrature rule used.

2.5.4. A proof of condition (C3) on a uniform mesh with constant a.

We show that, on a uniform mesh, α_h satisfies condition (C3) if a is constant. Without loss of generality we take $a \equiv 1$.

Lemma 2.10.
Assume $a \equiv 1$. If the mesh is uniform, then

$$\| \sigma_h \|^2_{V_h} \leqslant \frac{48}{5} \alpha_h(\sigma_h, \sigma_h) \leqslant \frac{96}{5} \| \sigma_h \|^2_{V_h} .$$

Proof.
If we write σ_h as a linear combination of basis functions $\boldsymbol{\eta}$,

$$\sigma_h = \sum_{m \in K_W \bigcup K_E} s_{1,m} \boldsymbol{\eta}_m + \sum_{m \in K_N \bigcup K_S} s_{2,m} \boldsymbol{\eta}_m ,$$

then we find,

$$\alpha_h(\sigma_h, \sigma_h) =$$

$$\alpha_h \left[\sum_{k \in K_W \bigcup K_E} s_{1,k}\boldsymbol{\eta}_k \, , \sum_{m \in K_W \bigcup K_E} s_{1,m}\boldsymbol{\eta}_m \right] + \qquad (2.41)$$
$$\alpha_h \left[\sum_{k \in K_N \bigcup K_S} s_{2,k}\boldsymbol{\eta}_k \, , \sum_{m \in K_N \bigcup K_S} s_{2,m}\boldsymbol{\eta}_m \right] .$$

where K_N etc. are defined in (38). For the term in (41) corresponding to the e_1 component, we find:

$$\alpha_h \left[\sum_{k \in K_W \bigcup K_E} s_{1,k}\boldsymbol{\eta}_k \, , \sum_{m \in K_W \bigcup K_E} s_{1,m}\boldsymbol{\eta}_m \right] =$$
$$h_1 h_2 \sum_{m \in K_W \bigcap K_E} s_{1,m} \left[\frac{1}{12} s_{1,m-(1,0)} + \frac{10}{12} s_{1,m} + \frac{1}{12} s_{1,m+(1,0)} \right] +$$
$$h_1 h_2 \sum_{m \in K_W - K_E} s_{1,m} \left[\frac{7}{24} s_{1,m} + \frac{6}{24} s_{1,m+(1,0)} + \frac{-1}{24} s_{1,m+(2,0)} \right] +$$
$$h_1 h_2 \sum_{m \in K_E - K_W} s_{1,m} \left[\frac{-1}{24} s_{1,m-(2,0)} + \frac{6}{24} s_{1,m-(1,0)} + \frac{7}{24} s_{1,m} \right] ,$$

where $m-(1,0) = (i-1,j-½)$ if $m = (i,j-½)$ etc.

Next, we interpret the coefficients $s_{1,m}$ with $m \in K_W \bigcup K_E$ as a vector $\mathbf{s}$ in $\mathbb{R}^{(N_1+1)N_2}$. We introduce the notation $\mathbf{f}_{1,m}$ for the unit vector along the coordinate axis corresponding to $s_{1,m}$. We define the matrix A by,

$$\mathbf{f}_{1,k}^T A \mathbf{f}_{1,m} = \alpha_h(\boldsymbol{\eta}_{1,k}, \boldsymbol{\eta}_{1,m}) .$$

We can write $\alpha_h(\sigma_h,\sigma_h)$ as follows,

$$\mathbf{s}^T A \mathbf{s} = ½\mathbf{s}^T(A+A^T)\mathbf{s} .$$

According to the fundamental theorem on symmetric matrices, this implies that all eigenvalues of $A+A^T$ are real and that,

$$A+A^T = O^T D O ,$$

where O is an orthogonal matrix and D is a diagonal matrix with as diagonal elements the eigenvalues of A. Gershgorin's theorem implies that all eigenvalues are larger than

$$½h_1h_2\left(\frac{14}{24} - \frac{8}{24} - \frac{1}{24}\right) = \frac{5}{48}h_1h_2 ,$$

and smaller than

$$h_1h_2\left(\frac{10}{12} + \frac{1}{12} + \frac{1}{12}\right) = h_1h_2 .$$

The same reasoning can be applied to the $\mathbf{e}_2$ component of σ_h. We find,

$$\alpha_h(\sigma_h,\sigma_h) = ½\mathbf{s}^T(A+A^T)\mathbf{s} = ½\mathbf{s}^T O^T D O \mathbf{s} \geqslant \frac{5}{48} \| \mathbf{s} \|_{V_h}^2 .$$

□

Lemma 2.11.
For a constant coefficient a, the bilinear form α_h satisfies condition (C3).
Proof.
This follows immediately from lemma 9 and lemma 10.

□

2.6 The effect of a non-zero c.

We use a one-dimensional example to illustrate the problems associated with a zero order term mentioned in the introduction (cf. Polak, Schilders and Couperus) [5]. The one dimensional problem is studied, because we can easily obtain the discrete system of equations in u. We see, that, for the quadrature rule given in section 2.4.1, $ch^2/a > 6$ results in the loss of the conditions for the local maximum principle for u_h. For our new quadrature rule, the corresponding bound for satisfying the local maximum principle is $ch^2/a < 12$. As any one-dimensional problem can be trivially extended to an example for two dimensions, the same difficulties will appear in two dimensions.

If we write down our modified discretisation in one dimension on a uniform grid with $a=\epsilon$, $c=1$, $f=0$ and $g(0)=0$, $g(1)=U$, then we find the following system of equations:

$$\frac{7h}{24\epsilon}\sigma_0 + \frac{6h}{24\epsilon}\sigma_1 - \frac{h}{24\epsilon}\sigma_2 + u_{1/2} = 0 \,, \tag{2.42a.0}$$

$$\frac{h}{12\epsilon}\sigma_{i-1} + \frac{10h}{12\epsilon}\sigma_i + \frac{h}{12\epsilon}\sigma_{i+1} - u_{i-1/2} + u_{i+1/2} = 0 \tag{2.42a.i}$$

$$\text{for } i=1,2,...,N-1 \,,$$

$$-\frac{h}{24\epsilon}\sigma_{N-2} + \frac{6h}{24\epsilon}\sigma_{N-1} + \frac{7h}{24\epsilon}\sigma_N - u_{N-1/2} = -U \,, \tag{2.42a.N}$$

$$-\sigma_{i-1}+\sigma_i + hu_{i-1/2} = 0 \text{ for } i=1,2,...,N \,. \tag{2.42b}$$

Elimination of σ yields,

$$3u_{1/2}(1+\frac{4h^2}{24\epsilon}) - (1-\frac{4h^2}{24\epsilon})u_{1+1/2} = 0 \tag{2.43a.1}$$

$$(1-\frac{h^2}{12\epsilon})u_{i-1/2} - 2(1+\frac{5h^2}{12\epsilon})u_{i+1/2} + (1-\frac{h^2}{12\epsilon})u_{i+1+1/2} = 0, \tag{2.43a.i}$$

$$i=1,2,...,N-2 \,,$$

$$-u_{N-1-1/2}(1-\frac{4h^2}{24\epsilon}) + 3u_{N-1/2}(1+\frac{4h^2}{24\epsilon}) = 2U \,. \tag{2.43a.N}$$

We see, that the matrix is always diagonal dominant, but for $h^2/\epsilon > 12$ it is not an M-matrix.

If we use exact integration for Raviart Thomas mixed finite elements, then we find,

$$3u_{1/2}(1+\frac{h^2}{6\epsilon}) - (1-\frac{h^2}{6\epsilon})u_{1+1/2} = 0 \qquad (2.44a.1)$$

$$(1 - \frac{h^2}{6\epsilon})u_{i-1/2} - 2(1 + \frac{h^2}{3\epsilon})u_{i+1/2} + (1 - \frac{h^2}{6\epsilon})u_{i+1+1/2} = 0, \quad (2.44a.i)$$

$$\text{for } i=1,2,...,N-2\ ,$$

$$- u_{N-1-1/2}(1-\frac{h^2}{6\epsilon}) + 3u_{N-1/2}(1+\frac{h^2}{6\epsilon}) = 2U\ . \qquad (2.44a.N)$$

Here we see, that there is no qualitative difference in sensitivity to the ratio $\frac{h^2}{\epsilon}$ between our method and the standard method. However, for the trapezoidal rule we find:

$$3u_{1/2}(1+\frac{h^2}{3\epsilon}) - u_{1+1/2} = 0 \qquad (2.45a.1)$$

$$- u_{i-1/2} + 2(1 + \frac{h^2}{2\epsilon})u_{i+1/2} - u_{i+1+1/2} = 0 \qquad (2.45a.i)$$

$$\text{for } i=1,2,...,N-2\ ,$$

$$- u_{N-1-1/2} + 3u_{N-1/2}(1+\frac{h^2}{3\epsilon}) = 2U\ . \qquad (2.45a.N)$$

In this case, we do get an M-matrix.

We recall from section 2.5, that the accuracy of a method is determined by the accuracy of the quadrature rule used in α_h. If σ is sufficiently smooth, then we find the following orders for the above schemes, $O(h^{3\frac{1}{2}})$ for (42) ($O(h^4)$ if the error is not concentrated at the edges), $O(h^2)$ for (45) and for (44). The latter result may seem strange, because this scheme is based on exact integration of products of test and trial functions. However, by inspection of the formulas, we see that the need to integrate products of continuous piecewise linear functions results in coefficients, that are not optimal for approximate integration of products of smooth functions and continuous, piecewise linear functions.

In scheme (42) and (44), we find the same equations for boundary cells. The equations for boundary cells in (45) however, are different. As (42) is $O(h^{3\frac{1}{2}})$ accurate, the equations for the boundary given by this scheme are more accurate than those given by (45). So, on the same mesh, we expect the error in the boundary cells for scheme (45) to be larger than for scheme (44), but we expect to find the same order behaviour for both schemes. Our experiments confirm this expectation.

2.7 Numerical experiments.

This section gives numerical results for problem (1) on a uniform mesh. We take $c \equiv 0$, $\Omega =$ the unit square and f, g such that

$$u = \frac{(\exp(x - \tfrac{1}{2}) - 1)\,(\exp(y - \tfrac{1}{2}) - 1)}{a},$$

is the solution of the continuous problem. First we give results for $a=1$ on the unit square, then we divide the unit square into four smaller squares and give results for a discontinuous coefficient a, $a=1$ in the lower left square, $a=10$ in the upper left square, $a=100$ in the lower right square and $a=1000$ in the upper right square (Figure 1).

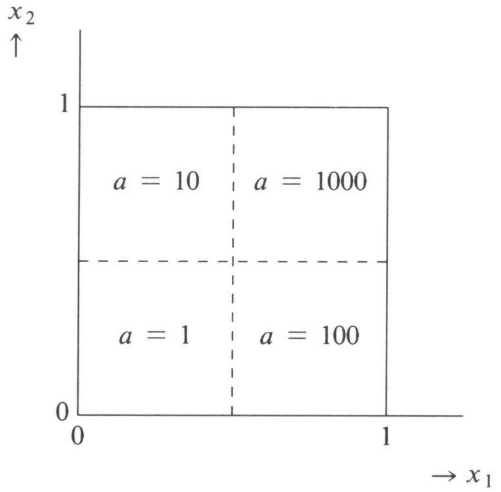

Figure 1.

For u_h, the size of the error is expressed in the $L^2(\Omega)$ norm. For σ_h, the size of the error is expressed as the Euclidean norm in the space of vectors of coefficients of the $\boldsymbol{\eta}$ basis vectors, scaled by the square root of the area of one cell.

We give results for the discretisation from section 2.3 and the two discretisations from section 2.4.
We indicate the quadrature rule used in the discretisation by roman numbers, I denotes the quadrature given in section 2.3.4, number II denotes exact quadrature (section 2.4.1) and discretisation III denotes the trapezoidal rule (section 2.4.2).

h	$\log_2 \| P_h u - u_h \|_E$			$\log_2 \| \Pi_h \sigma - \sigma_h \|_E$		
	I	II	III	I	II	III
1/2	-13.03	-13.03	-7.22	-9.18	-7.43	-5.13
1/4	-16.32	-14.88	-9.13	-12.67	-9.25	-6.42
1/8	-20.06	-16.81	-11.04	-16.41	-11.21	-8.05
1/16	-23.95	-18.79	-12.99	-20.25	-13.21	-9.83
1/32	-27.90	-20.78	-14.97	-24.14	-15.21	-11.67
1/64	-31.87	-22.78	-16.97	-28.07	-17.21	-13.56
1/128	-35.86	-24.78	-18.97	-32.01	-19.21	-15.47
1/256	-39.86	-26.78	-20.97	-35.96	-21.21	-17.39

Table 1. *Errors for the three methods for the constant coefficient case.*

h	$\log_2 \| P_h u - u_h \|_E$			$\log_2 \| \Pi_h \sigma - \sigma_h \|_E$		
	I	II	III	I	II	III
1/2	-16.23	-16.23	-8.23	-9.26	-7.59	-4.91
1/4	-17.89	-16.94	-10.22	-12.53	-9.31	-6.12
1/8	-21.75	-18.72	-12.20	-16.24	-11.26	-7.71
1/16	-25.69	-20.67	-14.18	-20.06	-13.25	-9.47
1/32	-29.67	-22.65	-16.17	-23.94	-15.25	-11.31
1/64	-33.67	-24.65	-18.17	-27.85	-17.25	-13.18
1/128	-37.67	-26.65	-20.17	-31.78	-19.25	-15.08
1/256	-41.67	-28.65	-22.17	-35.72	-21.25	-17.00

Table 2 *Errors for the three methods for the discontinuous coefficient case.*

Starting at $h=1/8$, we see, for case I, convergence of order 4 as predicted in section 2.5.2 for a uniform mesh and large areas with constant coefficients. The other schemes show second order behaviour. We recall, that the error analysis in section 2.5 shows that the accuracy of a method is determined by the accuracy of the quadrature rule α_h applied to σ. Our σ is smooth, so we indeed expect the following orders for the above schemes, $O(h^4)$ for (I), $O(h^2)$ for (II), $O(h^2)$ for (III).

2.8 An a-posteriori error estimate.

We see that there is a difference in order of accuracy between our special method, given in section 2.3.4 and the method based on the use of the trapezoidal rule, given in section 2.4.2. This suggests that the special scheme may be used to obtain an a-posteriori estimate of the error in the solution of the trapezoidal scheme.

In this section, we shall use the following notation, $\alpha_{h,3}$ is the bilinear form we obtain if we use the three point rule given in section 2.3.4 to evaluate α_h and $\alpha_{h,1}$ is the bilinear form we obtain if we use the trapezoidal rule given in section 2.4.2. Furthermore, let $(\boldsymbol{\sigma},u)$ be the solution of problem (19), let $(\boldsymbol{\sigma}_h,u_h)$ be the solution of the discretisation (21) given in section 2.3.3 with

$\alpha_h = \alpha_{h,1}$

and let $(\tilde{\boldsymbol{\sigma}}_h, \tilde{u}_h)$ be the solution of the same discretisation, with $\alpha_h = \alpha_{h,3}$.

The simplest way to obtain an a-posteriori error estimate is to solve both schemes. Given the solution of both schemes, we can obtain estimates for

$$\| \Pi_h \boldsymbol{\sigma} - \boldsymbol{\sigma}_h \|_{H(\mathrm{div},\Omega)} ,$$

and

$$\| P_h u - u_h \|_{L^2(\Omega)} ,$$

as follows, we insert an extra term in the abcve expressions and use the triangle inequality to find,

$$\| \tilde{\boldsymbol{\sigma}}_h - \boldsymbol{\sigma}_h \|_{H(\mathrm{div},\Omega)} - \| \Pi_h \boldsymbol{\sigma} - \tilde{\boldsymbol{\sigma}}_h \|_{H(\mathrm{div},\Omega)} \leqslant \| \Pi_h \boldsymbol{\sigma} - \boldsymbol{\sigma}_h \|_{H(\mathrm{div},\Omega)} \leqslant \| \tilde{\boldsymbol{\sigma}}_h - \boldsymbol{\sigma}_h \|_{H(\mathrm{div},\Omega)} + \| \Pi_h \boldsymbol{\sigma} - \tilde{\boldsymbol{\sigma}}_h \|_{H(\mathrm{div},\Omega)} ,$$

and

$$\| \tilde{u}_h - u_h \|_{L^2(\Omega)} - \| P_h u - \tilde{u}_h \|_{L^2(\Omega)} \leqslant \| P_h u - u_h \|_{L^2(\Omega)} \leqslant \| \tilde{u}_h - u_h \|_{L^2(\Omega)} + \| P_h u - \tilde{u}_h \|_{L^2(\Omega)} .$$

Next, we assume that $\boldsymbol{\sigma}$ is sufficiently smooth and we recall that

$$\| \Pi_h \boldsymbol{\sigma} - \boldsymbol{\sigma}_h \|_{H(\mathrm{div},\Omega)} + \| P_h u - u_h \|_{L^2(\Omega)} = \mathcal{O}(h^k)$$

and

$$\| \Pi_h \boldsymbol{\sigma} - \tilde{\boldsymbol{\sigma}}_h \|_{H(\mathrm{div},\Omega)} + \| P_h u - \tilde{u}_h \|_{L^2(\Omega)} = \mathcal{O}(h^{l+2}) ,$$

where $k,l=2$ if the mesh is uniform and a is constant and otherwise $k=1$ or $2, l=1$ or 2 depending on the mesh and a. This implies, that

$$\| \Pi_h \boldsymbol{\sigma} - \boldsymbol{\sigma}_h \|_{H(\mathrm{div},\Omega)} = (1 + \mathcal{O}(h)) \| \tilde{\boldsymbol{\sigma}}_h - \boldsymbol{\sigma}_h \|_{H(\mathrm{div},\Omega)} ,$$

and

$$\| P_h u - u_h \|_{L^2(\Omega)} = (1 + \mathcal{O}(h)) \| \tilde{u}_h - u_h \|_{L^2(\Omega)} .$$

where h is the maximum cell diameter of the mesh.

2.9 Conclusions.

For equation (1), we have increased the accuracy of the mixed finite element approximation of $(\Pi_h \sigma, P_h u)$ by introducing a particular quadrature rule for $\alpha(\sigma, \tau_h)$. This leads to a scheme, that has the same complexity as standard mixed finite elements for lowest order Raviart-Thomas elements, but that is of $O(h^3)$ in stead of $O(h^2)$ if σ is sufficiently smooth. This behaviour is confirmed by numerical experiments.

In section 2.8, we show that this difference in order can be used to give an a posteriori error estimator for the less accurate version.

If we compare the usual method (section 2.4.1) with the other two methods, we see, that the only advantage of the method given in section 2.4.1 over the method that uses the trapezoidal rule (section 2.4.2) is a better treatment of boundary cells (see the discussion in section 2.6). The only advantage of the method given in section 2.4.1 over our modified method is, that the method from section 2.4.1 may give exact results for less smooth solutions, viz. for solutions with $\sigma \in V_h$.

To decide whether to use the method based on the trapezoidal rule or our modified method, we must weigh the advantage of a simpler matrix, that reduces to an M-matrix for u_h for all $c \geqslant 0$, against the loss of accuracy. The numerical experiments show the loss of accuracy to be considerable for smooth σ. So, only if it is known, that the combination of a, c and h may lead to instability (for instance if $ch^2/a \geqslant 1$), or if σ is not smooth enough, is it more efficient to use the method based on the trapezoidal rule. In all other cases our modified method would be the better choice.

The choice between our method and the method discussed in section 2.4.1 is simple. Both methods are equally sensitive to a zero order term. Both methods also have the same sparsity pattern in their matrices, so they roughly need the same amount of work to solve. As the method in section 2.4.1 is of lower order than the modified method, the modified method is more efficient if we look at accuracy obtained versus complexity.

References

1. P. A. Raviart and J. M. Thomas, "A mixed finite element method for 2-nd order elliptic problems," in *Mathematical aspects of the finite element method*, Lecture Notes in Mathematics, vol. 606, pp. 292-315, Springer, 1977.

2. Mie Nakata, Alan Weiser, and Mary Fanett Wheeler, "Some superconvergence results for mixed finite element methods for elliptic problems on rectangular domains," in *The Mathematics of Finite Elements and Applications*, ed. J. R. Whiteman, vol. 5, pp. 367-389, 1985.

3. J. Douglas, Jr. and J. Wang, "Superconvergence of mixed finite element methods on rectangular domains," *Calcolo*, vol. 26, pp. 121-133, 1989.

4. Junping Wang, "Superconvergence and extrapolation for mixed finite element methods on rectangular domains.," *Math. Comp.*, vol. 56, pp. 477-503, 1991.

5. S. J. Polak, W. H. A. Schilders, and H. D. Couperus, "A finite element method with current conservation," in *Simulation of semiconductor devices and processes*, ed. M. Rudan, vol. 3, pp. 453-462, Tecnoprint, Bologna, 1988.

6. M. Fortin, "An analysis of the convergence of mixed finite element methods," *RAIRO Numerical Analysis*, vol. 11, no. 4, pp. 341-354, 1977.

7. J. Douglas, Jr. and J. E. Roberts, "Global estimates for mixed methods for second order elliptic equations," *Mathematics of computation*, vol. 44, no. 169, pp. 39-52, 1985.

8. Jean E. Roberts and Jean-Marie Thomas, "Mixed and Hybrid Finite Element Methods," RR 737, INRIA, Rocquencourt, October 1987.

9. J. H. Bramble and S. R. Hilbert, "Estimation of linear functionals on Sobolev spaces with application to Fourier transforms and spline interpolation," *SIAM J. Numer. Anal.*, vol. 7, no. 1, pp. 112-124, 1970.

10. P. R. Halmos, *Measure Theory,* Springer Verlag, 1974.

11. H. L. Royden, *Real Analysis, second edition,* MacMillan Company, 1963.

12. V. Girault and P. Raviart, *Finite Element Methods for Navier-Stokes Equations,* Springer series in computational mathematics, 5, Springer-Verlag, 1986.

13. D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order,* Springer-Verlag, 1977.

14. William P. Ziemer, *Weakly Differentiable Functions,* Springer-Verlag, 1989.

15. Robert A. Adams, *Sobolev Spaces,* Academic Press, 1975.

3. The one-dimensional convection-diffusion equation

3.1 Introduction.

In this chapter we give a general technique to obtain a discretisation scheme for the one-dimensional convection-diffusion equation starting from Raviart-Thomas [1] or Brezzi-Douglas-Marini [2] type elements. The technique can also be applied in two or more dimensions. The resulting schemes are equivalent to the schemes based on transformed variables (called Slotboom variables in semiconductor context) introduced by Brezzi, Marini and Pietra [3] but without the Lagrange multipliers used in the latter schemes. The purpose of this chapter is to give an error analysis for such schemes that yields information on their local accuracy. For this purpose we adapt the technique used by Douglas and Roberts [4]. Our analysis differs in following two respects from the approach by O'Riordan and Stynes [5-10] or the approach by Reinhardt [11]. One: it deals with mixed finite elements as opposed to finite elements. And two: it attempts to deal with problems with localised singular perturbation. This last aspect is very important for semi-conductor problems, where we find such a situation in the continuity equations for the charge carriers. In that case the convection is given by the electric field. Singular perturbation may occur around junctions between differently doped materials, where very localised and very large electric fields can appear. We analyse the model equation,

$$-(au'-bu)'=f \quad \text{on} \quad \Omega \,, \tag{3.1a}$$

on the domain $\Omega=(0,L)$ with homogeneous boundary conditions,

$$u(0) = u(L) = 0 \,. \tag{3.1b}$$

Note the absence of a zero order term. In this respect our analysis is less general than that of the approaches of Stynes and O'Riordan and Reinhardt. Our analysis makes use of the regularity of the continuous problem and its adjoint. We take the adjoint problem to be

$$-((av')'+bv')=F \quad \text{on} \quad \Omega \,, \tag{3.2a}$$

with homogeneous boundary conditions,

$$v(0) = v(L) = 0 \,. \tag{3.2b}$$

We proceed as follows. To derive error bounds for the discrete problem, we need to know the regularity of the solution of (1), upper bounds on the norm of the solution of (1) and upper bounds on the norm of the solution of the adjoint problem. In section 3.2, we discuss the regularity of problem (1) under the condition that b/a is strictly positive. Section 3.3 derives upper bounds for the norm of the solution of the adjoint problem. In section 3.4 we describe the discretisation. Section 3.5 derives special estimates for projections of the solution of the adjoint problem that are needed later. Section 3.6 uses the results from the sections 3.2 to 3.4 to derive a priori error estimates. In section 3.7 we give our conclusions.

3.2 Regularity of the problem.

We formulate a theorem on the regularity of problem (1), which gives general formulas for the solution u of (1) and its flux $\sigma = -(au' - bu)$. We postpone its proof to sections 3.2.2 and 3.2.3. In section 3.2.1 we recall some facts concerning differentiation and integration needed in the proof of this theorem.

Theorem 3.1.
We assume that,

$$\frac{1}{a} \in \mathrm{L}^p(\Omega)\ , \quad \operatorname*{ess\,inf}_{x\in\Omega} \frac{1}{a} > 0\ , \tag{3.3a}$$

$$b \in \mathrm{L}^q(\Omega)\ , \quad \operatorname*{ess\,inf}_{x\in\Omega} b > 0\ , \tag{3.3b}$$

$$\frac{1}{p} + \frac{1}{q} = \frac{1}{r} \text{ with } r \geqslant 1\ , \tag{3.3c}$$

$$f \in \mathrm{W}^{k,1}(\Omega)\ , \tag{3.3d}$$

where

$$\operatorname*{ess\,inf}_{x\in\Omega} f = -\operatorname*{ess\,sup}_{x\in\Omega} -f = -\inf_{M\subset\Omega\,,\,\lambda(M)=0}\ \sup_{x\in\Omega-M} -f(x)\ ,$$

with λ the Lebesgue measure on $\mathbb{R}$. Note that (3b) implies $1/b \in \mathrm{L}^\infty(\Omega)$. Under the conditions (3a-d), equation (1) has a unique solution $u \in \mathrm{W}^{1,1}(\Omega)$ and

$$\|u\|_{\mathrm{L}^\infty(\Omega)} \leqslant \|1/b\|_{\mathrm{L}^\infty(\Omega)}\, \|f\|_{\mathrm{L}^1(\Omega)}\ , \tag{3.4a}$$

$$\|u\|_{\mathrm{W}^{1,1}(\Omega)} \leqslant \tag{3.4b}$$

$$(\|1\|_{\mathrm{L}^1(\Omega)} + \|b/a\|_{\mathrm{L}^1(\Omega)})\|1/b\|_{\mathrm{L}^\infty(\Omega)}\, \|f\|_{\mathrm{L}^1(\Omega)} + \|1/a\|_{\mathrm{L}^1(\Omega)}\, \|f\|_{\mathrm{L}^1(\Omega)}\ ,$$

$$\|\sigma\|_{\mathrm{L}^\infty(\Omega)} \leqslant \|f\|_{\mathrm{L}^1(\Omega)}\ , \tag{3.4c}$$

$$\|\sigma\|_{\mathrm{W}^{k+1,1}(\Omega)} \leqslant \|1\|_{\mathrm{L}^1(\Omega)}\, \|f\|_{\mathrm{L}^1(\Omega)} + \|f\|_{\mathrm{W}^{k,1}(\Omega)}\ . \tag{3.4d}$$

Moreover, if we introduce

$$\psi(x) = \int_{t=0}^{x} \frac{b(t)}{a(t)}\, dt\ , \tag{3.5}$$

$$S(\xi,\eta) = \int_{t=\xi}^{\eta} \frac{\exp(-\psi(t))}{a(t)}\, dt\ , \tag{3.6}$$

then the functions ψ and S are well-defined and the solution and the flux have the following absolutely continuous representations,

$$u(x) = \tag{3.7}$$

$$\frac{\exp(\psi(x))}{S(0,L)} \left[\int_{y=x}^{L} S(y,L)S(0,x)f(y)\,dy + \int_{y=0}^{x} S(0,y)S(x,L)f(y)\,dy \right],$$

$$-\sigma(x) = \tag{3.8}$$

$$a(x)u'(x) - b(x)u(x) =$$

$$\frac{1}{S(0,L)} \int_{y=x}^{L} S(y,L)f(y)\,dy - \frac{1}{S(0,L)} \int_{y=0}^{x} S(0,y)f(y)\,dy \,.$$

The above results stay valid as long as a and b are of fixed sign and are bounded away from zero. Section 3.2.1 recalls some important facts concerning the integration and differentiation of Lebesgue integrable functions. In section 3.2.2 we use the Green's function for (1) to derive the formulas for the solution and the flux. In section 3.2.3 we prove the rest of the theorem.

3.2.1. Facts on integration and differentiation of Lebesgue integrable functions.

In preparation for our proof of theorem 1, we recall some facts concerning the integration and differentiation of Lebesgue integrable functions. We recall the definition of weak differentiability and the definition of the Sobolev space $W^{k,p}(\Omega)$. We assume that Ω is a bounded interval.

Definition 3.1.
Let the absolute value of u be integrable on compact subsets of Ω. A function v, whose absolute value is integrable on compact subsets of Ω, is called the k^{th} weak derivative of u if it satisfies,

$$\int_{\Omega} \phi v \, d\mu = (-1)^k \int_{\Omega} u \frac{d^k \phi}{dx^k} d\mu \quad \forall \;\; \phi \in C_0^{\infty}(\Omega) \,.$$

Cf. section 1, chapter 2 [12].

Definition 3.2.
The Sobolev space $W^{k,p}(\Omega)$ is the space of $L^p(\Omega)$ functions for which all weak derivatives up to order k are $L^p(\Omega)$ functions. We use the following norm on this space,

$$||f||_{W^{k,p}(\Omega)} = \left[||f||^p_{L^p(\Omega)} + \sum_{j=1}^{k} || \frac{d^j f}{dx^j} ||^p_{L^p(\Omega)} \right]^{1/p} \quad \forall \;\; f \in W^{k,p}(\Omega) \,.$$

Cf. section 1, chapter 2 [12].

Definition 3.3.
A real-valued function f defined on a closed bounded interval $\overline{\Omega}$ is said to be

absolutely continuous on $\overline{\Omega}$ if, given $\epsilon > 0$, there is a $\delta > 0$ such that $\sum_{i=1}^{n} |f(y_i)-f(x_i)| < \epsilon$, for every finite collection of non-overlapping sub-intervals $\{ (x_i,y_i) \}_{i=1}^{n}$ of Ω with $\sum_{i=1}^{n} |y_i-x_i| < \delta$. Cf. section 4, chapter 5 [13].

Theorem 3.2.
A function F is an indefinite integral if and only if it is absolutely continuous. Theorem 13, section 4, chapter 5 of [13].

Theorem 3.3.
Every absolutely continuous function F is the indefinite integral of its derivative F' and if f is an integrable function on $\overline{\Omega}$,

$$F(x) = F(0) + \int_{t=0}^{x} f(t)\, dt \ ,$$

then $F'(x)=f(x)$ for almost all x in Ω. Corollary 14, section 4, chapter 5 and Theorem 9, section 3, chapter 5 [13].

Lemma 3.1.
If f and g are absolutely continuous on Ω, then fg and $\exp(f)$ are absolutely continuous.
Proof.
Consider the condition

$$\sum_{i=1}^{n} |fg(y_i)-fg(x_i)| < \epsilon \ .$$

Continuous functions on a closed interval are bounded, so f and g are bounded. Take $M=\max(\|f\|_{L^\infty(\Omega)}, \|g\|_{L^\infty(\Omega)})$. Now there exists by definition a δ such that, for every finite collection of non-overlapping sub-intervals $\{ (x_i,y_i) \}_{i=1}^{n}$ of Ω with $\sum_{i=1}^{n} |y_i-x_i| < \delta$,

$$\sum_{i=1}^{n} |f(y_i)-f(x_i)| < \frac{\epsilon}{2M} \quad \text{and} \quad \sum_{i=1}^{n} |g(y_i)-g(x_i)| < \frac{\epsilon}{2M} \ .$$

This implies that

$$\sum_{i=1}^{n} |fg(y_i)-fg(x_i)| \leqslant \sum_{i=1}^{n} |g(y_i)f(y_i)-g(x_i)f(y_i)+g(x_i)f(y_i)-g(x_i)f(x_i)| \leqslant$$

$$\sum_{i=1}^{n} M\,|g(y_i)-g(x_i)| + \sum_{i=1}^{n} M\,|f(y_i)-f(x_i)| \leqslant \epsilon \ .$$

Moreover, there is a δ such that, for every finite collection of non-overlapping sub-intervals $\{ (x_i,y_i) \}_{i=1}^{n}$ of Ω with $\sum_{i=1}^{n} |y_i-x_i| < \delta$,

$$\sum_{i=1}^{n} |f(y_i)-f(x_i)| \; < \; \epsilon \exp(-3M) \; .$$

In that case,

$$\sum_{i=1}^{n} |\exp(f(y_i))-\exp(f(x_i))| \; \leqslant$$

$$\sum_{i=1}^{n} \exp(f(x_i))\,|f(y_i)-f(x_i)|\,\exp(|f(y_i)-f(x_i)|) \; \leqslant \; \epsilon \; .$$

□

Theorem 3.4.
Let $\Omega=(\xi,\eta)$ be a bounded interval of $\mathbb{R}$. Let $C_0^\infty(\Omega)$ be the space of all $C^\infty(\Omega)$ functions with a compact support in Ω. Let $\mathrm{W}_0^{1,p}(\Omega)$ be the closure of $C_0^\infty(\Omega)$ in $\mathrm{W}^{k,p}(\Omega)$. All elements of $\mathrm{W}_0^{1,p}(\Omega)$, where $1 \leqslant p \leqslant \infty$, are absolutely continuous. Cf. Gilbarg and Trudinger, page 148 [14].
Proof.
We prove this to get an idea of the character of the space in question. For each $t \in \mathrm{W}_0^{1,p}(\Omega)$ there is by definition a Cauchy sequence $\{\, t_n \,\}_{n=1}^{\infty} \subset C_0^\infty(\Omega)$, that converges to t in the $\mathrm{W}^{k,p}(\Omega)$-norm. We denote the first derivative of a function g by g'. We have, $t_n \rightarrow t$ in $\mathrm{L}^p(\Omega)$ and $t_n' \rightarrow t'$ in $\mathrm{L}^p(\Omega)$, so, if we define

$$T_n(x) \;=\; \int_{y=\xi}^{x} t_n'(y)\,dy \quad \text{and} \quad T(x) \;=\; \int_{y=\xi}^{x} t'(y)\,dy \quad \text{for} \;\; x \in \Omega \; ,$$

then for all elements of the sequence $\{\, t_n \,\}$, we have $t_n \;=\; T_n$ and theorem 3 implies that

$$\|\, t' - T' \,\|_{\mathrm{L}^p(\Omega)} \;=\; 0 \; .$$

Moreover, for a given n,

$$\|\, t - T \,\|_{\mathrm{L}^p(\Omega)} \;\leqslant\; \|\, t - t_n \,\|_{\mathrm{L}^p(\Omega)} \;+\; \|\, T_n - T \,\|_{\mathrm{L}^p(\Omega)} \; ,$$

so

$$\|\, t - T \,\|_{\mathrm{L}^p(\Omega)} \;\leqslant\; \|\, t - t_n \,\|_{\mathrm{L}^p(\Omega)} \;+\; \|\, \int_{y=\xi}^{x} (t' - t_n')\,dy \,\|_{\mathrm{L}^p(\Omega)} \;\leqslant$$

$$\|\, t - t_n \,\|_{\mathrm{L}^p(\Omega)} \;+\; (\eta-\xi)\,\|\, t' - t_n' \,\|_{\mathrm{L}^p(\Omega)} \;\leqslant\; (1+\eta-\xi)\,\|\, t - t_n \,\|_{\mathrm{W}^{1,p}(\Omega)} \; .$$

This holds for all n, so $\|\, t - T \,\|_{\mathrm{L}^p(\Omega)} = 0$. This proves that t is the indefinite integral of t'. By theorem 2 this implies that t is absolutely continuous. □

3.2.2. The derivation of expressions for the solution and the flux.

We derive the expressions (7) and (8), we show that these functions satisfy (1), and we prove the statement about absolute continuity from theorem 1. We proceed as follows. In theorem 5 we construct the Green's function [15, 16] of (1) and use this to derive (7) and (8). We then substitute (7) in (1) and use theorem 3 to show that (7) and (8) satisfy (1). Absolute continuity of (7) and (8) is shown to follow from theorem 2. First we show that ψ is well-defined.

Lemma 3.2.
If (3a-d) hold, then the function ψ, defined by (5) is an absolutely continuous function on Ω and its derivative $\psi'(x)$ lies in $L^r(\Omega)$ and is equal to $b(x)/a(x)$.
Proof.
The Hölder inequality implies that

$$\|fg\|_{L^r(\Omega)} \leqslant \|f\|_{L^p(\Omega)}\|g\|_{L^q(\Omega)} \,. \tag{3.9}$$

for all $p,q,r \in [1,\infty]$ with $\frac{1}{p}+\frac{1}{q}=\frac{1}{r}$. For superscripts of $L^p(\Omega)$ spaces only, we use the convention that $1/0=\infty$ and $1/\infty=0$. We assumed that $\frac{1}{a} \in L^p(\Omega)$, $b \in L^q(\Omega)$, so, according to (9) $b/a \in L^r(\Omega)$. According to theorem 2, ψ is absolutely continuous. Theorem 3 implies that $\psi' = b/a$ in almost all points of Ω. □

Theorem 3.5.
Assume (3a-d), take ψ as in (5) and S as in (6). If $f \in L^1(\Omega)$ then the function u defined below is a solution of the equation (1) with right hand side f and with homogeneous boundary conditions.

$$u(x) = \int_{y=0}^{1} G(x,y)f(y)\,dy \,, \tag{3.10}$$

with

$$G(x,y) := \frac{\exp(\psi(x))}{S(0,L)}\Big[\theta(y-x)S(y,L)S(0,x) + \theta(x-y)S(0,y)S(x,L)\Big] \tag{3.11}$$

where θ is the Heaviside function,

$$\theta(z) = \begin{cases} 0 \text{ if } z < 0\,, \\ \tfrac{1}{2} \text{ if } z=0\,, \\ 1 \text{ if } z > 0\,. \end{cases} \tag{3.12}$$

Proof.
We see immediately that $G \in C([0,L]\times[0,L])$. We use theorem 3 and the chain rule to derive (8) from (7). According to the chain rule and theorem 3,

$$u'(x) = \tag{3.13}$$

$$\frac{b(x)}{a(x)}u(x) + \frac{1}{a(x)S(0,L)}\int_{y=x}^{L} S(y,L)f(y)\,dy - \frac{1}{a(x)S(0,L)}\int_{y=0}^{x} S(0,y)f(y)\,dy\ ,$$

equation (8) follows immediately from (13) and the definition of the flux. Absolute continuity of the solution constructed with the aid of the Green's function follows from theorem 2, lemma 1 and equations (7) and (8). We see immediately that u satisfies the homogeneous Dirichlet boundary conditions. When we apply theorem 3 to equation (8), we find, $\sigma'(x) = f(x)$. □
See also the books by Roach and Yosida [15, 16].

3.2.3. Upper bounds on the norms of the solution and the flux.

We complete the proof of theorem 1 by proving that (1) has a unique solution in $W^{1,1}(\Omega)$ and deriving the upper bounds on the norm of the solution and the flux from (7) and (8).

First we verify uniqueness of the solution as follows. Suppose (1) has two solutions $u_1, u_2 \in W_0^{1,p}(\Omega)$ for a given f. This implies that $w_0 = u_2 - u_1 \in W_0^{1,p}(\Omega)$ is a solution of (1) with $f = 0$. Now by definition,

$$((\exp(-\psi)w_0)', a\exp(\psi)\phi') = 0 \quad \forall\ \phi \in C_0^\infty(\Omega)\ ,$$

so $(\exp(-\psi)w_0)' = 0$. According to theorem 4, the function w_0 is absolutely continuous and according to lemma 1 the function $\exp(-\psi)$ is absolutely continuous. Theorem 3 now implies that $\exp(-\psi)w_0$ is constant. The only $w_0 \in W_0^{1,p}(\Omega)$ that can give this result is $w_0 = 0$.

Before we can derive upper bounds on the norms of (7) and (8), we need to derive some bounds on $S(\xi,\eta)$.

Lemma 3.3.
Assume (3a-d) and take ψ as in (5). Let S be the function on $\Omega \times \Omega$ defined by (6). Then $S(0,x)$ and $S(x,L)$ are absolutely continuous functions. If $0 \leqslant \xi_0 \leqslant \xi < \eta \leqslant \eta_0 \leqslant L$ then

$$0 < S(\xi,\eta) \leqslant S(\xi_0,\eta_0)\ , \tag{3.14}$$

and

$$S(\xi,\eta) \leqslant \|b^{-1}\|_{L^\infty(\Omega)}(\exp(-\psi(\xi)) - \exp(-\psi(\eta)))\ . \tag{3.15}$$

Proof.
From (3a, b) and the positivity of the integrand (14) follows immediately. From (6) it follows that

$$S(\xi,\eta) = \int_{x=\xi}^{\eta} \frac{1}{b(x)} \frac{-d\exp(-\psi(x))}{dx}(x)\,dx\ .$$

We see, that $(-\exp(-\psi))' = \psi'\exp(-\psi) > 0$, so

$$S(\xi,\eta) \leqslant \|b^{-1}\|_{L^\infty(\Omega)} \int_{x=\xi}^{\eta} \frac{-d\exp(-\psi(x))}{dx}(x)\,dx\ .$$

As $\exp(-\psi)$ is absolutely continuous according to lemma 1, we find from theorem 3 that

$$\int_{t=\xi}^{\eta} (\exp(-\psi(t)))'(t)\, dt = \exp(-\psi(\eta)) - \exp(-\psi(\xi)) .$$

□

Next, we can prove the inequalities (4a-d). We assume that (3a-d) hold. Application of (14) to (7) yields the following upper bound on u,

$$|u(x)| \leqslant \frac{\exp(\psi(x))S(x,L)S(0,x)}{S(0,L)} \|f\|_{L^1(\Omega)} .$$

We use (14) and (15) to write this as,

$$|u(x)| \leqslant \|1/b\|_{L^\infty(\Omega)} \|f\|_{L^1(\Omega)} .$$

This proves (4a). Now (4b) follows immediately from (13). Next, we derive (4c). From (14) and (8) an estimate for σ follows immediately:

$$|a(x)u'(x) - b(x)u(x)| \leqslant \|f\|_{L^1(\Omega)} .$$

And (4d) follows from (4c) and the fact that (1) implies $\sigma' = f$.

3.3 The adjoint problem.

First, we derive a Green's function for (2). Then we give expressions for the solution and the flux of (2). Finally we derive upper bounds on the norms of the solution and the flux. The following theorem accomplishes our first two goals.

Theorem 3.6.
Assume (3a-d), take ψ as in (5) and S as in (6). If $F \in L^1(\Omega)$ then the function v defined below is a solution of the equation (2) with right hand side F and with homogeneous boundary conditions.

$$v(x) = \tag{3.16}$$

$$\frac{1}{S(0,L)} \left[\int_{y=0}^{x} S(x,L)S(0,y)\exp(\psi(y))F(y)dy + \int_{y=x}^{L} S(0,x)S(y,L)\exp(\psi(y))F(y)dy \right] .$$

Proof.
The Green's function for the adjoint problem (2) is given by, $\overline{G}(x,y) = G(y,x)$. See also Roach or Yosida [15, 16]. According to theorem (2) v is absolutely continuous on $[0,L]$, so $v(0)=v(L)=0$. Moreover,

$$\tau(x) = -a(x)v'(x) = \tag{3.17}$$

$$-\frac{\exp(-\psi(x))}{S(0,L)} \left[-\int_{y=0}^{x} S(0,y)\exp(\psi(y))F(y)dy + \int_{y=x}^{L} S(y,L)\exp(\psi(y))F(y)dy \right] .$$

And by differentiation of integrals, $\tau'(x) = -\dfrac{b(x)}{a(x)}\tau(x) + F(x)$. This in turn implies that v satisfies the adjoint problem. □

It now remains to give upper bounds on the norms of the solution and the flux.

Theorem 3.7.

Assume (3a-d), take ψ as in (5) and S as in (6). Assume $F \in \mathrm{W}^{k,1}(\Omega)$. Now (2) has a unique solution $v \in \mathrm{W}^{1,1}(\Omega)$. The solution v and the corresponding flux τ, defined by $\tau = -av'$, have the following properties:

$$\| v \|_{\mathrm{L}^\infty(\Omega)} \leqslant \| 1/b \|_{\mathrm{L}^\infty(\Omega)} \| F \|_{\mathrm{L}^1(\Omega)} \ , \tag{3.18a}$$

$$\| v \|_{\mathrm{W}^{1,1}(\Omega)} \leqslant \tag{3.18b}$$

$$\left(\| 1 \|_{\mathrm{L}^1(\Omega)} \| 1/b \|_{\mathrm{L}^\infty(\Omega)} + \left(1 + \frac{\| 1/b \|_{\mathrm{L}^\infty(\Omega)}}{S(0,L)} \right) \| 1/a \|_{\mathrm{L}^1(\Omega)} \right) \| F \|_{\mathrm{L}^1(\Omega)} \ ,$$

$$\| \tau \|_{\mathrm{L}^\infty(\Omega)} \leqslant \left(1 + \frac{\| 1/b \|_{\mathrm{L}^\infty(\Omega)}}{S(0,L)} \right) \| F \|_{\mathrm{L}^1(\Omega)} \ . \tag{3.18c}$$

Moreover, the solution and the flux are absolutely continuous.

Proof.

The solution is unique, because if it is not, then (2) with $F \equiv 0$ has a non-trivial solution in $\mathrm{W}_0^{1,1}(\Omega)$. This in turn would imply that there is an absolutely continuous w_0 such that

$$(aw_0')' + bw_0' = 0 \ \text{ on } \ \Omega \ ,$$

$$w_0(0) = w_0(L) = 0 \ .$$

According to theorem 1 there is a unique absolutely continuous $v \in \mathrm{W}_0^{1,1}(\Omega)$ such that

$$(av' - bv)' = w_0 \ .$$

But this implies that

$$(w_0, w_0) = (w_0, (av' - bv)') = -(w_0', av' - bv) = -(aw_0', v') + (bw_0', v) \ .$$

We use the definition of weak differentiability to write this as,

$$(w_0, w_0) = ((aw_0')' + bw_0', v) = 0 \ .$$

This implies that $w_0 = 0$. Absolute continuity of the solution constructed with the aid of the Green's function follows from theorem 2, lemma 1 and equations (16) and (17). Uniqueness of the solution implies that we may derive upper bounds on the norm of the solution and the flux from the previously given expressions. We proceed as follows. Application of (14) and (15) to (16) yields the following estimate for v,

$$| v(x) | \leqslant \| 1/b \|_{\mathrm{L}^\infty(\Omega)} \| F \|_{\mathrm{L}^1(\Omega)} \ .$$

This proves (18a). The inequality (18b) follows immediately from (17). Next, we derive the (18c). From (14) and (17) an estimate for τ follows immediately:

$$|a(x)v'(x)| \leqslant \left(1+\frac{\|b^{-1}\|_{L^{\infty}(\Omega)}}{S(0,L)}\right)\|F\|_{L^{1}(\Omega)} .$$

□

3.4 The discretisation.

We construct a Petrov-Galerkin mixed finite element discretisation. Our derivation uses trial spaces V_h and W_h that are defined as the ranges of the projections $\Pi_h: V \rightarrow V_h$ and $P_h: W \rightarrow W_h$, where we take $V = W^{1,1}(\Omega)$ and $W = L^1(\Omega)$. This approach was first used by Raviart and Thomas [1] and Fortin [17]. Our test spaces are derived from the trial spaces by multiplication with an exponential function. The final result will be equivalent to the standard mixed finite element discretisaticn for the symmetrised form of the equation but the special derivation allows us to obtain better a-priori error estimates. We proceed as follows. First we give conditions on the projections P_h and Π_h. We show that these conditions guarantee that $\frac{d}{dx}(V_h) = W_h$. Next we give an example of such projections. Finally we derive the discrete scheme and verify that the resulting discrete problem has a unique solution.

3.4.1. The projections onto the trial spaces for the solution and its flux.

As mentioned earlier, we derive our trial spaces from projections $P_h: W \rightarrow W$ and $\Pi_h: V \rightarrow V$. We assume these projections have finite dimensional ranges and satisfy the following conditions:

$$(s, P_h t) = (P_h s, t) \quad \forall \; s,t \in W , \tag{3.19}$$

and

$$P_h \frac{d}{dx} v = \frac{d}{dx} \Pi_h v \quad \forall \; v \in V , \tag{3.20a}$$

$$\Pi_h v(0) = v(0) \quad \forall \; v \in V . \tag{3.20b}$$

We define our approximation spaces as follows. We set $V_h = \mathscr{R}(\Pi_h)$ and $W_h = \mathscr{R}(P_h)$.

Theorem 3.8.
The map $\frac{d}{dx}: W^{1,1}(\Omega) \rightarrow L^1(\Omega)$ is continuous and surjective.
Proof.
Continuity follows immediately from the norms on these spaces. The map is surjective because, for all $f \in L^1(\Omega)$, theorem 2 shows that the function F, defined by

$$F(x) = \int_{y=0}^{x} f(y)\, dy \quad \text{for} \quad x \in \Omega ,$$

is an element of $W^{1,1}(\Omega)$ with derivative f. □

Corollary 3.1.
The map $\frac{d}{dx}: V_h \rightarrow W_h$ is surjective.
Proof.
From (20) it follows that the image of V_h under $\frac{d}{dx}$ lies in W_h. From theorem 8 and (20) it follows that the image is in fact equal to W_h. □

The above use of projections can be found in [1, 17].

3.4.2. An example of a set of trial spaces.

An example of a set of spaces and projections that meet these criteria are the lowest order Raviart-Thomas spaces with the projections given in [1]. For the one dimensional case, this simply means that the image of a function under Π_h is obtained by linear interpolation between the values in mesh nodes and for P_h the image is obtained by taking cell-wise averages. Now V_h is the space of continuous functions that are linear on the mesh cells and W_h is the space of functions that are constant on mesh cells.

3.4.3. The discrete scheme.

We construct a Petrov-Galerkin mixed finite element method as follows. We take V_h as trial space for σ. As test space for σ we take $X_h = \exp(-\psi)V_h$. For u we take W_h as test space and $Y_h = \exp(\psi)W_h$ as trial space. Here ψ is defined as in (5). We define projections onto X_h and Y_h.

$$\hat{\Pi}_h \tau = \exp(-\psi)\Pi_h(\exp(\psi)\tau) , \tag{3.21}$$

$$\hat{P}_h t = \exp(\psi)P_h(\exp(-\psi)t) . \tag{3.22}$$

From (19) it follows that,

$$(s, \hat{P}_h t) = (\exp(-\psi)P_h(\exp(\psi)s), t) = (\exp(-\psi)P_h(\exp(\psi)s), \hat{P}_h t) = (\hat{P}_h^* s, \hat{P}_h t) \tag{3.23}$$

where $\hat{P}_h^*$ is the adjoint operator of $\hat{P}_h$. By application of the defining formulas we find,

$$\left(\frac{d}{dx}\tau + \frac{b}{a}\tau, \hat{P}_h t \right) = \left(\left(\frac{d}{dx} + \frac{b}{a} \right) \hat{\Pi}_h \tau, \hat{P}_h t \right) . \tag{3.24}$$

The continuous solution of (1) satisfies

$$(\sigma, u) \in V \times W , \tag{3.25a}$$

$$\left(\sigma, \frac{\tau}{a} \right) - \left(\frac{d}{dx}\tau + \frac{b}{a}\tau, u \right) = 0 \quad \forall \ \tau \in H^1(\Omega) , \tag{3.25b}$$

$$\left(\frac{d}{dx}\sigma,t\right)=(f,t) \quad \forall \;\; t \in \mathrm{L}^2(\Omega) \; . \tag{3.25c}$$

Our discrete scheme has the following form.

$$(\sigma_h,u_h) \in V_h \times Y_h \; , \tag{3.26a}$$

$$\left(\sigma_h,\frac{\tau_h}{a}\right) \; - \; \left(\frac{d}{dx}\tau_h+\frac{b}{a}\tau_h,u_h\right)=0 \quad \forall \;\; \tau_h \in X_h \; , \tag{3.26b}$$

$$\left(\frac{d}{dx}\sigma_h,t_h\right)=(f,t_h) \quad \forall \;\; t_h \in W_h \; . \tag{3.26c}$$

We see that this scheme is equivalent to,

$$(\sigma_h, U_h) \in V_h \times W_h \; , \tag{3.27a}$$

$$\left(\sigma_h,\frac{\exp(-\psi)\tau_h}{a}\right) \; - \; \left(\frac{d}{dx}\tau_h, U_h\right)=0 \quad \forall \;\; \tau_h \in V_h \; , \tag{3.27b}$$

$$\left(\frac{d}{dx}\sigma_h,t_h\right)=(f,t_h) \quad \forall \;\; t_h \in W_h \; . \tag{3.27c}$$

This last system has a unique solution. This can be demonstrated as follows. Suppose $f=0$. As σ_h is continuous, (20) and (27c) imply that σ_h is constant. Now take $\tau_h=1$, from (27b) it follows that $\sigma_h \equiv 0$. Now corollary 1 implies $U_h \; = \; 0$. This completes the demonstration.

3.5 Properties of the projections.

In the section on a priori error estimates we shall need estimates of terms containing the difference between a function and its projection under one of the projections introduced in the previous section. In this section we give estimates for those terms. We start by considering $\boldsymbol{\tau}-\hat{\Pi}_h\boldsymbol{\tau}$. To do this we need the following auxiliary lemma.

Lemma 3.4.
If $f \in \mathrm{W}^{1,1}(\Omega)$ then

$$f(x)-\Pi_h f(x) \; = \; \int_{y=0}^{x} (f' \; - \; \mathrm{P}_h(f'))dy \; . \tag{3.28}$$

Proof.
The function f is continuous and differentiable, so

$$f(x)=f(0) \; + \; \int_{y=0}^{x} f'(y) \; dy \; .$$

Moreover (20) implies that,

$$\Pi_h f(x) \; = \; \Pi_h f(0) \; + \; \int_{y=0}^{x} \mathrm{P}_h(f')(y) \; dy \; .$$

□

Lemma 3.5.
If $(v,\tau=-av')$ is the solution of the adjoint equation for the right hand side F, then

$$|(t,(\tau-\hat{\Pi}_h\tau)/a)| \leqslant \|b^{-1}\|_{L^\infty(\Omega)} \|F-\hat{P}_hF\|_{L^1(\Omega)} \|t\|_{L^\infty(\Omega)} \quad \forall \ t \in L^\infty(\Omega) \ .$$

Proof.
We know that

$$v(x) = \int_{y=0}^{L} \overline{G}(x,y)F(y)\,dy \ ,$$

where $\overline{G}(x,y)$ is Green's function for the adjoint problem. Now consider

$$av'(x)-\exp(-\psi(x))\Pi_h(\exp(\psi)av') \ .$$

We can write this as,

$$\exp(-\psi(x))\left[\exp(\psi)av'(x)-\Pi_h(\exp(\psi)av')\right] \ .$$

We wish to apply the previous lemma. To do this we need the first derivative of $\exp(\psi)av'$. Equation (17) implies that

$$(\exp(\psi)av')' = -\exp(\psi(x))F(x) \ .$$

We use this to evaluate the expression $(t,(\tau-\hat{\Pi}_h\tau)/a)$,

$$(t,(\tau-\hat{\Pi}_h\tau)/a) = \int_{x=0}^{L} \frac{\exp(-\psi(x))}{a(x)} \int_{y=0}^{x} \exp(\psi)F-\mathrm{P}_h(\exp(\psi)F)\,dy\ t(x)dx =$$

$$\int_{x=0}^{L} \frac{\exp(-\psi(x))}{a(x)} \int_{y=0}^{x} \exp(\psi)(F-\hat{P}_hF)\,dy\ t(x)dx \ .$$

This implies,

$$(t,(\tau-\hat{\Pi}_h\tau)/a) \leqslant \left| \int_{y=0}^{L}\int_{x=y}^{L} \frac{\exp(\psi(y)-\psi(x))}{a(x)}(F-\hat{P}_hF)(y)\ t(x)dxdy \right| \leqslant$$

$$\int_{y=0}^{L} \|t\|_{L^\infty(\Omega)} \|b^{-1}\|_{L^\infty(\Omega)} |(F-\hat{P}_hF)(y)|\,dy \ .$$

□

Next, we consider $v-\mathrm{P}_hv$.

Lemma 3.6.
If $(v,\tau=-av')$ is the solution of the adjoint equation for the right hand side F,

then

$$\| v - \mathrm{P}_h v \|_{\mathrm{L}^2(\Omega)} \leqslant \| 1 \|_{\mathrm{L}^2(\Omega)} \| 1/b \|_{\mathrm{L}^\infty(\Omega)} \| F \|_{\mathrm{L}^1(\Omega)} \,. \tag{3.29}$$

Proof.
This follows immediately from $\| v - \mathrm{P}_h v \|_{\mathrm{L}^2(\Omega)} \leqslant \| v \|_{\mathrm{L}^2(\Omega)}$ and (18a). □

3.6 A priori error estimates.

We derive estimates for $\| \sigma - \sigma_h \|_{\mathrm{L}^\infty(\Omega)}$ and $\| \hat{P}_h u - u_h \|_{\mathrm{W}^{k,1}(\Omega)}$. We start by giving estimates for $\| \sigma - \sigma_h \|_{\mathrm{L}^\infty(\Omega)}$ and $\| \sigma - \sigma_h \|_{\mathrm{W}^{1,1}(\Omega)}$. We proceed as follows. First we show that there is a point $\xi \in \Omega$ where the function $\sigma - \sigma_h$ is zero, then we determine the first derivative of the function and use this to determine the desired estimates.

Lemma 3.7.
Given that (3a-d) are satisfied and σ satisfies (25c) and σ_h is a solution of (26c), there is at least one point ξ such that $(\sigma - \sigma_h)(\xi) = 0$.
Proof.
We see immediately that $\exp(-\psi) \in X_h$. The solution of (1) satisfies (25b), so

$$\left(\sigma - \sigma_h, \frac{\exp(-\psi)}{a} \right) = 0 \,.$$

We know that $\exp(-\psi)$ and a are strictly positive and bounded from below, so there must be places where $\sigma - \sigma_h$ is negative. We know that $\sigma \in \mathrm{W}^{1,1}(\Omega)$ from $-\sigma' = f$, and $\sigma_h \in V_h \subset \mathrm{W}^{1,1}(\Omega)$ so $\sigma - \sigma_h$ is continuous. This implies that there is a ξ such that $(\sigma - \sigma_h)(\xi) = 0$.

□

Theorem 3.9.
If (3a-d) hold and

$$C(f) = | \sup_{\xi,\eta \in \Omega} \int_{y=\xi}^{\eta} (f - \mathrm{P}_h f)(y) dy | \,, \tag{3.30}$$

then $C(f) \leqslant \| f - \mathrm{P}_h f \|_{\mathrm{L}^1(\Omega)}$ and

$$\| \sigma - \sigma_h \|_{\mathrm{L}^\infty(\Omega)} \leqslant C(f) \,, \tag{3.31}$$

$$\| \sigma - \sigma_h \|_{\mathrm{W}^{1,1}(\Omega)} \leqslant \| 1 \|_{\mathrm{L}^1(\Omega)} C(f) + \| f - \mathrm{P}_h f \|_{\mathrm{L}^1(\Omega)} \,. \tag{3.32}$$

Proof.
We take ξ to be a zero of $\sigma - \sigma_h$. We know that $\sigma - \sigma_h \in \mathrm{W}^{1,1}(\Omega)$, so we may write,

$$(\sigma - \sigma_h)(x) - (\sigma - \sigma_h)(\xi) = \int_{y=\xi}^{x} (\sigma - \sigma_h)'(y) dy \,.$$

From (25) and (26) we see immediately that

$$(\sigma-\sigma_h)' = f-\mathrm{P}_h f \, .$$

This implies,

$$(\sigma-\sigma_h)(x)-(\sigma-\sigma_h)(\xi) = \int_{y=\xi}^{x} (f-\mathrm{P}_h f)(y)dy \, .$$

This implies that

$$\|\sigma-\sigma_h\|_{\mathrm{L}^\infty(\Omega)} \leqslant \sup_{x\in\Omega} \int_{y=\xi}^{x} (f-\mathrm{P}_h f)(y)dy \, .$$

□

We give an estimate for $\|\hat{P}_h u-u_h\|_{\mathrm{W}^{k,1}(\Omega)^*}$. To derive this estimate we use the dual problem.

Theorem 3.10.
Under the conditions given in(3a-d),

$$\|\hat{P}_h u-u_h\|_{\mathrm{L}^\infty(\Omega)} \leqslant 2\|b^{-1}\|_{\mathrm{L}^\infty(\Omega)}(1 + C_k(\psi))\|f-\mathrm{P}_h f\|_{\mathrm{L}^1(\Omega)} \, ,$$

where

$$C_k(\psi) := \sup_{F\in\mathrm{L}^1(\Omega)} \frac{\|F-\hat{P}_h F\|_{\mathrm{L}^1(\Omega)}}{\|F\|_{\mathrm{L}^1(\Omega)}} \, ,$$

and

$$\|\hat{P}_h u-u_h\|_{\mathrm{W}^{k,1}(\Omega)^{2*}} \leqslant 2\|b^{-1}\|_{\mathrm{L}^\infty(\Omega)}(1 + D_k(\psi))\|f-\mathrm{P}_h f\|_{\mathrm{L}^1(\Omega)} \, ,$$

where

$$D_k(\psi) := \sup_{F\in\mathrm{W}^{k,1}(\Omega)} \frac{\|F-\hat{P}_h F\|_{\mathrm{L}^1(\Omega)}}{\|F\|_{\mathrm{W}^{k,1}(\Omega)}} \, .$$

Proof.
Regularity of the adjoint problem gives us a solution $(t,\tau=-at')$ of (2) for all $F\in\mathrm{L}^1(\Omega)$. For this solution, we see that according to (26),

$$(\hat{P}_h u-u_h,F) = (\tau'+\psi'\tau,\hat{P}_h u-u_h) = ((\hat{\Pi}_h\tau)'+\psi'(\hat{\Pi}_h\tau),\hat{P}_h u-u_h) =$$
$$(\sigma-\sigma_h,\hat{\Pi}_h\tau/a) = (\sigma-\sigma_h,\tau/a) - (\sigma-\sigma_h,(\tau-\hat{\Pi}_h\tau)/a) =$$
$$((\sigma-\sigma_h)',t) - (\sigma-\sigma_h,(\tau-\hat{\Pi}_h\tau)/a) = (f-\mathrm{P}_h f,t-\mathrm{P}_h t) - (\sigma-\sigma_h,(\tau-\hat{\Pi}_h\tau)/a) \, .$$

We use lemma 5 and 6 and theorem 9 to derive from this that,

$$|(\hat{P}_h u-u_h,F)| \leqslant$$
$$(\|1\|_{\mathrm{L}^2(\Omega)}\|F\|_{\mathrm{L}^1(\Omega)} + \|F-\hat{P}_h F\|_{\mathrm{L}^1(\Omega)})2\|b^{-1}\|_{\mathrm{L}^\infty(\Omega)}\|f-\mathrm{P}_h f\|_{\mathrm{L}^1(\Omega)} \, .$$

□

Corollary 3.2.
Assume that W_h contains the characteristic functions $\chi_{(x_{i-1},x_i)}$ of the cells of the partition $P = \{ \, 0 = x_0 < x_1 < x_2 < \cdots < x_n = L \, \}$. As a direct consequence of theorem 10 and under the same conditions, we find

$$\frac{\| \hat{P}u - u_h \|_{L^1((x_{i-1},x_i))}}{\| 1 \|_{L^1((x_{i-1},x_i))}} \leqslant 2 \| b^{-1} \|_{L^\infty(\Omega)} \| f - P_h f \|_{L^1(\Omega)} \, . \qquad (3.33)$$

Proof.
We prove this as follows. For F in the proof of theorem 10, take $F = \exp(\psi)\chi_{(x_{i-1},x_i)}$. According to the Riesz representation theorem $L^1(\Omega)^* = L^\infty(\Omega)$. We find,

$$\| \exp(2\psi) \|_{L^1((x_{i-1},x_i))} \| \exp(-\psi)(\hat{P}u - u_h) \|_{L^\infty((x_{i-1},x_i))} \leqslant \qquad (3.34)$$
$$2 \| b^{-1} \|_{L^\infty(\Omega)} \| \exp(\psi) \|_{L^1((x_{i-1},x_i))} \| f - P_h f \|_{L^1(\Omega)} \, .$$

We see immediately that,

$$0 \leqslant (g - P_h g, g - P_h g) = \| g \|^2_{L^2(\Omega)} - \| P_h g \|^2_{L^2(\Omega)} \, .$$

This implies that,

$$\| \exp(2\psi) \|_{L^1((x_{i-1},x_i))} \geqslant \frac{\| \exp(\psi) \|^2_{L^1((x_{i-1},x_i))}}{\| 1 \|_{L^1((x_{i-1},x_i))}} \, .$$

We apply this to (34) and find,

$$\frac{\| \exp(\psi) \|_{L^1((x_{i-1},x_i))}}{\| 1 \|_{L^1((x_{i-1},x_i))}} \| \exp(-\psi)(\hat{P}u - u_h) \|_{L^\infty((x_{i-1},x_i))} \leqslant$$
$$2 \| b^{-1} \|_{L^\infty(\Omega)} \| \exp(\psi) \|_{L^1((x_{i-1},x_i))} \| f - P_h f \|_{L^2(\Omega)} \, .$$

This implies,

$$\frac{\| \exp(\psi) \|_{L^1((x_{i-1},x_i))}}{\| 1 \|_{L^1((x_{i-1},x_i))}} \| \exp(-\psi)(\hat{P}u - u_h) \|_{L^\infty((x_{i-1},x_i))} \leqslant$$
$$2 \| b^{-1} \|_{L^\infty(\Omega)} \| f - P_h f \|_{L^2(\Omega)} \, .$$

Which in turn implies that

$$\frac{\| \hat{P}u - u_h \|_{L^1((x_{i-1},x_i))}}{\| 1 \|_{L^1((x_{i-1},x_i))}} \leqslant 2 \| b^{-1} \|_{L^\infty(\Omega)} \| f - P_h f \|_{L^2(\Omega)} \, .$$

□

3.7 Conclusions.

We see that the accuracy of the solution of the problem with homogeneous Dirichlet boundary conditions is entirely determined by two factors. One being the accuracy of the approximation of the right hand side f by $\mathrm{P}_h f$ and the other being the quality of the approximation of $F \in W_1^k(\Omega)$ by $\hat{P}_h F$. As mentioned in the introduction, in the semi-conductor continuity equations the convection is given by the electric field. Singular perturbation may occur around junctions between differently doped materials, where locally very large electric fields can appear. From the uniform $\mathrm{L}^\infty(\Omega)$ error estimate for the total error in the flux in theorem 9 it follows that local singular perturbation on the approximation has no influence on the error in the flux. In corollary 2 we get a uniform cell-wise estimate for the discretisation error with respect to a problem dependent projection that is close to $\mathrm{L}^2(\Omega)$ projection on cells where the convection does not dominate the diffusion. The main problem that must be faced when extending this analysis to two or more dimensions, is the derivation of a useful estimate for $\|\sigma - \sigma_h\|$.

References

1. P. A. Raviart and J. M. Thomas, “A mixed finite element method for 2-nd order elliptic problems,” in *Mathematical aspects of the finite element method*, Lecture Notes in Mathematics, vol. 606, pp. 292-315, Springer, 1977.
2. F. Brezzi, J. Douglas Jr., and L. D. Marini, “Two Families of Mixed Finite Elements for Second Order Elliptic Problems,” *Numerische Mathematik*, vol. 47, pp. 217-235, 1985.
3. F. Brezzi, L. D. Marini, and P. Pietra, “Mixed exponential fitting schemes for current continuity equations,” in *Proceedings of the sixth international NASECODE conference*, ed. J. J. H. Miller, Boole Press Ltd, 1989.
4. J. Douglas, Jr. and J. E. Roberts, “Global estimates for mixed methods for second order elliptic equations,” *Mathematics of computation*, vol. 44, no. 169, pp. 39-52, 1985.
5. Eugene O’Riordan, “Singularly Perturbed Finite Element methods,” *Numerische Mathematik*, vol. 44, pp. 425-434, Springer-Verlag, 1984.
6. Eugene O’Riordan and Martin Stynes, “A finite element method for a singularly perturbed boundary value problem in conservative form,” in *BAIL III*, ed. J. J. H. Miller, pp. 271-275, Boole Press, Dublin, 1984.
7. Martin Stynes and Eugene O’Riordan, “A Finite Element Method for a Singular Perturbed Boundary Value Problem,” *Numerische Mathematik*, vol. 50, pp. 1-15, Springer-Verlag, 1986.
8. Eugene O’Riordan and Martin Stynes, “A Uniformly Accurate Finite-Element Method for a Singularly Perturbed One-Dimensional Reaction-Diffusion Problem,” *Mathematics of Computation*, vol. 47, no. 176, pp.

555-570, 1986.

9. Eugene O'Riordan and Martin Stynes, "An analysis of a Superconvergence Result for a Singularly Perturbed Boundary Value Problem," *Mathematics of Computation*, vol. 46, no. 173, pp. 81-92, 1986.

10. Eugene O'Riordan and Martin Stynes, "A uniform finite element method for a conservative singularly perturbed problem," *Journal of Computational and Applied Mathematics*, vol. 18, pp. 163-174, 1987.

11. H.-J. Reinhardt, "A-Posteriori Error Analysis and Adaptive Finite Element Methods for Singularly Perturbed Convection-Diffusion Equations," *Math. Meth. in the Appl. Sci.*, vol. 4, pp. 529-548, 1982.

12. William P. Ziemer, *Weakly Differentiable Functions,* Springer-Verlag, 1989.

13. H. L. Royden, *Real Analysis, second edition,* MacMillan Company, 1963.

14. D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order,* Springer-Verlag, 1977.

15. K. Yosida, *Lectures on Differential and Integral Equations,* Interscience, 1960.

16. G. F. Roach, *Green's Functions: introductory theory with applications,* Van Nostrand Reinhold Company, 1970.

17. M. Fortin, "An analysis of the convergence of mixed finite element methods," *RAIRO Numerical Analysis*, vol. 11, no. 4, pp. 341-354, 1977.

4. A Finite Volume Discretisation Scheme with a-posteriori Error Estimates for the Symmetrised Continuity Equation.

4.1 Introduction.

Before we can discuss the approach used to obtain an a-posteriori error estimate, we must give our interpretation of the well-known extension of the one-dimensional Scharfetter-Gummel scheme [1] to two dimensions. We take as our starting point the continuity equation for electrons in the stationary case,

$$-a(\mathbf{grad}\, u + u\, \mathbf{grad}\, \psi) = \boldsymbol{\sigma} ,$$

$$\operatorname{div} \boldsymbol{\sigma} = f .$$

We sketch the derivation of the discretisation on a rectangular grid. Consider two adjacent cells. We assume that a and the component of $\mathbf{grad}\, \psi$ along the line segment $\hat{\Gamma}$ connecting the cell centres are constant. Furthermore we assume that the component of $\boldsymbol{\sigma}$ parallel to $\hat{\Gamma}$ is constant on $\hat{\Gamma}$ and on the common cell edge. Under these assumptions we can give an expression for $\boldsymbol{\sigma}$ in terms of u. Furthermore we can calculate the integral over the common cell edge of the component of $\boldsymbol{\sigma}$ orthogonal to the common cell edge. This gives us a finite volume scheme for the above equations. Note that along the line segment $\hat{\Gamma}$ we get an exponential fitting scheme as described by Il'in [2]. The resulting discretisation scheme is equivalent to one of the schemes discussed in the articles by Bank et al. [3, 4].

For the error analysis we choose a trial space $V_h \times W_h$ and write the finite volume scheme as a saddle point problem which has a solution in that trial space. We use theorem 3.1 from the article by Nicolaides [5] to prove stability of the problem and existence of the solution. We then choose a projection $\Pi_h \times \overline{\mathrm{P}}_h$ of the solution $(\boldsymbol{\sigma}, u)$ of the continuous problem. We use the stability of the problem to give an upper bound on the global discretisation error in terms of the local discretisation error. We show that we may express the local discretisation error in terms of partial derivatives of $\boldsymbol{\sigma}$. Consistency follows immediately from the expression obtained. We then use the expression for the residual to construct a deferred correction scheme, based on the finite volume scheme in that form. We prove that, if the original scheme gives an $\mathcal{O}(h^k)$ accurate approximation, then this deferred correction scheme gives an $\mathcal{O}(h^{k+1})$ accurate approximation.

Our analysis shows, that the discretisation error for the Scharfetter-Gummel scheme is second order in areas of constant cell size and slowly varying electrical potential (i.e. the jump in ψ - the scaled potential - between cell centres is smaller than 2). It also shows the scheme to be only first order accurate if the ratio of adjacent cell edges differs too much from one or if the number of boundary cells is a large fraction of the total number of cells.

In section 4.2 we formulate a model problem. Section 4.3 discusses the discretisation spaces to be used. We give a description of the discretisation in section 4.4. Section 4.5 gives conditions that imply existence of the solution and stability of the problem. In section 4.6 we describe a quadrature rule. Section 4.7 shows consistency and section 4.8 gives the a-posteriori error estimate. In section 4.9 we summarise our results and draw some conclusions.

4.2 The model equation.

In this chapter we study a model equation for the semi-conductor continuity equation. For a discussion of both numerical and physical aspects of semi-conductor modelling, we refer to the books by Markowich [6], or Selberherr [7], or the papers by Polak et al. [8] or Engl et al. [9]. For a review of numerical aspects of such models, we refer to the articles by Bank et al. [3, 10, 11]. We consider a linearised model for the equations for one of the two charge carrier densities.

$$-a(\mathbf{grad}\, u + u\, \mathbf{grad}\, \psi) = \boldsymbol{\sigma} \text{ on } \Omega , \tag{4.1a}$$

$$\operatorname{div} \boldsymbol{\sigma} = f \text{ on } \Omega , \tag{4.1b}$$

with Dirichlet boundary conditions on some parts of the boundary

$$u \,|_{\Gamma_1} = g , \tag{4.1c}$$

and mixed or Robin boundary conditions on the remaining parts of the boundary

$$\mathbf{grad}\, u \cdot \mathbf{n}_{\partial\Omega} \,|_{\Gamma_2} + u\, \mathbf{grad}\, \psi \cdot \mathbf{n}_{\partial\Omega} \,|_{\Gamma_2} = 0 , \tag{4.1d}$$

where the notation $\mathbf{n}_{\partial A}$ denotes the outward unit normal vector on the boundary of a domain A. In equation (1a), ψ corresponds to the electrical potential scaled by the thermal voltage. We place the following restrictions on the coefficients. We assume that the coefficients a and ψ are continuous and differentiable, $a, \psi \in C^1(\overline{\Omega})$, we also assume that a is bounded away from zero, $\exists \; a_0 > 0 \in \mathbb{R} : a \geqslant a_0$ on Ω, and ψ is piecewise bilinear on Ω. We assume that the function g is continuous and differentiable, $g \in C^1(\partial\Omega)$. Note that the connected subsets of the Dirichlet boundary generally correspond to the contacts of the device. We assume that the right hand side f is square integrable, i.e. $f \in \mathrm{L}^2(\Omega)$, and that $\Gamma_1 \bigcup \Gamma_2 = \partial\Omega$ and $\Gamma_1 \bigcap \Gamma_2 = \varnothing$. We assume that the shape of Ω, Γ_1 and Γ_2 and the conditions on a, ψ, f and g guarantee that $\boldsymbol{\sigma} \in \mathrm{H}^1(\Omega)^2$ and $u \in C(\Omega)$.

For later reference, we give an equivalent system of equations, obtained by a transformation of the dependent variable,

$$\boldsymbol{\sigma} = -a\exp(-\psi)\,\mathbf{grad}\,U\,, \tag{4.2a}$$

$$\operatorname{div}\boldsymbol{\sigma} = f\,, \tag{4.2b}$$

$$U\,|_{\Gamma_1} = \exp(\psi)g\,|_{\Gamma_1}\,, \tag{4.2c}$$

$$\boldsymbol{\sigma}\cdot\mathbf{n}_{\partial\Omega}\,|_{\Gamma_2} = 0\,, \tag{4.2d}$$

where $U = \exp(\psi)u$. Note that, in these variables, condition (2d) implies homogeneous Neumann boundary conditions for U on Γ_2.

We assume that U is square integrable and that $\boldsymbol{\sigma}$ lies in the space,

$$\mathrm{H}(\mathrm{div};\Omega) := \{\, \boldsymbol{\tau} \in \mathbf{L}^2(\Omega) \mid \operatorname{div}\boldsymbol{\tau} \in \mathrm{L}^2(\Omega) \,\}\,,$$

with the inner product,

$$(\boldsymbol{\tau}_1,\boldsymbol{\tau}_2)_{\mathrm{H}(\mathrm{div};\Omega)} = (\boldsymbol{\tau}_1,\boldsymbol{\tau}_2)_{\mathbf{L}^2(\Omega)} + (\operatorname{div}\boldsymbol{\tau}_1, \operatorname{div}\boldsymbol{\tau}_2)_{\mathrm{L}^2(\Omega)} \quad \forall\ \tau_1,\tau_2 \in \mathbf{L}^2(\Omega)\,,$$

where

$$\mathbf{L}^2(\Omega) = \{\, \boldsymbol{\tau}:\Omega\rightarrow\mathbb{R}^2 \mid \int_\Omega \boldsymbol{\tau}\cdot\boldsymbol{\tau}\, d\mu < \infty \,\}$$

with the usual inner product,

$$(\boldsymbol{\tau}_1,\boldsymbol{\tau}_2)_{\mathbf{L}^2(\Omega)} = \int_\Omega \boldsymbol{\tau}_1\cdot\boldsymbol{\tau}_2\, d\mu \quad \forall\ \boldsymbol{\tau}_1,\boldsymbol{\tau}_2 \in \mathbf{L}^2(\Omega)\,.$$

Properties of H(div;Ω) are found in Girault and Raviart [12]. We wish to define a subspace V of H(div;Ω) that contains all elements that satisfy the homogeneous Neumann boundary condition given in (2d). To do this properly, we define this subspace as the closure in H(div;Ω) of the space $\mathscr{V}(\Omega)$ of $C^\infty(\bar{\Omega})$ functions that satisfy the condition (2d),

$$\mathscr{V}(\Omega) := \{\, \boldsymbol{\tau} \in C^\infty(\bar{\Omega})^2 \mid (\boldsymbol{\tau}\cdot\mathbf{n}_{\partial\Omega})\,|_{\partial\Omega} = 0 \ \text{ on } \ \Gamma_2 \,\}\,,$$

where we assume that Γ_1 and Γ_2 are such that this definition makes sense. Now V is by definition a closed subspace of H(div;Ω) and a Hilbert space for the H(div;Ω) inner product.

4.3 The discretisation spaces.

The Scharfetter-Gummel discretisation can best be interpreted as a finite volume scheme, so we need an mesh of finite volumes, which we call the primary or finite volume mesh, and a dual mesh with the cell centres of the original mesh as vertices. In addition, the dual mesh needs vertices on the centres of those edges of finite volumes that lie on the boundary of the domain. For that purpose we add cells of zero thickness to the finite volume mesh, to avoid the need for special formulas that refer to dual mesh vertices on the domain edge. We restrict ourselves to rectangular domains and to partitions of Ω that are Cartesian products of partitions of the sides of the rectangle. We assume, that the boundaries between Γ_1 and Γ_2 coincide with vertices of the mesh. We assume, that ψ is piecewise bilinear on the cells of the dual mesh.

We use a Cartesian coordinate system and we position our rectangular domain Ω as follows,

$$\Omega =]0,L_1[\times]0,L_2[\ . \tag{4.3}$$

We use the following naming conventions. The horizontal unit vector is denoted by $\mathbf{e}_1$ and the vertical unit vector is denoted by $\mathbf{e}_2$. All lower case bold letters are vectors, the corresponding lower case italic letters with subscript 1 or 2 are the vector components in the horizontal or vertical direction.

4.3.1. The partition.

To introduce names for the vertices and cells we need to specify the partitions of the sides of our domain. We use the letter P for the partition of the horizontal axis and the letter Q for the partition of the vertical axis,

$$P = \{ \ 0 = p_{-1} = p_0 < p_1 < \cdots < p_{N_1} = p_{N_1+1} = L_1 \ \} \ , \tag{4.4}$$

$$Q = \{ \ 0 = q_{-1} = q_0 < q_1 < \cdots < q_{N_2} = q_{N_2+1} = L_2 \ \} \ , \tag{4.5}$$

where we added p_{-1}, p_{N_1+1}, q_{-1}, q_{N_2+1} to take into account the zero-width boundary cells. The partition of Ω is given by $P \times Q$. In the obvious way we introduce a notation for particular points in the primary and in the dual mesh. First, the vertices of the primary mesh,

$$\begin{aligned} \mathbf{x}_{i,j} = (p_i, q_j)^T \ \text{ for } \ & i = -1,0,1,2, \ldots, N_1+1 \ , \\ & j = -1,0,1,2, \ldots, N_2+1 \ . \end{aligned} \tag{4.6}$$

We denote the vertices of the dual cells by,

$$\begin{aligned} \mathbf{x}_{i-½,j-½} = \frac{\mathbf{x}_{i-1,j-1} + \mathbf{x}_{i,j}}{2} \ \text{ for } \ & i = 0,1,2, \ldots, N_1+1 \ , \\ & j = 0,1,2, \ldots, N_2+1 \ . \end{aligned} \tag{4.7}$$

Finally, we introduce,

$$\mathbf{x}_{i,j-½} = \frac{\mathbf{x}_{i,j-1} + \mathbf{x}_{i,j}}{2} \ \text{ for } \ i = 0,1,2, \ldots, N_1 \ , j = 1,2, \ldots, N_2 \ . \tag{4.8}$$

and

$$\mathbf{x}_{i-½,j} = \frac{\mathbf{x}_{i-1,j} + \mathbf{x}_{i,j}}{2} \ \text{ for } \ i = 1,2, \ldots, N_1 \ , j = 0,1,2, \ldots, N_2 \ . \tag{4.9}$$

We denote the finite volumes, i.e. the cells of the partition $P \times Q$ by,

$$\begin{aligned} &\Omega_{i-½,j-½} = \\ &\{ \ \mathbf{x} \,|\, \mathbf{x}_{i-1,j-1} < \mathbf{x} < \mathbf{x}_{i,j} \ \} \ \text{ for } \ i = 1,2, \ldots, N_1 \ , j = 1,2, \ldots, N_2 \ , \end{aligned} \tag{4.10}$$

where the notation

$$\mathbf{a} < \mathbf{b} , \tag{4.11}$$

has its usual meaning, i.e.

$$\mathbf{a} < \mathbf{b} \Leftrightarrow a_1 < b_1 \text{ and } a_2 < b_2 . \tag{4.12}$$

Similarly,

$$\Gamma_{i-½,j} = \{ \mathbf{x} \,|\, \mathbf{x}_{i-1,j} \leqslant \mathbf{x} \leqslant \mathbf{x}_{i,j} \} \tag{4.13}$$
$$\text{for } i=1,2, \ldots ,N_1 , j=0,1,2, \ldots ,N_2 ,$$

and

$$\Gamma_{i,j-½} = \{ \mathbf{x} \,|\, \mathbf{x}_{i,j-1} \leqslant \mathbf{x} \leqslant \mathbf{x}_{i,j} \} \tag{4.14}$$
$$\text{for } i=0,1,2, \ldots ,N_1 , j=1,2, \ldots ,N_2 .$$

In our error analysis in section 4.7, we also need to identify the cells and edges of the dual mesh, these are denoted by,

$$\hat{\Omega}_{i,j} = \{ \mathbf{x} \,|\, \mathbf{x}_{i-½,j-½} < \mathbf{x} < \mathbf{x}_{i+½,j+½} \} \tag{4.15}$$
$$\text{for } i=0,1,2, \ldots ,N_1 , j=0,1,2, \ldots ,N_2 ,$$
$$\hat{\Gamma}_{i-½,j} = \{ \mathbf{x} \,|\, \mathbf{x}_{i-½,j-½} \leqslant \mathbf{x} \leqslant \mathbf{x}_{i-½,j+½} \} \tag{4.16}$$
$$\text{for } i=1,2, \ldots ,N_1 , j=0,1,2, \ldots ,N_2 ,$$

and

$$\hat{\Gamma}_{i,j-½} = \{ \mathbf{x} \,|\, \mathbf{x}_{i-½,j-½} \leqslant \mathbf{x} \leqslant \mathbf{x}_{i+½,j-½} \} \tag{4.17}$$
$$\text{for } i=0,1,2, \ldots ,N_1 , j=1,2, \ldots ,N_2 .$$

Note that, at the start of this section, we assumed $\psi|_{\hat{\Omega}_{i,j}}$ to be bilinear. This implies that $\psi|_{\hat{\Gamma}_r}$ is linear for all $r \in \tilde{E}$, where $\tilde{E}$ is the collection of index tuples of edge centres,

$$\tilde{E} = \{ e=(i,j-½) \,|\, i=0,1,2, \ldots ,N_1 , j=1,2, \ldots ,N_2 \} \cup$$
$$\{ e=(i-½,j) \,|\, i=1,2, \ldots ,N_1 , j=0,1,2, \ldots ,N_2 \} .$$

We indicate the set of indices of all edges that are not on the Neumann boundary by,

$$E = \{ e \in \tilde{E} \,|\, \Gamma_e \subset \bar{\Omega} - \Gamma_2 \} .$$

Finally, we define the set of index tuples of cell centres, by

$$M = \{ e=(i-½,j-½) \,|\, i=1,2, \ldots ,N_1 , j=1,2, \ldots ,N_2 \} ,$$

and we extend the definition of the Kronecker-δ to index tuples,

$$\delta_{rs} = \begin{cases} 1 \text{ if } r=s , \\ 0 \text{ if } r \neq s . \end{cases}$$

4.3.2. Local coordinates.

When we analyse the quadrature rules - in section 4.6 - and the discretisation - in section 4.7 - it is convenient to have at our disposal a local coordinate system with its origin at the intersection of a primary and a dual mesh line. We define this system as follows. Take a unit vector $\mathbf{e}_{x,r}$ parallel to $\hat{\Gamma}_r$, and let the direction of increasing coordinates correspond to the direction of increasing coordinates in the global coordinate system given at the start of section 4.3. Take a unit vector $\mathbf{e}_{y,r}$ parallel to Γ_r and directed to give a right hand coordinate system when combined with $\mathbf{e}_{x,r}$. I.e. $\mathbf{e}_{x,r}$ is a normal vector on Γ_r and $\mathbf{e}_{y,r}$ is a normal vector on $\hat{\Gamma}_r$. We shall use the letters x and y for local coordinates, so if $\mathbf{x}$ is an arbitrary position vector in the global coordinate system and $\mathbf{x}_r$ is the global coordinate vector of the intersection of Γ_r and $\hat{\Gamma}_r$, then

$$\mathbf{x} = \mathbf{x}_r + x\mathbf{e}_{x,r} + y\mathbf{e}_{y,r} \ .$$

When we use the terms left and right, we shall mean left and right with respect to the local coordinate system. We denote the length of Γ_r by $h_{r,y} = \lambda(\Gamma_r)$. So, the highest local coordinate on Γ_r is $y = \frac{1}{2}h_{r,y}$. We denote the width of the cell to the left of Γ_r - i.e. the cell to the left of the origin of the local coordinate system - by $h_{r,L}$, we denote the width of the cell to the right of Γ_r by $h_{r,R}$. So, the highest local coordinate on $\hat{\Gamma}_r$ is $x = \frac{1}{2}h_{r,R}$. If $\mathbf{x}_r$ lies on the boundary of Ω where the global coordinate along $\hat{\Gamma}_r$ is highest, then we have a cell with width $h_{r,R} = 0$ to the right of Γ_r. The same holds at the other boundary.

We construct a function ψ_r on each $\hat{\Gamma}_r$,

$$\psi_r(x) = \psi(\mathbf{x}_r + x\mathbf{e}_{x,r}) \ . \tag{4.18a}$$

By linearity, we can write this as,

$$\psi_r(x) = \beta_r x + \gamma_r \ , \ x \in [-\tfrac{1}{2}h_{r,L}, \tfrac{1}{2}h_{r,R}] \ . \tag{4.18b}$$

We define ϕ_r to be the difference between the values of ψ in the two cell centres,

$$\phi_r(x) = \beta_r \frac{h_{r,L} + h_{r,R}}{2} \ . \tag{4.18c}$$

To have a convenient notation, we introduce a special notation σ_r for the $\mathbf{e}_{x,r}$ component of a continuous vector valued function $\boldsymbol{\sigma}$ given as a function of local coordinates, i.e.

$$\sigma_r(x,y) = \boldsymbol{\sigma}(\mathbf{x}_r + x\mathbf{e}_{x,r} + y\mathbf{e}_{y,r})\cdot\mathbf{e}_{x,r} \ . \tag{4.19}$$

4.3.3. Some local projections.

In this chapter we need several projections that are mesh dependent. To simplify their definition, we introduce a notation for the average of a function over a given area or a given line segment. We denote the average over an area

A by,

$$P[A](f) = \frac{1}{\mu(A)} \int_A f \, d\mu \, , \tag{4.20}$$

for all measurable and bounded $A \subset \Omega$ with $\mu(A) > 0$ and all f, integrable over A, where μ is the Lebesgue measure on $\mathbb{R}^2$. We denote the average over a line segment Γ by,

$$P[\Gamma](f) = \frac{1}{\lambda(\Gamma)} \int_\Gamma f \, d\lambda \, , \tag{4.21}$$

for all measurable finite line segments Γ with $\lambda(\Gamma) > 0$ and all f, integrable over Γ, that lie in $\overline{\Omega}$. Here λ is the Lebesgue measure on $\mathbb{R}$.

4.3.4. Some global projections and the trial spaces.

In this chapter we examine the difference between the solution of (1) and a discrete approximation of that solution. To do this we need to compare a known discrete solution with an unknown continuous solution. We simplify the problem by using a projection $\Pi_h \times \overline{\mathrm{P}}_h$ of the continuous solution onto the trial space $V_h \times W_h$. The problem then reduces to the study of the interpolation error - i.e. the difference between the continuous solution and its projection - and the discretisation error - i.e difference between this projection and our discrete solution -. In general the projection can not be calculated numerically, but its properties and accuracy are known, so the problem reduces to finding a measure for the distance - in the trial space - between the projection and the discrete solution. This approach differs from the standard approach in Hilbert spaces, because the chosen projection is not necessarily orthogonal. However, the approach can also be found in Douglas and Roberts [13]. In this section we describe the trial space $V_h \times W_h$. Using the local projections defined earlier we then construct the global projection $\Pi_h \times \overline{\mathrm{P}}_h : \mathrm{H}^1(\Omega)^2 \times \mathrm{L}^2(\Omega) \rightarrow V_h \times W_h$. We use the lowest order Raviart-Thomas space [14] for our trial space. The subspace V_h of the trial space is spanned by vector valued functions that satisfy the homogeneous Neumann boundary conditions,

$$V_h = Span(\{ \boldsymbol{\eta}_r \mid r \in E \}) \, , \tag{4.22}$$

where the basis vectors $\boldsymbol{\eta}_r$ have a triangular prism shaped components (tent functions),

$$\boldsymbol{\eta}_r(\mathbf{x}_r + x\mathbf{e}_{x,r} + y\mathbf{e}_{y,r}) = \tag{4.23}$$

$$\begin{cases} \dfrac{h_{r,L}+x}{h_{r,L}}\mathbf{e}_{x,r} \ \text{ if } \ (x,y) \in [-h_{r,L},0] \times [-\tfrac{1}{2}h_{r,y}, \tfrac{1}{2}h_{r,y}] \\ \dfrac{h_{r,R}-x}{h_{r,R}}\mathbf{e}_{x,r} \ \text{ if } \ (x,y) \in [0,h_{r,R}] \times [-\tfrac{1}{2}h_{r,y}, \tfrac{1}{2}h_{r,y}] \\ \qquad 0 \ \text{ elsewhere} \end{cases}$$

for all $r \in E$. For W_h, we use a space of piece wise constant functions,

$$W_h = Span(\{ \chi_{\Omega_{i-½,j-½}} \,|\, i=1,2,\ldots,N_1 \,,\, j=1,2,\ldots,N_2 \}) \,, \qquad (4.24)$$

where χ_A is the characteristic function of $A \subset \Omega$, i.e.

$$\chi_A(\mathbf{x})=1 \text{ if } \mathbf{x} \in A \,,\, \chi_A(\mathbf{x})=0 \text{ if } \mathbf{x} \in \Omega - A \,. \qquad (4.25)$$

Next we define the projection $\Pi_h \times \overline{\mathrm{P}}_h$. The map $\overline{\mathrm{P}}_h : C(\Omega) \rightarrow W_h$ is a projection such that the function and its image coincide at cell centres:

$$\overline{\mathrm{P}}_h(f) = \sum_{s \in M} f(\mathbf{x}_s)\chi_{\Omega_s} \,. \qquad (4.26)$$

and the mapping $\Pi_h : \mathrm{H}^1(\Omega)^2 \rightarrow V_h$, taken from [14], is given by,

$$\Pi_h(\mathbf{f}) = \sum_{r \in E} P[\Gamma_r](\boldsymbol{f} \cdot \mathbf{e}_{x,r})\boldsymbol{\eta}_r \,. \qquad (4.27)$$

These are the basic ingredients for our calculations. However, we still need some other definitions associated with cell edges. We define the space E_h spanned by the characteristic functions of dual cell edges,

$$E_h = Span(\{ \chi_{\hat{\Gamma}_r} \,|\, r \in E \}) \,,$$

and the space G_h,

$$G_h = Span(\{ \chi_{\Gamma_r} \,|\, \Gamma_r \subset \Gamma_1 \}) \,,$$

and we introduce a map $Q_h : V_h \rightarrow E_h$, similar to $\overline{\mathrm{P}}_h$,

$$Q_h(\mathbf{f}) = \sum_{r \in E} \mathbf{f}(\mathbf{x}_r) \cdot \mathbf{e}_{x,r} \chi_{\hat{\Gamma}_r} \,. \qquad (4.28)$$

Finally, we define the additional global projection, $\mathrm{P_h} : \mathrm{L}^2(\Omega) \rightarrow W_h$,

$$\mathrm{P_h}(f) = \sum_{s \in M} P[\Omega_s](f)\chi_{\Omega_s} \,. \qquad (4.29)$$

and we notice that the pair of projections Π_h and $\mathrm{P_h}$ are those discussed by Raviart and Thomas [14].

4.4 Discretisation of the system.

We construct a scheme for the approximation of the solution $(\boldsymbol{\sigma},u)$ of (1). We proceed as follows. We formulate the set of integral equations that hold for the solution of (1) and that correspond to the classical finite volume equations. We then write this set of equations as a saddle-point problem. Finally we replace exact integration by quadrature rules where appropriate.

Given a (horizontal) dual mesh edge $\hat{\Gamma}_{i,j-½}$, the following formula holds for the solution $(\boldsymbol{\sigma},u)$ of (1):

$$(\exp(\psi)u)\left(\frac{p_{i+1}+p_i}{2}, \frac{q_{j-1}+q_j}{2}\right) - (\exp(\psi)u)\left(\frac{p_i+p_{i-1}}{2}, \frac{q_{j-1}+q_j}{2}\right) = (A)$$

$$- \int_{x_1=\frac{p_{i-1}+p_i}{2}}^{\frac{p_i+p_{i+1}}{2}} \left[\frac{\exp(\psi)}{a} \boldsymbol{\sigma} \cdot \mathbf{e}_1 \right] \left[x_1, \frac{q_{j-1}+q_j}{2} \right] dx_1 .$$

This follows immediately from (2a). An analogous formula is derived for a vertical dual mesh edge. In this way we find one equation for each dual mesh edge. Note that, if $\mathbf{x}_r$ lies on the Dirichlet part of the boundary, then one of the endpoints of the integration coincides with $\mathbf{x}_r$ and $u(\mathbf{x}_r)$ is given by $g(\mathbf{x}_r)$. For each cell Ω_m of the primary mesh, (2b) implies that

$$\int_{\partial\Omega_m} \boldsymbol{\sigma} \cdot \mathbf{n}_{\partial\Omega_m} \, d\lambda = \int_{\Omega_m} f \, d\mu . \tag{B}$$

This gives us an equation for each cell Ω_m. The set of equations given above is the starting point for our derivation of a finite volume version of the Scharfetter-Gummel scheme. Our derivation is a variation on the derivation of a finite volume scheme as given in [15].

We introduce some notation in order to write this in the form of a saddle point problem. We define two operators $\mathscr{E}: W_h \rightarrow W_h$ and $\mathscr{E}_{\partial\Omega}: C(\partial\Omega) \rightarrow G_h$, and two bilinear forms, $\alpha_{SG}: V \times E_h \rightarrow \mathbb{R}$ and $b: V_h \times W_h \rightarrow \mathbb{R}$:

$$\mathscr{E} t_h = \sum_{s \in M} \exp(\psi(\mathbf{x}_s)) t_h(\mathbf{x}_s) \chi_{\Omega_s} \quad \forall \; t \in W_h , \tag{4.30}$$

$$(\mathscr{E}_{\partial\Omega} g)|_{\Gamma_r} = \exp(\psi(\mathbf{x}_r)) g(\mathbf{x}_r) \chi_{\Gamma_r} \quad \forall \; \Gamma_r \subset \Gamma_1 ,$$

$$\alpha_{SG}(\boldsymbol{\sigma}, Q_h \boldsymbol{\eta}_r) := \lambda(\Gamma_r) \int_{\hat{\Gamma}_r} \exp(\psi) \, \boldsymbol{\sigma} \cdot \mathbf{e}_{x,r} \, d\lambda \quad \forall \; r \in E \quad \forall \; \boldsymbol{\sigma} \in \mathrm{H}^1(\Omega)^2 , \tag{4.31}$$

and

$$b(\boldsymbol{\tau}_h, t_h) := \int_{\Omega} \operatorname{div} \boldsymbol{\tau}_h \; t_h \, d\mu \quad \forall \; \boldsymbol{\tau}_h \in V_h , \; t_h \in W_h .$$

Using these definitions, we can write the equations as follows,

$$\alpha_{SG}(\boldsymbol{\sigma}, Q_h \boldsymbol{\eta}_r) - b(\boldsymbol{\eta}_r, \mathscr{E} \overline{\mathrm{P}}_h u) = - < \mathscr{E}_{\partial\Omega} g, \boldsymbol{\eta}_r \cdot \mathbf{n}_{\partial\Omega} > \quad \forall \; r \in E , \tag{4.32a}$$

and

$$b(\boldsymbol{\sigma}, t_h) = (f, t_h) \quad \forall \; t_h \in W_h . \tag{4.32b}$$

Equation (32a) corresponds with (A) and gives a relation between the current along an edge and the value of u at the endpoints of that edge. Equation (32b) corresponds with (B) and gives a relation between the currents through the different edges of a given cell. Note that in the form α_{SG} the basis vector $\boldsymbol{\eta}_r$ just serves to indicate the edge over which the integration takes place. We shall use the same convention in the quadrature rule for α_{SG}. We obtain our discrete system by replacing $\boldsymbol{\sigma}$ by $\boldsymbol{\sigma}_h$, $\overline{\mathrm{P}}_h u$ by u_h and α_{SG} by a quadrature rule α_h. The discrete system has the form,

$$(\boldsymbol{\sigma}_h,u_h) \in V_h \times W_h \ ,$$

$$\alpha_h(\boldsymbol{\sigma}_h,Q_h\boldsymbol{\tau}_h) \ - \ b(\boldsymbol{\tau}_h,\mathscr{E}u_h) \ = \ - \ < \ \mathscr{E}_{\partial\Omega}g,\boldsymbol{\tau}_h\cdot\mathbf{n}_{\partial\Omega} \ > \quad \forall \ \ \boldsymbol{\tau}_h \in V_h \ , \tag{4.33a}$$

$$b(\boldsymbol{\sigma}_h,t_h) \ = \ (f,t_h) \quad \forall \ \ t_h \in W_h \ . \tag{4.33b}$$

We discuss a specific quadrature rule α_h for α_{SG} in section 4.6. To facilitate the study of the properties of different versions of α_h, we introduce a bilinear form $(.,.)_h \colon V_h \times E_h \rightarrow \mathbb{R}$,

$$(\boldsymbol{\sigma}_h,Q_h\boldsymbol{\tau}_h)_h := \sum_{r \in E} \mu_r \Pi[\Gamma_r](\boldsymbol{\sigma}_h)\cdot\mathbf{e}_{x,r} \ \Pi[\Gamma_r](\boldsymbol{\tau}_h)\cdot\mathbf{e}_{x,r} \quad \forall \ \ \boldsymbol{\sigma}_h \ , \boldsymbol{\tau}_h \in V_h \ ,$$

where μ_r is

$$\mu_r \ = \ \lambda(\Gamma_r) \ \lambda(\hat{\Gamma}_r) \quad \forall \ \ r \in E \ ,$$

and λ is the Lebesgue measure on $\mathbb{R}$. The bilinear form $(.,.)_h$ is a weighted version of the Euclidean inner product on V_h. We prove that in V_h the norm derived from this inner product and the $\mathbf{L}^2(\Omega)$-norm are equivalent. Note that the norm corresponding to $(.,.)_h$ resembles the norm $\|.\|_h$ introduced in chapter 2.

Lemma 4.1.

$$\|\boldsymbol{\sigma}_h\|^2_{\mathbf{L}^2(\Omega)} \ \leqslant \ (\boldsymbol{\sigma}_h,Q_h\boldsymbol{\sigma}_h)_h \ \leqslant \ 3\,\|\boldsymbol{\sigma}_h\|^2_{\mathbf{L}^2(\Omega)} \ ,$$

where $\mathbf{L}^2(\Omega) = \mathrm{L}^2(\Omega)^2$.

Proof.

We start by determining the value of $(\boldsymbol{\sigma}_h,\boldsymbol{\sigma}_h)_{\mathrm{L}^2(\Omega)}$. To simplify matters, we introduce coordinates s_r for $\boldsymbol{\sigma}_h$ with respect to the basis $\boldsymbol{\eta}_r$ given in (23) and we split $\boldsymbol{\sigma}_h$ into mutually orthogonal $\mathbf{e}_i$ parts $(i=1,2)$,

$$\boldsymbol{\sigma}_{1,j} \ = \ \sum_{i=0}^{N_1} \boldsymbol{\eta}_{i,j-½} s_{i,j-½} \ ,$$

and

$$\boldsymbol{\sigma}_{2,i} \ = \ \sum_{j=0}^{N_2} \boldsymbol{\eta}_{i-½,j} s_{i-½,j} \ .$$

Now,

$$(\boldsymbol{\sigma}_h,\boldsymbol{\sigma}_h)_{\mathrm{L}^2(\Omega)} \ = \ \sum_{i=1}^{N_1} (\boldsymbol{\sigma}_{2,i},\boldsymbol{\sigma}_{2,i})_{\mathrm{L}^2(\Omega)} \ + \ \sum_{j=1}^{N_2} (\boldsymbol{\sigma}_{1,j},\boldsymbol{\sigma}_{1,j})_{\mathrm{L}^2(\Omega)} \ .$$

We see immediately, that

$$(\boldsymbol{\sigma}_{1,j},\boldsymbol{\sigma}_{1,j})_{\mathrm{L}^2(\Omega)} \ = \ \sum_{i=1}^{N_1} \frac{1}{3}(s^2_{i-1,j-½} + s^2_{i,j-½} + s_{i-1,j-½}s_{i,j-½})\mu(\Omega_{i-½,j-½}) \ ,$$

so

$$\frac{1}{6}\sum_{i=1}^{N_1}(s^2_{i-1,j-½} + s^2_{i,j-½})\mu(\Omega_{i-½,j-½}) \ \leqslant \ (\boldsymbol{\sigma}_{1,j},\boldsymbol{\sigma}_{1,j})_{\mathrm{L}^2(\Omega)} \ \leqslant$$

$$\frac{1}{2}\sum_{i=1}^{N_1}(s_{i-1,j-1/2}^2+s_{i,j-1/2}^2)\mu(\Omega_{i-1/2,j-1/2}) \; .$$

Furthermore,

$$(\boldsymbol{\sigma}_{1,j},\boldsymbol{\sigma}_{1,j})_h \;=\; \frac{1}{2}\sum_{i=0}^{N_1-1}\mu(\Omega_{i+1/2,j-1/2})s_{i,j-1/2}^2 \;+\; \frac{1}{2}\sum_{i=1}^{N_1}\mu(\Omega_{i-1/2,j-1/2})s_{i,j-1/2}^2 \; .$$

□

4.5 Existence and uniqueness of the solution.

In this section we give sufficient conditions for the existence and uniqueness of the solution of the discrete system.

We plan to use theorem 3.1 from Nicolaides [5] to prove existence, uniqueness and stability for the discrete scheme. To apply the theorem, we need to define norms on our discrete spaces and to verify the conditions (2.1, 2,2, 3.1 and 3.2) given in [5]. We shall use the norms associated with the following inner products on V_h and W_h,

$$(\boldsymbol{\sigma}_h,\boldsymbol{\tau}_h)_{V_h} \;=\; (\boldsymbol{\sigma}_h,\boldsymbol{\tau}_h)_h \;+\; (\operatorname{div}\boldsymbol{\sigma}_h, \operatorname{div}\boldsymbol{\tau}_h)_{L^2(\Omega)} \quad \forall\; \boldsymbol{\sigma},\boldsymbol{\tau}\in V_h \; , \qquad (4.34a)$$

and

$$(u_h,t_h)_{W_h} \;=\; (u_h,t_h)_{L^2(\Omega)} \quad \forall\; u_h,t_h\in W_h \; . \qquad (4.34b)$$

The conditions 1, 2 and 3 that follow are equivalent to the conditions (2.1, 2,2, 3.1 and 3.2) imposed by Nicolaides.

Condition 1.

The bilinear form α_h is bounded, i.e. there is a $0 \;<\; A\in\mathbb{R}$, independent of the mesh, such that

$$\alpha_h(\boldsymbol{\sigma}_h,\boldsymbol{\tau}_h) \;\leqslant\; A\,||\,\boldsymbol{\sigma}_h\,||_{V_h}\,||\,\boldsymbol{\tau}_h\,||_{V_h} \; ,$$

and α_h is coercive on the kernel of the divergence operator in V_h, i.e. there exists a $0 \;<\; \delta\in\mathbb{R}$, independent of the mesh, such that

$$\delta\;(\boldsymbol{\sigma}_h,\boldsymbol{\sigma}_h)_h \;\leqslant\; \alpha_h(\boldsymbol{\sigma}_h,\boldsymbol{\sigma}_h) \; ,$$

for all $\boldsymbol{\sigma}_h\in V_h\bigcap\mathcal{N}(\operatorname{div})$. Our condition 1 corresponds to conditions (3.1) and (3.2) in the paper by Nicolaides [5].

Condition 2

The bilinear form b is bounded and there exists a $0 \;<\; \gamma'\in\mathbb{R}$, independent of the mesh, such that

$$\sup_{0\neq\boldsymbol{\tau}_h\in V_h}\frac{|\,b(\boldsymbol{\tau}_h,t_h)\,|}{||\,\boldsymbol{\tau}_h\,||_{V_h}} \;\geqslant\; \gamma'\,||\,t_h\,||_{L^2(\Omega)} \; .$$

Condition 3

There exists a $0 < \gamma \in \mathbb{R}$, independent of the mesh, such that

$$\sup_{0 \neq w \in W_h} \frac{|b(\boldsymbol{\sigma}, w)|}{\| w \|_{W_h}} \geqslant \gamma \inf_{\mathbf{z} \in (\mathcal{N}(\mathrm{div}) \bigcap V_h)} \| \boldsymbol{\sigma} - \mathbf{z} \|_{V_h} \quad \forall \ \boldsymbol{\sigma} \in V_h \ .$$

Our conditions 2 and 3 correspond to conditions (2.1) and (2.2) in [5]. We can now give a version of theorem 3.1 of Nicolaides.

Theorem 4.1.
If the conditions 1, 2 and 3 are satisfied, then the discrete system (33) has a unique solution and the norm of the solution is bounded by,

$$\| \boldsymbol{\sigma}_h \|_{V_h} \leqslant \frac{1}{\delta} \| \mathscr{E} g \|_{V_h^*} + \frac{1}{\gamma} \left(1 + \frac{A}{\delta} \right) \| f \|_{\mathrm{L}^2(\Omega)} \ ,$$

$$\| \mathscr{E} u_h \|_{\mathrm{L}^2(\Omega)} \leqslant \frac{1}{\gamma'} \left(A \, \| \boldsymbol{\sigma}_h \|_{V_h} + \| \mathscr{E} g \|_{V_h^*} \right) .$$

Proof.
The proof is a direct application of theorem 3.1 in [5].

□

Now, we have to ask ourselves when these conditions are satisfied. In section 4.6, we shall introduce an α_h that satisfies condition 1. Because the remaining two conditions are not easily verified in the form given here, we give alternative conditions 2a and 3a that are easier to verify. Lemma 2 shows that 2a and 3a imply 2 and 3.

Condition 2a.

The corresponding Poisson problem is regular, i.e. Ω, $\Gamma_1, \Gamma_2 \subset \partial\Omega$ are such that there exists a $C > 0$, $C \in \mathbb{R}$ such that

$$\forall \ f \in \mathrm{L}^2(\Omega) \ \exists! \ u \in \mathrm{H}^2(\Omega) \ :$$

$$\Delta u = f \ \text{ on } \ \Omega \ ,$$

$$u = 0 \ \text{ on } \ \Gamma_1 \ ,$$

$$\mathbf{grad}\, u \cdot \mathbf{n}_{\partial\Omega} = 0 \ \text{ on } \ \Gamma_2 \ ,$$

$$\| u \|_{\mathrm{H}^2(\Omega)} \leqslant C \| f \|_{\mathrm{L}^2(\Omega)} \ ,$$

Condition 3a

The map Π_h has the following approximation property, there is a $K > 0$, $K \in \mathbb{R}$, independent of the mesh, such that

$$\| \mathbf{grad}\, u - \Pi_h \, \mathbf{grad}\, u \|_{\mathbf{L}^2(\Omega)} \leqslant KH \| u \|_{\mathrm{H}^2(\Omega)} \ ,$$

where H is the maximum mesh diameter.

Assuming that 2a and 3a do indeed imply 2 and 3, it remains to see when 2a and 3a are satisfied. For condition 2a we refer to Grisvard [16]. Condition 3a follows almost immediately from the assumption that $\boldsymbol{\sigma}$ has components in $\mathrm{H}^1(\Omega)$. To illustrate this, we prove Lemma 3. This lemma proves that Π_h has the approximation property.

Lemma 4.2.
If all mesh edges have a length that is bounded above by a constant H_0, then the conditions 2a and 3a imply conditions 2 and 3 with $\gamma = \gamma' = \dfrac{1}{3C(1+KH_0)}$.

Proof.
Assume that (2a) and (3a) hold and take a fixed $w \in W_h$. If we solve the Poisson problem for $f = w$ then, according to (2a), the solution u_w satisfies,

$$\| \mathbf{grad}\, u_w \|_{\mathrm{H(div;\Omega)}} \leqslant C \| w \|_{\mathrm{L^2(\Omega)}} \ ,$$

and

$$(\operatorname{div} \mathbf{grad}\, u_w, w) = \| w \|^2_{\mathrm{L^2(\Omega)}} \ .$$

Furthermore, (3a) implies that for all $u \in \mathrm{H}^2(\Omega)$

$$\| \mathbf{grad}\, u - \Pi_h \mathbf{grad}\, u \|_{\mathbf{L}^2(\Omega)} \leqslant KH_0 \| u \|_{\mathrm{H^2(\Omega)}} \ .$$

So we find, that

$$\mathrm{P_h} \operatorname{div} \mathbf{grad}\, u_w = \mathrm{P_h} w \ ,$$

and

$$\| \Pi_h \mathbf{grad}\, u_w \|_{\mathrm{H(div;\Omega)}} \leqslant$$
$$H_0 K \| u_w \|_{\mathrm{H^2(\Omega)}} + \| \mathbf{grad}\, u_w \|_{\mathbf{L}^2(\Omega)} + \| \operatorname{div} \Pi_h \mathbf{grad}\, u_w \|_{\mathrm{L^2(\Omega)}} \ .$$

Our w lies in W_h, so special properties of $\mathrm{P_h}$ and Π_h imply,

$$\operatorname{div} \Pi_h \mathbf{grad}\, u_w = \mathrm{P_h} w = w \ .$$

We combine this with condition (3a) to find,

$$\| \Pi_h \mathbf{grad}\, u_w \|_{\mathrm{H(div;\Omega)}} \leqslant$$
$$\| u_w \|_{\mathrm{H^2(\Omega)}} + H_0 K \| u_w \|_{\mathrm{H^2(\Omega)}} \leqslant C(1+H_0K) \| w \|_{\mathrm{L^2(\Omega)}} \ .$$

We see immediately, that

$$\| \operatorname{div} \|_{\mathscr{L}(V_h, W_h)} \leqslant 1 \ ,$$

and

$$\forall \ w \in W_h \ \ \exists \ \boldsymbol{\tau}_h \in V_h : \operatorname{div} \boldsymbol{\tau} = w \ \text{ and } \ \| \boldsymbol{\tau}_h \|_{\mathrm{H(div;\Omega)}} \leqslant C(1+H_0K) \| w \|_{\mathrm{L^2(\Omega)}} \ .$$

Lemma 1 implies, that

$$\| \boldsymbol{\tau}_h \|_{\mathrm{H(div;\Omega)}} \leqslant \| \boldsymbol{\tau}_h \|_{V_h} \leqslant \sqrt{3} \| \boldsymbol{\tau}_h \|_{\mathrm{H(div;\Omega)}} \quad \forall \ \boldsymbol{\tau}_h \in V_h \ .$$

We find, that

$$\frac{|\,b(\Pi_h\,\mathbf{grad}\,u_w, w)\,|}{\|\,\Pi_h\,\mathbf{grad}\,u_w\,\|_{V_h}} \geqslant \frac{1}{3C(1+KH_0)}\,\|\,w\,\|_{\mathrm{L}^2(\Omega)} \ .$$

Now suppose $\boldsymbol{\sigma}_h \in V_h$. Then the above derivation implies, that there is a $\boldsymbol{\tau}_h \in V_h$ such that $\operatorname{div}\boldsymbol{\tau}_h = \operatorname{div}\boldsymbol{\sigma}_h$ and $\|\,\boldsymbol{\tau}_h\,\|_{\mathrm{H}(\mathrm{div};\Omega)} \leqslant C(1+H_0K)\,\|\,\operatorname{div}\boldsymbol{\sigma}_h\,\|_{\mathrm{L}^2(\Omega)}$. Moreover, $\boldsymbol{\tau}_h - \boldsymbol{\sigma}_h \in \mathcal{N}(\operatorname{div})$, so

$$\inf_{z_h \in V_h \bigcap \mathcal{N}(\operatorname{div})} \|\,\boldsymbol{\sigma}_h + z_h\,\|_{V_h} \leqslant \|\,\boldsymbol{\tau}_h\,\|_{V_h} \leqslant 3C(1+H_0K)\,\|\,\operatorname{div}\boldsymbol{\sigma}_h\,\|_{\mathrm{L}^2(\Omega)} \ .$$

So we find,

$$\frac{(\operatorname{div}\boldsymbol{\sigma}_h, \operatorname{div}\boldsymbol{\sigma}_h)_{\mathrm{L}^2(\Omega)}}{\|\,\operatorname{div}\boldsymbol{\sigma}_h\,\|_{\mathrm{L}^2(\Omega)}} = \|\,\operatorname{div}\boldsymbol{\sigma}\,\|_{\mathrm{L}^2(\Omega)} \geqslant$$

$$\frac{1}{3C(1+H_0K)} \inf_{z_h \in (\mathcal{N}(\operatorname{div}) \bigcap V_h)} \|\,\boldsymbol{\sigma}_h - z_h\,\|_{V_h} \quad \forall\ \boldsymbol{\sigma} \in {}_hV_h \ .$$

This implies,

$$\sup_{0 \neq w \in W_h} \frac{|\,b(\boldsymbol{\sigma},w)\,|}{\|\,w\,\|_{W_h}} \geqslant \frac{1}{3C(1+H_0K)} \inf_{z \in (\mathcal{N}(\operatorname{div}) \bigcap V_h)} \|\,\boldsymbol{\sigma} - z\,\|_{V_h} \quad \forall\ \boldsymbol{\sigma} \in V_h \ .$$

□

The inequality proved in the following lemma is also included in Lemma 5.1 of chapter 5.

Lemma 4.3.
If f is a square integrable function with square integrable derivatives on a rectangle $\Omega=[0,h_1]\times[0,h_2]$ with sides $\Gamma_{1,1}=\{\ h_1\ \}\times[0,h_2]$, $\Gamma_{2,1}=[0,h_1]\times\{\ h_2\ \}$, $\Gamma_{1,0}=\{\ 0\ \}\times[0,h_2]$ and $\Gamma_{2,0}=[0,h_1]\times\{\ 0\ \}$, then the following inequality holds for all $s \in \mathrm{L}^\infty([0,h_1])$ and $\mathscr{R}(s) \subset [0,1]$,

$$\|\,f-(1-s)\Pi[\Gamma_{1,0}]f - s\Pi[\Gamma_{1,1}]f\,\|_{\mathrm{L}^2(\Omega)} \leqslant \sqrt{2}\,(h_1^2+h_2^2)^{1/2}\,\|\,\mathbf{grad}\,f\,\|_{\mathrm{L}^2(\Omega)} \ .$$

Proof.
We start by proving the above inequality for $f \in C^1(\Omega)$. Then we can extend the inequality by density to $\mathrm{H}^1(\Omega)$. We see that,

$$\|\,f-(1-s)\Pi[\Gamma_{1,0}]f - s\Pi[\Gamma_{1,1}]f\,\|^2_{\mathrm{L}^2(\Omega)} =$$

$$\int_{x=0}^{h_1}\int_{y=0}^{h_2} \left\{ \frac{1}{h_2}\int_{z=0}^{h_2} \Big[(1-s(x))(f(x,y)-f(0,z)) \ + \ s(x)(f(x,y)-f(h_1,z)) \Big] dz \right\}^2 dxdy \ .$$

We use partial derivatives to rewrite the expression,

$$\|\,f-(1-s)\Pi[\Gamma_{1,0}]f - s\Pi[\Gamma_{1,1}]f\,\|^2_{\mathrm{L}^2(\Omega)} =$$

$$\int_{x=0}^{h_1}\int_{y=0}^{h_2}\left[\frac{1}{h_2}\int_{z=0}^{h_2}\left[(1-s(x))\left\{\int_{a=0}^{x}\frac{\partial f}{\partial a}(a,z)da + \int_{b=z}^{y}\frac{\partial f}{\partial b}(x,b)db\right\} + s(x)\left\{\int_{a=h_1}^{x}\frac{\partial f}{\partial a}(a,z)da + \int_{b=z}^{y}\frac{\partial f}{\partial b}(x,b)db\right\}\right]dz\right]^2 dxdy \ .$$

We use Hölder and extend the integrals where appropriate,

$$||f-(1-s)\Pi[\Gamma_{1,0}]f-s\Pi[\Gamma_{1,1}]f||^2_{\mathrm{L}^2(\Omega)} \leqslant$$
$$\int_{x=0}^{h_1}\int_{y=0}^{h_2}\left[\frac{h_1^{\frac12}}{h_2^{\frac12}}\,||\partial f/\partial x_1||_{\mathrm{L}^2(\Omega)} + h_2^{\frac12}\left[\int_{b=0}^{h_2}\left[\frac{\partial f}{\partial b}(x,b)\right]^2 db\right]^{\frac12}\right]^2 dxdy \ .$$

We use $(|A|+|B|)^2 \leqslant 2(A^2+B^2)$ to write this as,

$$||f-(1-s)\Pi[\Gamma_{1,0}]f-s\Pi[\Gamma_{1,1}]f||^2_{\mathrm{L}^2(\Omega)} \leqslant$$
$$2\int_{x=0}^{h_1}\int_{y=0}^{h_2}\left[\frac{h_1}{h_2}\,||\partial f/\partial x_1||^2_{\mathrm{L}^2(\Omega)} + h_2\int_{b=0}^{h_2}\left[\frac{\partial f}{\partial b}(x,b)\right]^2 db\right] dxdy \ .$$

This reduces to,

$$||f-(1-s)\Pi[\Gamma_{1,0}]f-s\Pi[\Gamma_{1,1}]f||^2_{\mathrm{L}^2(\Omega)} \leqslant$$
$$2h_1^2\,||\partial f/\partial x_1||^2_{\mathrm{L}^2(\Omega)} + 2h_2^2\,||\partial f/\partial x_2||^2_{\mathrm{L}^2(\Omega)} \ .$$

□

4.6 The quadrature rule.

In the previous section we left open the choice of the quadrature rule for the computation of α_h. In this section we select a quadrature rule and we check whether it meets condition 1 from section 4.5. If it is to satisfy the normal addition rules for integrals, the quadrature rule must respect the local support and vector character of the basis vector functions given in (24), so it must satisfy,

$$\alpha_h(\boldsymbol{\eta}_{i,j-\frac12}, Q_h\boldsymbol{\eta}_{k-\frac12,l}) \equiv 0 \ , \ j\neq l \Rightarrow \alpha_h(\boldsymbol{\eta}_{i,j-\frac12}, Q_h\boldsymbol{\eta}_{k,l-\frac12}) \equiv 0 \ ,$$
$$i\neq k \Rightarrow \alpha_h(\boldsymbol{\eta}_{i-\frac12,j}, Q_h\boldsymbol{\eta}_{k-\frac12,l}) \equiv 0 \ .$$

On $\hat{\Gamma}_r$ we use a one-point rule with $\mathbf{x}_r$ as nodal point. We choose the weight at the node in such a way, that

$$\alpha_{h,1}(\mathbf{e}_{x,r}, Q_h\boldsymbol{\eta}_r) = \alpha_{SG}(\mathbf{e}_{x,r}, Q_h\boldsymbol{\eta}_r) \ , \qquad (4.35a)$$

i.e. the rule is exact for all $\boldsymbol{\sigma}$ that have a constant component along $\hat{\Gamma}_r$.

The discretisation obtained in this way can be derived in several other ways, see e.g. the papers by Bank et al. [3, 4], the discretisation is closely related to the method given by MacNeal [15]. If we use the quadrature rule given above, we find the following formula for $\alpha_{h,1}(\boldsymbol{\eta}_r, Q_h\boldsymbol{\eta}_s)$,

$$\alpha_{h,1}(\boldsymbol{\eta}_r, Q_h\boldsymbol{\eta}_s) = \delta_{rs}\lambda(\Gamma_r)\int_{\hat{\Gamma}_s} \frac{\exp(\psi)}{a}\, d\lambda\ , \tag{4.35b}$$

this shows that the corresponding matrix is a diagonal matrix. It is clear that this rule corresponds to the use of a Scharfetter-Gummel scheme for each of the two directions $\mathbf{e}_1$ and $\mathbf{e}_2$ separately.

Lemma 4.4.
If ψ is piecewise linear, then

$$\|\boldsymbol{\sigma}_h\|^2_{L^2(\Omega)} \min_{r\in E} P[\hat{\Gamma}_r](\exp(\psi)/a) \leqslant \alpha_{h,1}(\boldsymbol{\sigma}_h,\boldsymbol{\sigma}_h) \leqslant$$
$$3\,\|\boldsymbol{\sigma}_h\|^2_{L^2(\Omega)} \max_{r\in E} P[\hat{\Gamma}_r](\exp(\psi)/a)\ .$$

Proof.
This follows immediately from (35b), the definition of $(.,.)_h$ and lemma 1.

□

So, formally $\alpha_{h,1}$ satisfies condition 1 from section 4.5 and we can apply theorem 1 from section 4.5 to the discrete scheme based on this quadrature rule. We use the word "formally" to indicate that the constant A in condition 1 may need to be very large. This is due to the appearance of the exponential weighting function in the nodal weight for the quadrature rule. In general we shall use the words "formal" and "formally" to indicate that certain statements hold, but only for very small h.

4.7 Consistency.

As discussed in section 4.3.4, we use a projection onto the trial space to split the difference between the solution of (1) and its discrete approximation into an interpolation error and a discretisation error as follows,

$$\|\boldsymbol{\sigma}-\boldsymbol{\sigma}_h\|_{H(\mathrm{div};\Omega)} \leqslant \|\boldsymbol{\sigma}-\Pi_h\boldsymbol{\sigma}\|_{H(\mathrm{div};\Omega)} + 3\,\|\Pi_h\boldsymbol{\sigma}-\boldsymbol{\sigma}_h\|_{V_h}\ ,$$
$$\|U-\mathscr{E}u_h\|_{L^2(\Omega)} \leqslant \|U-\bar{P}_hU\|_{L^2(\Omega)} + \|\bar{P}_hU-\mathscr{E}u_h\|_{W_h}\ .$$

The interpolation error can be estimated by standard approximation theory. Here we study the discretisation error.

4.7.1. Effects of piecewise bilinear interpolation for ψ.

At the start of section 4.3 we assumed that ψ was piecewise bilinear. If this does not hold then we can estimate the error caused by approximation of ψ with the aid of the following lemma.

Lemma 4.5.
If $\psi \in C^2([0,h])$ and we replace ψ in

$$\int_0^h \exp(\psi)\, d\lambda \ ,$$

by ψ_I, defined as

$$\psi_I(x) = \frac{h-x}{h}\psi(0) + \frac{x}{h}\psi(h) \quad \forall \;\; x \in [0,h] \ ,$$

then

$$\left| \int_0^h \exp(\psi)\, d\lambda - \int_0^h \exp(\psi_I)\, d\lambda \right| \leqslant$$

$$\left| \int_0^h \exp(\psi_I)\, d\lambda \right| \left(\exp(h^2 \,\|\, d^2\psi / dx^2 \,\|_{\mathrm{L}^\infty([0,h])}) - 1 \right] .$$

Proof.
We start by giving an estimate for

$$\|\,\psi - \psi_I\,\|_{\mathrm{L}^\infty([0,h])} \ .$$

If $\psi \in C^2([0,h])$, then

$$\psi(x) - \psi_I(x) = \psi(0) + \int_{y=0}^{x} \frac{d\psi}{dx}(y)dy \ - \ \psi(0) \ - \ \frac{x}{h} \int_{z=0}^{h} \frac{d\psi}{dx}(z)dz =$$

$$\frac{1}{h} \int_{z=0}^{h} \int_{y=0}^{x} \frac{d\psi}{dx}(y) \ - \ \frac{d\psi}{dx}(z)dydz \ = \ \frac{1}{h} \int_{z=0}^{h} \int_{y=0}^{x} \int_{w=z}^{y} \frac{d^2\psi}{dx^2}(w)dwdydz \ =$$

$$\frac{1}{h} \int_{z=0}^{h} \int_{y=0}^{x} \int_{w=0}^{y} \frac{d^2\psi}{dx^2}(w)dwdydz \ - \ \frac{1}{h} \int_{z=0}^{h} \int_{y=0}^{x} \int_{w=0}^{z} \frac{d^2\psi}{dx^2}(w)dwdydz$$

$$= \int_{y=0}^{x} (x-y) \, \frac{d^2\psi}{dx^2}(y)dy \ - \ \frac{x}{h} \int_{z=0}^{h} (h-z)\frac{d^2\psi}{dx^2}(z)dz \ .$$

This implies that

$$\|\,\psi - \psi_I\,\|_{\mathrm{L}^\infty([0,h])} \leqslant h^2 \,\|\, d^2\psi / dx^2 \,\|_{\mathrm{L}^\infty([0,h])} \ .$$

We combine integrals and reorder terms to find,

$$\left| \int_0^h \exp(\psi)\, d\lambda - \int_0^h \exp(\psi_I)\, d\lambda \right| = \left| \int_0^h \exp(\psi_I)\, [\exp(\psi - \psi_I) - 1]\, d\lambda \right| .$$

We use our estimate for

$$\| \psi - \psi_I \|_{L^\infty([0,h])} \cdot$$

and move the resulting constant term out of the integral to find the desired estimate.

□

This implies that the approximation of ψ by its bilinear interpolator causes a relative error in the coefficients of our quadrature rules that is formally of order $\mathcal{O}(h^2)$, where h is the maximum edge length and the error constant is dependent on $\partial^2\psi / \partial x_1^2$ and $\partial^2\psi / \partial x_2^2$. As we shall see in section 4.7.2, this is comparable in order to the local error resulting from the use of the quadrature rule $\alpha_{h,1}$.

4.7.2. The discretisation error.

In section 4.5, theorem 1, we gave an expression for the norm of the solution of a saddle-point system in terms of the right hand side. If we insert of the difference between the projection $(\Pi_h \boldsymbol{\sigma}, \overline{\mathrm{P}}_h u)$ of the solution of (1) and the solution $(\boldsymbol{\sigma}_h, u_h)$ of (33) into the saddle-point problem corresponding to the discrete system, the norm of the right hand side is given by

$$\sup_{0 \neq \boldsymbol{\tau}_h \in V_h} \frac{| \alpha_h(\Pi_h \boldsymbol{\sigma}, Q_h \boldsymbol{\tau}_h) - \alpha_{SG}(\boldsymbol{\sigma}, Q_h \boldsymbol{\tau}_h) |}{\| \boldsymbol{\tau}_h \|_{V_h}} , \tag{4.36}$$

for (33a) and 0 for the (33b). In this section, we consider this expression for $\alpha_h = \alpha_{h,1}$ and $a \equiv 1$.

We consider the above expression for $\boldsymbol{\tau}_h = \boldsymbol{\eta}_r$ and express it in the local coordinates and local functions defined in section 4.3.2. As we consider the expression for one fixed edge $\hat{\Gamma}_r$, we may omit the subscript r. The two bilinear forms of interest take the following form,

$$\alpha_{SG}(\boldsymbol{\sigma}, Q_h \boldsymbol{\eta}_r) = h_y \int_{x=-\frac{1}{2}h_L}^{\frac{1}{2}h_R} \sigma(x,0) \exp(\beta x + \gamma)\, dx , \tag{4.37}$$

$$\alpha_{h,1}(\boldsymbol{\sigma}, Q_h \boldsymbol{\eta}_r) = \left[\int_{y=-\frac{1}{2}h_y}^{\frac{1}{2}h_y} \sigma(0,y)\, dy \right] \int_{x=-\frac{1}{2}h_L}^{\frac{1}{2}h_R} \exp(\beta x + \gamma)\, dx . \tag{4.38}$$

We assume, that σ_1, σ_2 are elements of $C^4(\overline{\Omega})$. In the following lemma we give a formula for the difference between (37) and (38) for an arbitrary vector-valued function $\boldsymbol{\sigma} \in C^4(\overline{\Omega})^2$. To simplify notation, we introduce the moments of $\exp(\psi)$ on all dual mesh edges,

$$\tilde{L}_{r,n} = \int_{x=-\frac{1}{2}h_{r,L}}^{\frac{1}{2}h_{r,R}} x^n \exp(\psi_r)\, dx ,$$

we see immediately, that

$$\alpha_{h,1}(\boldsymbol{\eta}_r, Q_h\boldsymbol{\eta}_r) = \lambda(\Gamma_r)\tilde{L}_{r,0} \ .$$

We also introduce scaled versions of these moments,

$$L_{r,n} = \frac{\tilde{L}_{r,n}}{\tilde{L}_{r,0}} = \frac{\displaystyle\int_{x=-\frac{1}{2}h_{r,L}}^{\frac{1}{2}h_{r,R}} x^n \exp(\beta_r x)\, dx}{\displaystyle\int_{x=-\frac{1}{2}h_{r,L}}^{\frac{1}{2}h_{r,R}} \exp(\beta_r x)\, dx} \ .$$

Lemma 4.6.
We consider a vector-valued function $\boldsymbol{\sigma}$ in local coordinates around $\mathbf{x}_r$. Let $H = \max(\lambda(\Gamma_r), \lambda(\hat{\Gamma}_r))$. If we assume that $\sigma_r \in C^4(\mathbb{R}^2)$, then we can expand σ_r in a Taylor series around the origin of the local coordinate system, as r is fixed we omit the subscript r on σ, x and y. We write σ_x, σ_y for the partial derivatives in the local x and y directions.

$$\alpha_{SG}(\boldsymbol{\sigma}, Q_h\boldsymbol{\eta}_r) - \alpha_{h,1}(\boldsymbol{\sigma}, Q_h\boldsymbol{\eta}_r) = \tag{4.39}$$

$$\alpha_{h,1}(\boldsymbol{\eta}_r, Q_h\boldsymbol{\eta}_r)\Big[\sigma_x(0,0)L_{r,1} + \tfrac{1}{2}\sigma_{xx}(0,0)L_{r,2} -$$

$$\tfrac{1}{2}f_{yy}(0,0)\frac{{h_y}^2}{12} + \frac{1}{6}\sigma_{xxx}(0,0)L_{r,3} + \frac{1}{24}\sigma_{xxxx}(\mu_x,0)L_{r,4} - \frac{1}{24}\sigma_{yyyy}(0,\mu_y)\frac{{h_y}^4}{80}\Big] \ ,$$

with $\mu_x \in [-\tfrac{1}{2}h_L, \tfrac{1}{2}h_R]$, $\mu_y \in [-\tfrac{1}{2}h_y, \tfrac{1}{2}h_y]$.
Proof.
to verify this, we subtract (38) from (37), expand all occurrences of σ in Taylor series around the local origin and carry out all integrations over y. After integration, we are left with the above expression for $\alpha_{SG} - \alpha_{h,1}$,

□

The form of this expression and the earlier expression for the error due to bilinear approximation of ψ suggest that the formal order behaviour is best studied by dividing the part of the error corresponding to a given dual edge $\hat{\Gamma}_r$ by $\alpha_{h,1}(\boldsymbol{\eta}_r, Q_h\boldsymbol{\eta}_r)$.

The use of a one point rule for α_h can affect accuracy. We specify three cases where

$$\frac{(\alpha_{SG} - \alpha_{h,1})(\boldsymbol{\sigma}, Q_h\boldsymbol{\eta}_r)}{\alpha_{h,1}(\boldsymbol{\eta}_r, Q_h\boldsymbol{\eta}_r)} \ ,$$

may be $\mathcal{O}(h)$ in stead of $\mathcal{O}(h^2)$.

Case I.

If $\Gamma_r \subset \Gamma_1$, i.e. we are dealing with an edge on the Dirichlet boundary, then there are $\mu \in (0,1)$, $k_r \in [-K,K]$, where K depends only on the $L^\infty(\Omega)$ norm of derivatives of $\boldsymbol{\sigma}$, such that

$$\frac{|\alpha_{SG}(\boldsymbol{\sigma}, Q_h\boldsymbol{\eta}_r) - \alpha_{h,1}(\Pi_h\boldsymbol{\sigma}, Q_h\boldsymbol{\eta}_r)|}{\alpha_{h,1}(\boldsymbol{\eta}_r, Q_h\boldsymbol{\eta}_r)} = \mu H\sigma_x(0,0) + k_r H^2 ,$$

this contains a first order error term in the right hand side.

Case II.

If a vertical edge Γ_r lies in the interior on Ω and the width of the cell on the left side of he edge differs from that of the cell on the right side by more than a factor of order $\mathcal{O}(H^2)$, then there are $\mu \in (0,1)$, $k_r \in [-K,K]$, where K depends only on the $L^\infty(\Omega)$ norm of derivatives of $\boldsymbol{\sigma}$, such that

$$\frac{|\alpha_{SG}(\boldsymbol{\sigma}, Q_h\boldsymbol{\eta}_r) - \alpha_{h,1}(\Pi_h\boldsymbol{\sigma}, Q_h\boldsymbol{\eta}_r)|}{\alpha_{h,1}(\boldsymbol{\eta}_r, Q_h\boldsymbol{\eta}_r)} = \mu H\sigma_x(0,0) + k_r H^2 ,$$

because the first order error terms for these cells do not cancel even when $\beta=0$.

Case III.

Lastly, if an interval vertical edge Γ_r lies in the interior on Ω and the jump in ψ over the edge is larger than 2, then there is a $k_r \in [-K,K]$, where K depends only on the $L^\infty(\Omega)$ norm of derivatives of $\boldsymbol{\sigma}$, such that

$$\frac{|\alpha_{SG}(\boldsymbol{\sigma}, Q_h\boldsymbol{\eta}_r) - \alpha_{h,1}(\Pi_h\boldsymbol{\sigma}, Q_h\boldsymbol{\eta}_r)|}{\alpha_{h,1}(\boldsymbol{\eta}_r, Q_h\boldsymbol{\eta}_r)} = C\sigma_x(0,0) + k_r H^2 ,$$

because of the asymmetry of $\exp(\psi)$. The coefficient C of $\sigma_x(0,0)$ is given by

$$C = L_{r,1} ,$$

this is equivalent to

$$L_{r,1} = \tfrac{1}{2}\left[\tfrac{1}{2}(h_R - h_L) + \tfrac{1}{2}(h_R + h_L)G\left(\beta\frac{h_R + h_L}{4}\right)\right] ,$$

with

$$G(z) = \frac{z\cosh z - \sinh z}{z \sinh z} .$$

We see that,

$$\frac{dG}{dz}(z) = \frac{(\sinh z)^2 - z^2}{z^2(\sinh z)^2} .$$

The behaviour of G is as follows,

$$G(z) = -G(-z) ,$$

$$\lim_{z\to\infty} G(z) = 1 ,$$

$$0 \leqslant \frac{dG}{dz}(z) < 1 ,$$

so $-1 \leqslant G(z) \leqslant 1$. If we assume that $h_R = h_L$, then

$$L_{r,1} = \tfrac{1}{2}(h_R+h_L)G\left(\beta\frac{h_R+h_L}{4}\right) .$$

So the order behaviour of $L_{r,1}$ is determined by the order behaviour of G. Assume that $h = h_L + h_R < 1$. The order behaviour of G is as follows. If $\beta h < 2$, then

$$G(\tfrac{1}{2}\beta h) \leqslant \frac{\tfrac{1}{2}\beta h(1+\tfrac{1}{2}(\tfrac{1}{2}\beta h)^2\cosh(1)) - \tfrac{1}{2}\beta h}{(\tfrac{1}{2}\beta h)^2} = \frac{\beta h\cosh(1)}{2} < \frac{\beta\cosh(1)}{2}h ,$$

so $G(\tfrac{1}{2}\beta h)$ is $\mathcal{O}(h)$. On the other hand, as long as $\beta h > 20$,

$$G(\tfrac{1}{2}\beta h) \geqslant G(10) = \frac{10\cosh(10) - \sinh(10)}{10\sinh(10)} > \frac{9}{10} ,$$

so for all meshes with $h > 20/\beta$, we have $G(\tfrac{1}{2}\beta h)$ is $\mathcal{O}(1)$. We see that $L_{r,1}$ is at worst $\mathcal{O}(h)$ and at best $\mathcal{O}(h^2)$. If $h < |2/\beta|$ then $L_{r,1}$ is $\mathcal{O}(H^2)$. If $h > |20/\beta|$ then is $\mathcal{O}(H)$.

4.8 An a-posteriori error estimator.

In this section we study an a-posteriori error estimate for the discretisation based on $\alpha_h = \alpha_{h,1}$. We calculate a correction to an initial solution and use this to improve the order of approximation, this method is related to the deferred correction scheme as described by Fox and Mayers in chapter 6 of [17].

4.8.1. A derivation of a deferred correction scheme.

In this section we give a deferred correction scheme. The discussion takes into account formal order only, i.e. it assumes that h is "small enough". In equation (33) we take $\alpha_{h,1}$, given in (35b), as our α_h. If we insert $(\Pi_h\boldsymbol{\sigma} - \boldsymbol{\sigma}_h, \bar{\mathrm{P}}_h u - u_h)$ in (33) to determine an expression for the right hand side, then we find

$$\alpha_{h,1}(\Pi_h\boldsymbol{\sigma} - \boldsymbol{\sigma}_h, Q_h\boldsymbol{\tau}_h) - (\operatorname{div}\boldsymbol{\tau}_h, \mathscr{E}(\bar{\mathrm{P}}_h u - u_h)) = \qquad (4.40a)$$

$$\alpha_{h,1}(\Pi_h\boldsymbol{\sigma}, Q_h\boldsymbol{\tau}_h) - \alpha_{SG}(\boldsymbol{\sigma}, Q_h\boldsymbol{\tau}_h) \quad \forall\ \boldsymbol{\tau}_h \in V_h ,$$

$$(\operatorname{div}(\Pi_h\boldsymbol{\sigma} - \boldsymbol{\sigma}_h), t_h) = 0 \quad \forall\ t_h \in W_h . \qquad (4.40b)$$

Our approach is the following. We assume that $\sigma_1, \sigma_2 \in C^4(\bar{\Omega})$ and we assume that we have Dirichlet boundary conditions on the entire boundary of our domain. We see from (39) that we can approximate the right hand side of (40a) by an expression in the partial derivatives of $\boldsymbol{\sigma}$. If we can justify the use of finite difference approximations based on $\boldsymbol{\sigma}_h$ for these derivatives, then we can solve (33) with a adjusted right hand side and obtain a better solution. First we show that we can approximate partial derivatives of $\boldsymbol{\sigma}$ of first or second order in a given direction by finite differences of $\Pi_h\boldsymbol{\sigma}$. Next we show that we can use finite differences of $\boldsymbol{\sigma}_h$ to approximate the finite differences of $\Pi_h\boldsymbol{\sigma}$. We introduce the following special notation.

$$\partial_{h,\kappa}f(\mathbf{x}) = \frac{f(\mathbf{x}+h\mathbf{e}_\kappa)-f(\mathbf{x})}{h} ,$$

$$\partial^2_{h,\kappa}f(\mathbf{x}) = \frac{f(\mathbf{x}+h\mathbf{e}_\kappa)-2f(\mathbf{x})+f(\mathbf{x}-h\mathbf{e}_\kappa)}{h^2} .$$

Lemma 4.7.
If $f \in C^3([0,1]\times[0,1])$, $h \in (0,1/4)$,

$$\Gamma(\mathbf{x}) = \{ \ (x,y) \mid x_1=x \ , y \in [x_2-h/2,x_2+h/2] \ \} ,$$

and $\mathbf{x} \in [h,1-h]\times[h,1-h]$, then

$$| \, \partial_{h,\kappa}f(\mathbf{x}) - \frac{\partial f}{\partial x_\kappa}(\mathbf{x}+\tfrac{1}{2}h\mathbf{e}_\kappa) | = \mathcal{O}(h^2) , \tag{4.41a}$$

$$| \, \partial^2_{h,\kappa}f(\mathbf{x}) - \frac{\partial^2 f}{\partial x_\kappa^2}(\mathbf{x}+\tfrac{1}{2}h\mathbf{e}_\kappa) | = \mathcal{O}(h^2) , \tag{4.41b}$$

$$| \, \partial_{h,1}(f-P[\Gamma(\mathbf{x})](f))(\mathbf{x}) | = \mathcal{O}(h^2) , \tag{4.41c}$$

$$| \, \partial_{h,2}(f-P[\Gamma(\mathbf{x})](f))(\mathbf{x}) | = \mathcal{O}(h^2) , \tag{4.41d}$$

$$| \, \partial^2_{h,1}(f-P[\Gamma(\mathbf{x})](f))(\mathbf{x}) | = \mathcal{O}(h^2) , \tag{4.41e}$$

$$| \, \partial^2_{h,2}(f-P[\Gamma(\mathbf{x})](f))(\mathbf{x}) | = \mathcal{O}(h^2) . \tag{4.41f}$$

Proof.
The above statements are easily verified through the use of Taylor expansions.

□

For a special case we justify the use of finite differences of $\boldsymbol{\sigma}_h$ to approximate the finite differences of $\Pi_h\boldsymbol{\sigma}$. We assume that the mesh is uniform, i.e. $\lambda(\Gamma_{i-\frac{1}{2},j})=h$ and $\lambda(\Gamma_{i,j-\frac{1}{2}})=h$ for a fixed $h \in \mathbb{R}$ for all edges. We also assume that ψ is linear and increasing on the entire domain, i.e.

$$\psi(\mathbf{x}) = \beta_1 x_1 + \beta_2 x_2 + \gamma ,$$

with fixed $\beta_1,\beta_2,\gamma \in \mathbb{R}$ and $\beta_1,\beta_2 > 0$. We introduce two vectors $\boldsymbol{R}_h,\boldsymbol{S}_h$ in V_h. The bilinear form $\alpha_{h,1}$ acting on the sum of these vectors generates the right hand side of (40a) up to third order. We define

$$L_n = \frac{\displaystyle\int_{x=-\frac{1}{2}h}^{\frac{1}{2}h} x^n \exp(x\beta_\kappa)\,dx}{\displaystyle\int_{x=-\frac{1}{2}h}^{\frac{1}{2}h} \exp(x\beta_\kappa)\,dx} ,$$

this is equal to $L_{r,n}$ if Γ_r does not lie on the Dirichlet boundary Γ_1. We define the vectors by giving their value for each edge $\Gamma_r, r \in E$, we express this value in terms of local coordinates,

$$R_{h,r} = L_1\sigma_{r,x}(0,0) + \tfrac{1}{2}L_2\sigma_{r,xx}(0,0) - \frac{h^2}{24}\sigma_{r,yy}(0,0)\,. \tag{4.42}$$

$$S_{h,r} = \left[(L_{r,1}-L_1)\sigma_{r,x}(0,0) + (L_{r,2}-L_2)\tfrac{1}{2}\sigma_{r,xx}(0,0)\right], \tag{4.43}$$

note that $\boldsymbol{S}_h$ is non-zero only on the Dirichlet part Γ_1 of the boundary. If we compare (42) and (43) with (39), then we see immediately that,

$$\alpha_{SG}(\boldsymbol{\sigma},Q_h\boldsymbol{\eta}_r) - \alpha_{h,1}(\boldsymbol{\sigma},Q_h\boldsymbol{\eta}_r) = \tag{4.44a}$$

$$\alpha_{h,1}(\boldsymbol{R}_h + \boldsymbol{S}_h,Q_h\boldsymbol{\eta}_r) + \mathcal{O}\left[h^3\alpha_{h,1}(\boldsymbol{\eta}_r,Q_h\boldsymbol{\eta}_r)\right] \quad \forall\ r \in E\,.$$

We wish to approximate $\alpha_{h,1}(\boldsymbol{R}_h+\boldsymbol{S}_h,Q_h\boldsymbol{\eta}_r)$ by $\alpha_{h,1}(\Pi_h\boldsymbol{R},Q_h\boldsymbol{\eta}_r)$ $+ < \mathcal{E}_{\partial\Omega}S,\boldsymbol{\eta}_r\cdot\mathbf{n}_{\partial\Omega} >$, where $\boldsymbol{R}$ and S are continuous functions on Ω and the Dirichlet part of the boundary (Γ_1) respectively, we define,

$$R_\kappa(\mathbf{x}) = L_1\frac{\partial\boldsymbol{\sigma}\cdot\mathbf{e}_\kappa}{\partial x_\kappa}(\mathbf{x}) + \tfrac{1}{2}L_2\frac{\partial^2\boldsymbol{\sigma}\cdot\mathbf{e}_\kappa}{\partial x_\kappa^2}(\mathbf{x}) - \frac{h^2}{24}\frac{\partial^2\boldsymbol{\sigma}\cdot\mathbf{e}_\kappa}{\partial x_{3-\kappa}^2}(\mathbf{x})\,,$$

$$S(\mathbf{x}_r)|_{\Gamma_1} =$$

$$\frac{1}{h}\exp(-\psi_r(0))\alpha_{h,1}(\boldsymbol{\eta}_r,Q_h\boldsymbol{\eta}_r)\left[(L_{r,1}-L_1)\frac{\partial\boldsymbol{\sigma}\cdot\mathbf{e}_\kappa}{\partial x_\kappa}(\mathbf{x}_r) + (L_{r,2}-L_2)\tfrac{1}{2}\frac{\partial^2\boldsymbol{\sigma}\cdot\mathbf{e}_\kappa}{\partial x_\kappa^2}(\mathbf{x}_r)\right]$$

$$\forall\ \mathbf{x}_r \in \Gamma_1\,.$$

On each straight part of the Dirichlet boundary, we can extend S to a C^1 function on that part of the Dirichlet boundary by replacing $\mathbf{x}_r$ by $\mathbf{x}$. We see immediately that,

$$\alpha_{h,1}(\Pi_h\boldsymbol{R}-\boldsymbol{R}_h,Q_h\boldsymbol{\eta}_r) = \mathcal{O}\left[h^2(L_1 + L_2 + L_3)\alpha_{h,1}(\boldsymbol{\eta}_r,Q_h\boldsymbol{\eta}_r)\right], \tag{4.44b}$$

and

$$< \mathcal{E}_{\partial\Omega}S,\boldsymbol{\eta}_r\cdot\mathbf{n}_{\Gamma_1} > \ = \alpha_{h,1}(\boldsymbol{S}_h,Q_h\boldsymbol{\eta}_r)\,, \tag{4.44c}$$

for all r such that Γ_r is a part of the Dirichlet boundary. We see that, if problem (1) is solvable for all ($f{=}F, g{=}G$), then , according to (33), the solution of (1) ($\boldsymbol{\rho},v$) for $F=\operatorname{div}\boldsymbol{R}$ on Ω, $G=S$ on Γ_1 will satisfy the equations,

$$\alpha_{SG}(\boldsymbol{\rho},Q_h\boldsymbol{\eta}_r) - b(\boldsymbol{\eta}_r,\mathcal{E}\overline{\mathrm{P}}_h v) = - < \mathcal{E}_{\partial\Omega}S,\boldsymbol{\eta}_r\cdot\mathbf{n}_{\Gamma_1} > \quad \forall\ r \in E\,, \tag{4.45a}$$

$$b(\boldsymbol{\rho},t_h) = (\operatorname{div}\boldsymbol{R},t_h) \quad \forall\ t_h \in W_h\,. \tag{4.45b}$$

If we subtract $\boldsymbol{R}$ from $\boldsymbol{\rho}$ in these equations, then we find,

$$\alpha_{SG}(\boldsymbol{\rho}-\boldsymbol{R},Q_h\boldsymbol{\eta}_r) - b(\boldsymbol{\eta}_r,\mathcal{E}\overline{\mathrm{P}}_h v) = \tag{4.46a}$$

$$-\alpha_{SG}(\boldsymbol{R},Q_h\boldsymbol{\eta}_r) - \ < \mathcal{E}_{\partial\Omega}S,\boldsymbol{\eta}_r\cdot\mathbf{n}_{\Gamma_1} > \quad \forall\ r \in E\,,$$

$$b(\boldsymbol{\rho}-\boldsymbol{R},t_h) = 0 \quad \forall \;\; t_h \in W_h \; . \tag{4.46b}$$

We can write (46a) as follows,

$$\alpha_{h,1}(\Pi_h(\boldsymbol{\rho}-\boldsymbol{R}),Q_h\boldsymbol{\eta}_r) - b(\boldsymbol{\eta}_r,\mathscr{E}\overline{\mathrm{P}}_h v) = \tag{4.47a}$$

$$\alpha_{h,1}(\Pi_h(\boldsymbol{\rho}-\boldsymbol{R}),Q_h\boldsymbol{\eta}_r) - \alpha_{SG}(\boldsymbol{\rho}-\boldsymbol{R},Q_h\boldsymbol{\eta}_r)-\alpha_{SG}(\boldsymbol{R},Q_h\boldsymbol{\eta}_r) -$$

$$< \mathscr{E}_{\partial\Omega}\boldsymbol{S}\cdot\mathbf{n}_{\Gamma_1},\boldsymbol{\eta}_r\cdot\mathbf{n}_{\Gamma_1} > \quad \forall \;\; r \in E \; ,$$

for $r \in E$, this equation can be written as follows,

$$\alpha_{h,1}(\Pi_h(\boldsymbol{\rho}-\boldsymbol{R}),Q_h\boldsymbol{\eta}_r) - b(\boldsymbol{\eta}_r,\mathscr{E}\overline{\mathrm{P}}_h v) = \tag{4.48a}$$

$$\alpha_{h,1}(\Pi_h\boldsymbol{\rho},Q_h\boldsymbol{\eta}_r) - \alpha_{SG}(\boldsymbol{\rho},Q_h\boldsymbol{\eta}_r) - \alpha_{h,1}(\Pi_h\boldsymbol{R}-\boldsymbol{R}_h,Q_h\boldsymbol{\eta}_r) - \alpha_{h,1}(\boldsymbol{R}_h+\boldsymbol{S}_h,Q_h\boldsymbol{\eta}_r) \; .$$

According to 47a, if we subtract 46 from 40 and we use (44,a,b,c) then we get,

$$\alpha_{h,1}(\Pi_h\boldsymbol{\sigma}-\boldsymbol{\sigma}_h-\Pi_h(\boldsymbol{R}-\boldsymbol{\rho}),Q_h\boldsymbol{\eta}_r) - (\operatorname{div}\boldsymbol{\eta}_r,\mathscr{E}(\overline{\mathrm{P}}_h u-u_h-\overline{\mathrm{P}}_h v)) = \tag{4.49a}$$

$$\alpha_{SG}(\boldsymbol{\rho},Q_h\boldsymbol{\eta}_r) - \alpha_{h,1}(\Pi_h\boldsymbol{\rho},Q_h\boldsymbol{\eta}_r) + \mathscr{O}\left[(L_1+L_2+L_3)h^2\alpha_{h,1}(\boldsymbol{\eta}_r,Q_h\boldsymbol{\eta}_r)\right] +$$

$$\mathscr{O}\left[h^3\alpha_{h,1}(\boldsymbol{\eta}_r,Q_h\boldsymbol{\eta}_r)\right] \quad \forall \;\; r \in E \; ,$$

$$(\operatorname{div}(\Pi_h\boldsymbol{\sigma}-\boldsymbol{\sigma}_h) - \Pi_h(\boldsymbol{R}-\boldsymbol{\rho}),t_h) = 0 \quad \forall \;\; t_h \in W_h \; . \tag{4.49b}$$

If we assume that L_1 is $\mathscr{O}(h)$ but not $\mathscr{O}(h^2)$, - this holds if e.g. $\beta_1,\beta_2 > 20/h$ - and that problem (1) satisfies the following regularity condition for all $f \in \mathrm{L}^2(\Omega), g \in \mathrm{H}^{3/2}(\partial\Omega)$,

$$\|u\|_{\mathrm{L}^2(\Omega)} + \|\sigma_1\|_{\mathrm{H}^1(\Omega)} + \|\sigma_2\|_{\mathrm{H}^1(\Omega)} \leqslant C(\|f\|_{\mathrm{L}^2(\Omega)} + \|g\|_{\mathrm{H}^{3/2}(\partial\Omega)}) \; ,$$

then we can derive an estimate for

$$|\,\alpha_{SG}(\boldsymbol{\rho},Q_h\boldsymbol{\eta}_r) - \alpha_{h,1}(\Pi_h\boldsymbol{\rho},Q_h\boldsymbol{\eta}_r)\,| \; .$$

To prove that the problem has this regularity, we use theorem 5.2.2. by P. Grisvard [16], for our problem, the theorem states that, if we have Dirichlet boundary conditions everywhere and (1) has a unique solution, then operator (1) - with $\Gamma_1 = \partial\Omega$ is a bijective continuous mapping from $\mathrm{H}^2(\Omega)$ to $\mathrm{L}^2(\Omega) \times \mathrm{H}^{3/2}(\partial\Omega)$. From the equivalence of (1) and (2) and the ellipticity of (2), we see that - for Dirichlet boundary conditions - equation (1) has a unique solution. Now the above mentioned theorem states that the operator is bounded, so bounded inverse theorem (Schechter [18], theorem 4.1) implies that the operator has a bounded inverse. This in turn implies that the above regularity condition holds.

According to our assumption that $\sigma_1,\sigma_2 \in C^4(\overline{\Omega})$ and the fact that L_1 is $\mathscr{O}(h)$, we have $\boldsymbol{R} = h\boldsymbol{F}$ with $F_1,F_2 \in \mathrm{H}^4(\overline{\Omega})$, $\operatorname{div}\boldsymbol{R} = hf$ with $f \in \mathrm{H}^3(\Omega)$ and $S = h^2 g$ with $g \in \mathrm{H}^3(\partial\Omega)$. According to lemma 6, this implies that,

$$|\,\alpha_{SG}(\boldsymbol{\rho},Q_h\boldsymbol{\eta}_r) - \alpha_{h,1}(\Pi_h\boldsymbol{\rho},Q_h\boldsymbol{\eta}_r)\,| \leqslant$$

$$hCL_1(\|f\|_{\mathrm{L}^2(\Omega)} + \|g\|_{\mathrm{H}^{3/2}(\partial\Omega)})\,|\,\alpha_{h,1}(\boldsymbol{\eta}_r,Q_h\boldsymbol{\eta}_r)\,| = \mathscr{O}(h^2\alpha_{h,1}(\boldsymbol{\eta}_r,Q_h\boldsymbol{\eta}_r)) \quad \forall \;\; r \in E \, ,$$

this implies that $(\boldsymbol{\sigma}_h + \Pi_h(\boldsymbol{R}-\boldsymbol{\rho}), u_h + \overline{\mathrm{P}}_h v)$ considered as an approximation to $(\Pi_h\boldsymbol{\sigma}, \overline{\mathrm{P}}_h u)$ is one order of h more accurate than $(\boldsymbol{\sigma}_h, u_h)$, i.e. it is $\mathcal{O}(h^2)$.

Now assume, that L_1 *is* $\mathcal{O}(h^2)$ - this holds if e.g. $\beta_1, \beta_2 < 2/h$ - and problem (1) satisfies the above regularity condition. Now according to our assumption that $\sigma_1, \sigma_2 \in C^4(\overline{\Omega})$ and the fact that L_1 is $\mathcal{O}(h^2)$, we have $\boldsymbol{R} = h^2\boldsymbol{F}$ with $F_1, F_2 \in \mathrm{H}^4(\overline{\Omega})$, $\operatorname{div}\boldsymbol{R} = h^2 f$ with $f \in \mathrm{H}^3(\Omega)$ and $S = h^2 g$ with $g \in \mathrm{H}^3(\partial\Omega)$. Note that $L_{r,1}$ is $\mathcal{O}(h)$ if Γ_r is a part of the Dirichlet edge. According to lemma 6, away from the Dirichlet edge,

$$\begin{gathered}|\alpha_{SG}(\boldsymbol{\rho}, Q_h\boldsymbol{\eta}_r) - \alpha_{h,1}(\Pi_h\boldsymbol{\rho}, Q_h\boldsymbol{\eta}_r)| \leqslant \\ C(L_1+L_2)h(\|f\|_{\mathrm{L}^2(\Omega)} + \|g\|_{\mathrm{H}^{3/2}(\partial\Omega)})|\alpha_{h,1}(\boldsymbol{\eta}_r, Q_h\boldsymbol{\eta}_r)| = \\ \mathcal{O}(h^3\alpha_{h,1}(\boldsymbol{\eta}_r, Q_h\boldsymbol{\eta}_r)) \quad \forall \ r \in E \ ,\end{gathered}$$

and on the Dirichlet edge,

$$\begin{gathered}|\alpha_{SG}(\boldsymbol{\rho}, Q_h\boldsymbol{\eta}_r) - \alpha_{h,1}(\Pi_h\boldsymbol{\rho}, Q_h\boldsymbol{\eta}_r)| \leqslant \\ CL_{r,1}h(\|f\|_{\mathrm{L}^2(\Omega)} + \|g\|_{\mathrm{H}^{3/2}(\partial\Omega)})|\alpha_{h,1}(\boldsymbol{\eta}_r, Q_h\boldsymbol{\eta}_r)| = \mathcal{O}(h^2\alpha_{h,1}(\boldsymbol{\eta}_r, Q_h\boldsymbol{\eta}_r)) \quad \forall \ r \in E \, .\end{gathered}$$

This implies that

$$\|\Pi_h\boldsymbol{\sigma} - \boldsymbol{\sigma}_h - \Pi_h(\boldsymbol{R}-\boldsymbol{\rho})\|_{\mathbf{L}^2(\Omega)} = \mathcal{O}(h^3) \, ,$$

because expression (36) for the above case is bounded by

$$CN_1N_2h^5 + 2D(N_1+N_2)h^4 \, .$$

We can summarise the two results given above as follows,

$$\|\Pi_h\boldsymbol{\sigma} - \boldsymbol{\sigma}_h - \Pi_h(\boldsymbol{R}-\boldsymbol{\rho})\|_{\mathbf{L}^2(\Omega)} = \mathcal{O}(h^{k+1}) \, ,$$

where k is the order of L_1, i.e. $L_1 = \mathcal{O}(h^k)$.

This in turn implies that,

$$\|\Pi_h\boldsymbol{\sigma} - \boldsymbol{\sigma}_h - \Pi_h(\boldsymbol{R}-\boldsymbol{\rho})\|_{\mathbf{L}^\infty(\Omega)} = \mathcal{O}(h^k) \, ,$$

on at most a $\mathcal{O}(h)$ part of Ω and

$$\|\Pi_h\boldsymbol{\sigma} - \boldsymbol{\sigma}_h - \Pi_h(\boldsymbol{R}-\boldsymbol{\rho})\|_{\mathbf{L}^\infty(\Omega)} = \mathcal{O}(h^{k+1})$$

elsewhere.

We use this to justify the approximation of the partial derivatives $\partial^n/\partial x_\kappa^n$ of $\boldsymbol{\sigma}$ in $\boldsymbol{R}_h$ by divided differences $\partial_{h,\kappa}^n$ of $\boldsymbol{\sigma}_h$. As

$$\Pi_h\boldsymbol{\sigma} - \boldsymbol{\sigma}_h = \Pi_h(\boldsymbol{R}-\boldsymbol{\rho}) + \mathcal{O}(h^{k+1}) \, ,$$

and $\boldsymbol{R}-\boldsymbol{\rho} = h^k\boldsymbol{t}$ with $\boldsymbol{t} \in C^2(\overline{\Omega})^2$, we find,

$$\partial_{h,\kappa}^n\Big[\Pi_h\sigma - \boldsymbol{\sigma}_h\Big](\mathbf{x}_r) = \partial_{h,\kappa}^n\Big[\Pi_h(\boldsymbol{R}-\boldsymbol{\rho})\Big](\mathbf{x}_r) + \mathcal{O}(h^{k-n}) = h^k\frac{\partial^n}{\partial x_\kappa^n}\boldsymbol{t}(\mathbf{x}_r) + \mathcal{O}(h^{k-n}) \, ,$$

for $\kappa = 1,2$ on a $\mathcal{O}(h)$ part of the domain Ω and

$$\partial^n_{h,\kappa}\left[\Pi_h\sigma-\boldsymbol{\sigma}_h\right](\mathbf{x}_r) = \partial^n_{h,\kappa}\left[\Pi_h(\boldsymbol{R}-\boldsymbol{\rho})\right](\mathbf{x}_r) + \mathcal{O}(h^{k+1-n}) =$$
$$h^k\frac{\partial^n}{\partial x^n_\kappa}\boldsymbol{t}(\mathbf{x}_r) + \mathcal{O}(h^{k+1-n}) \,,$$

for $\kappa=1{,}2$ elsewhere. Combined with lemma 7 we find that for $\kappa=1{,}2$ and $n=1{,}2$,

$$|\frac{\partial^n\boldsymbol{\sigma}}{\partial x_\kappa}(\mathbf{x}_r) - \partial_{h,\kappa}\boldsymbol{\sigma}(\mathbf{x}_r)| = \mathcal{O}(h^{k-n}) \,,$$

on a $\mathcal{O}(h)$ part of the domain and

$$|\frac{\partial^n\boldsymbol{\sigma}}{\partial x_\kappa}(\mathbf{x}_r) - \partial_{h,\kappa}\boldsymbol{\sigma}(\mathbf{x}_r)| = \mathcal{O}(h^{k+1-n}) \,,$$

elsewhere.

Let us denote by $\tilde{\boldsymbol{R}}_h$ the approximation of $\boldsymbol{R}_h$ and by $\tilde{\boldsymbol{S}}_h$ the approximation of $\boldsymbol{S}_h$, obtained by substituting $\partial^n_{h,\kappa}\boldsymbol{\sigma}_h$ for $\partial^n/\partial x^n_\kappa$ with $n=1{,}2$. We see that

$$\tilde{\boldsymbol{R}}_h+\tilde{\boldsymbol{S}}_h-\boldsymbol{R}-\boldsymbol{S} = L_1\,\mathcal{O}(h^{k-1}) + L_2\,\mathcal{O}(h^{k-2}) \,,$$

on $\mathcal{O}(h)$ of all cells and

$$\tilde{\boldsymbol{R}}_h+\tilde{\boldsymbol{S}}_h-\boldsymbol{R}-\boldsymbol{S} = L_1\,\mathcal{O}(h^{k}) + L_2\,\mathcal{O}(h^{k-1}) \,,$$

elsewhere. Let $(\tilde{\boldsymbol{\sigma}}_h,\tilde{u}_h)$ be the solution of

$$\alpha_{h,1}(\tilde{\boldsymbol{\sigma}}_h,Q_h\boldsymbol{\tau}_h) - (\operatorname{div}\boldsymbol{\tau}_h,\mathcal{E}(\tilde{u}_h)) = \tag{4.50a}$$
$$\alpha_{h,1}(\tilde{\boldsymbol{R}}_h+\tilde{\boldsymbol{S}}_h,Q_h\boldsymbol{\tau}_h) - \ < \mathcal{E}_{\partial\Omega}g,\boldsymbol{\eta}_r\!\cdot\!\mathbf{n}_{\Gamma_1} > \quad \forall\ \boldsymbol{\tau}_h \in V_h \,,$$
$$(\operatorname{div}(\tilde{\boldsymbol{\sigma}}_h),t_h) = (f,t_h) \quad \forall\ t_h \in W_h \,, \tag{4.50b}$$

then

$$\alpha_{h,1}(\Pi_h\boldsymbol{\sigma} - \tilde{\boldsymbol{\sigma}}_h,Q_h\boldsymbol{\tau}_h) - (\operatorname{div}\boldsymbol{\tau}_h,\mathcal{E}(\bar{\mathrm{P}}_h u - \tilde{u}_h)) = \tag{4.51a}$$
$$\alpha_{h,1}(\boldsymbol{R}_h+\boldsymbol{S}_h,Q_h\boldsymbol{\tau}_h) - \alpha_{h,1}(\tilde{\boldsymbol{R}}_h+\tilde{\boldsymbol{S}}_h,Q_h\boldsymbol{\tau}_h)$$
$$\forall\ \boldsymbol{\tau}_h \in V_h \,,$$
$$(\operatorname{div}(\Pi_h\boldsymbol{\sigma}-\tilde{\boldsymbol{\sigma}}_h),t_h) = 0 \quad \forall\ t_h \in W_h \,, \tag{4.51b}$$

so - in $\mathrm{L}^2(\Omega)$ norm - $(\tilde{\boldsymbol{\sigma}}_h,\tilde{u}_h)$ is formally one order of h closer to $(\Pi_h\boldsymbol{\sigma},\bar{\mathrm{P}}_h u)$ than $(\boldsymbol{\sigma}_h,u_h)$.

We can derive an a-posteriori error estimate by calculating the difference between the discrete solution with and without a tilde. It may be possible to derive a mesh-refinement criterion from $\tilde{\boldsymbol{R}}_h$.

4.8.2. Numerical results.

In this section we show how the deferred correction method works in practice. We consider problem (1) with Dirichlet boundary conditions on the entire boundary,

$$\Gamma_1 = \partial\Omega \ , \ a = 0.01 \ \text{ and } \ \psi = 100(x_1+x_2) \ ,$$

and data derived from a known solution,

$$u = \tanh(8(x_1-x_2)) \ .$$

It follows that,

$$g = u\,|_{\partial\Omega} \ ,$$
$$f = -\frac{\operatorname{div}(\mathbf{grad}\,u + u\,\mathbf{grad}\,\psi)}{100} \ .$$

We show results for the Scharfetter-Gummel version of the discretisation and the results obtained after applying the correction discussed in section 4.8.1 once, twice, thrice or ten times.

We take the unit square for Ω. On a mesh of $n \times n$ cells, with mesh width $h=1/n$, we have $4n-4$ Dirichlet edge cells and a total of n^2 cells. We use the 2-norm as norm for the error vectors,

$$\| v \| = \left(\frac{1}{|I|} \sum_{i\in I} v_i^2 \right) ,$$

where $|I|$ is the number of elements in the index set.
All experiments satisfied the expected symmetry relation

$$\log_2 \| (\Pi_h \boldsymbol{\sigma} - \boldsymbol{\sigma}_h)\cdot \mathbf{e}_1 \| = \log_2 \| (\Pi_h \boldsymbol{\sigma} - \boldsymbol{\sigma}_h)\cdot \mathbf{e}_2 \| \ ,$$

for the accuracy given in the tables.

the $\log_2$ **of the errors for** $\boldsymbol{\alpha}_h = \boldsymbol{\alpha}_{h,1}$.		
meshwidth	$\log_2 \| \mathrm{P_h} u - u_h \|$	$\log_2 \| (\Pi_h \boldsymbol{\sigma} - \boldsymbol{\sigma}_h)\cdot \mathbf{e}_1 \|$
1 / 2	-1.6	-0.9
1 / 4	-1.5	-1.4
1 / 8	-1.9	-1.9
1 / 16	-2.6	-2.6
1 / 32	-3.8	-3.8
1 / 64	-5.5	-5.4
1 / 128	-7.3	-7.3

We see that the large jump in ψ per cell on the coarsest meshes, combined with the large gradient of the solution relative to the coarsest meshes result in convergence slower than $\mathcal{O}(h)$. For a fine mesh, $h < 1/32$, we see that the convergence behaviour tends to $\mathcal{O}(h^2)$. For intermediate meshes intermediate convergence rates are found.

the $\log_2$ of the errors after one correction.		
meshwidth	$\log_2 \| \mathrm{P_h} u - u_h \|$	$\log_2 \| (\Pi_h \boldsymbol{\sigma} - \boldsymbol{\sigma}_h)\cdot \mathbf{e}_1 \|$
1 / 2	-2.2	-1.6
1 / 4	-2.1	-2.0
1 / 8	-2.7	-2.8
1 / 16	-3.9	-3.9
1 / 32	-6.0	-6.0
1 / 64	-9.1	-9.1
1 / 128	-12.9	-12.8

We still see slow convergence rates at the coarsest meshes, probably due to the relative steepness of the solution on that mesh. Convergence speed on the finer meshes is improved by the correction. We see that - as predicted below equation (4.51b) in section 4.8.1 - where the previous table shows first order behaviour between meshes, we now find second order convergence. And where the previous table shows second order behaviour between meshes, we now find third order convergence.

the $\log_2$ of the errors after two corrections		
meshwidth	$\log_2 \| \mathrm{P_h} u - u_h \|$	$\log_2 \| (\Pi_h \boldsymbol{\sigma} - \boldsymbol{\sigma}_h)\cdot \mathbf{e}_1 \|$
1 / 2	-2.6	-1.8
1 / 4	-2.4	-2.4
1 / 8	-3.1	-3.3
1 / 16	-4.7	-4.8
1 / 32	-7.6	-7.6
1 / 64	-11.8	-11.7
1 / 128	-15.4	-15.3

We still see slow convergence at the coarsest meshes. Again we find $k+1$-th order behaviour between meshes where the previous table shows k-th order order behaviour between meshes.

the $\log_2$ of the errors after ten corrections.		
meshwidth	$\log_2 \| \mathrm{P_h} u - u_h \|$	$\log_2 \| (\Pi_h \boldsymbol{\sigma} - \boldsymbol{\sigma}_h)\cdot \mathbf{e}_1 \|$
1 / 2	-2.8	-2.0
1 / 4	-2.8	-2.7
1 / 8	-3.7	-3.9
1 / 16	-5.9	-6.1
1 / 32	-9.2	-9.2
1 / 64	-12.6	-12.6
1 / 128	-15.4	-15.4

After ten iterations no further significant changes occurred. We see that we have third order behaviour from $h = 1/16$ onward.

4.9 Conclusions.

In section 4.4 and 4.6 we have seen that the Scharfetter-Gummel discretisation in two dimensions can be written as a saddle point problem. We can use theorem 3.1 by Nicolaides [5] to show that this discretisation is at least formally stable and consistent. We then showed consistency. In section 4.8 we presented a technique to obtain a local error indicator and we gave numerical results.

The results on a posteriori error estimates can be summarised as follows. We show that it gives an approximation of the error that is an $\mathcal{O}(h^{k+1})$ accurate approximation to the true error, when the true error is $\mathcal{O}(h^k)$. This can also be seen in the numerical results for this method.

We see that the two dimensional Scharfetter-Gummel scheme for the current continuity equation is stable and consistent. Our error analysis in section 4.7 yields the following information on the order of the error. For small enough h, he error is order two only if a cell is not adjacent to the boundary and has a size that differs at most $\mathcal{O}(h^2)$ from its neighbours. If these conditions do not hold the error is of order $\mathcal{O}(h)$. To be certain that the global order of the error is $\mathcal{O}(h^2)$ the change in ψ between cell centres must be smaller than 2. For semiconductors this means that the change in the voltage scaled by the thermal voltage must be smaller than 2. In Section 4.8 it is shown that is possible to calculate a correction to the solution of the Scharfetter-Gummel scheme. From this we can derive an a-posteriori error estimator.

As we mentioned in chapter one, a search of the literature shows that papers on a posteriori error estimates for finite volume or mixed finite element discretisations - other than for fluid dynamics - are rare. There are papers that deal with a posteriori error estimates for the mixed discretisation of the Navier-Stokes equations, see e.g. the paper by Verfürth, [19] but the techniques used there are geared to that type of problem.

References

1. D. L. Scharfetter and H. K. Gummel, "Large-Signal Analysis of a Silicon Read Diode Oscillator," *IEEE Transactions on Electron Devices*, vol. ED-16, no. 1, pp. 64-77, 1969.
2. A. M. Il'in, "Differencing scheme for a differential equation with a small parameter affecting the highest derivative," *Mathematical Notes of the Academy of Sciences of the USSR*, vol. 6, no. 1-2, pp. 596-602, 1969.
3. Wolfgang Fichtner, Donald J. Rose, and Randolph E. Bank, "Numerical methods for semiconductor device simulation," *SIAM journal of scientific and statistical computing*, vol. 4, no. 3, pp. 416-435, 1983.
4. Wolfgang Fichtner, Donald J. Rose, and Randolph E. Bank, "Semiconductor device simulation," *SIAM journal of scientific and statistical computing*, vol. 4, no. 3, pp. 391-415, 1983.

5. R. A. Nicolaides, "Existence, uniqueness and approximation for generalized saddle point problems," *SIAM J. Numer. Anal.*, vol. 19, no. 2, pp. 349-357, 1982.

6. Peter A. Markowich, *The Stationary Semiconductor Device Equations,* Springer-Verlag, Wien New York, 1986.

7. Siegfried Selberherr, *Analysis and simulation of semiconductor devices,* Springer-verlag, Wien New York, 1984.

8. S. J. Polak, C. den Heijer, H. A. Schilders, and P. Markowich, "Semiconductor device modelling from the numerical point of view," *International Journal for Numerical Methods in Engineering*, vol. 24, pp. 763-838, 1987.

9. Walter L. Engl, Heinz K. Dirks, and Bernd Meinerzhagen, "Device Modeling," *Proceedings of the IEEE*, vol. 71, no. 1, pp. 10-33, January 1983.

10. R. E. Bank, W. Fichtner, D. J. Rose, and R. K. Smith, "Algorithms for semiconductor device simulation," in *Large Scale Scientific Computation*, ed. B. Engquist, Progress in Scientific Computing, Birkhäuser, 1987.

11. Randolph E. Bank, Joseph W. Jerome, and Donald J. Rose, "Analytical and numerical aspects of semiconductor device modelling," in *Computing Methods in Applied Sciences and Engineering*, ed. J. L. Lions, vol. V, pp. 593-597, North-Holland, 1982.

12. V. Girault and P. Raviart, *Finite Element Methods for Navier-Stokes Equations,* Springer series in computational mathematics, 5, Springer-Verlag, 1986.

13. J. Douglas, Jr. and J. E. Roberts, "Global estimates for mixed methods for second order elliptic equations," *Mathematics of computation*, vol. 44, no. 169, pp. 39-52, 1985.

14. P. A. Raviart and J. M. Thomas, "A mixed finite element method for 2-nd order elliptic problems," in *Mathematical aspects of the finite element method,* Lecture Notes in Mathematics, vol. 606, pp. 292-315, Springer, 1977.

15. R. H. MacNeal, "An asymmetrical finite difference network," *Quart. Appl. Math.*, vol. XI, no. 3, pp. 295-310, 1953.

16. P. Grisvard, *Elliptic Problems in Nonsmooth Domains,* Pitman, 1985.

17. L . Fox and D. F. Mayers, *Numerical Solution of Ordinary Differential Equations for scientists and engineers,* Chapman and Hall, 1987.

18. M. Schechter, *Principles of Functional Analysis,* Academic Press, 1971.

19. R. Verfürth, "A Posteriori Error Estimators for the Stokes Equations," *Numerische Mathematik*, vol. 55, pp. 309-325, 1989.

5. A Petrov Galerkin Mixed Finite Element Method with exponential fitting.

5.1 Introduction.

The use of a form of exponential fitting for the semiconductor continuity equation is suggested by the success of the Scharfetter-Gummel discretisation [1] in one dimension and variations on that discretisation in two dimensions. Numerous derivations of Scharfetter-Gummel type discretisations are given in the literature, for instance by Selberherr [2], Markowich [3], Bank et al. [4], Brezzi et al. [5], and others. This chapter extends a one dimensional exponential fitting technique, discussed by Hemker [6], to the two dimensional problem.

In section 5.2 we introduce a model equation for the semiconductor continuity equations. We introduce several bilinear forms, related to the coefficients in this equation. In section 5.3 and 5.4 we treat the discretisation. In section 5.5 we collect some technical results and in section 5.6 we derive two error estimates. These error estimates are based on the techniques used by Douglas and Roberts [7]. The proofs in section 5.6 take all characteristics of our special discrete system into account, in particular the quadrature rule for the approximation of certain integrals in the discrete system. Note that the error estimates in section 5.6 are degenerate if the problem is singularly perturbed, i.e. if the convection dominates in the problem. On the other hand, an indication for good behaviour of the method for singular problems is that - for constant coefficients - it reproduces reproduces the solution $C\exp(-\beta_1 x_1 - \beta_2 x_2)$ exactly. In section 5.9, we develop an a posteriori error estimator. In the last section we discuss our findings.

5.2 The equation.

We consider the following problem, find $u \in \mathrm{H}^2(\Omega)$ such that:

$$-\operatorname{div}\left(\frac{1}{\alpha}(\mathbf{grad}\, u + u\boldsymbol{\beta})\right) + \gamma u = f \quad \text{on} \quad \Omega \quad \text{and} \tag{5.1}$$

$$u = -g \quad \text{on} \quad \partial\Omega ,$$

where Ω is a bounded rectangular domain in $\mathbb{R}^2$. We impose the following restrictions on the coefficients:

$$\alpha \in \mathrm{W}_1^\infty(\Omega) \quad \text{and} \quad \exists\ A \in \mathbb{R} : \alpha \geqslant A > 0 \ \text{on} \ \Omega , \tag{5.2}$$

$$\frac{1}{\alpha} \in W_1^\infty(\Omega) \text{ on } \Omega , \tag{5.3}$$

$$\boldsymbol{\beta} = (\beta_1,\beta_2)^T \text{ with } \beta_1,\beta_2 \in W_1^\infty(\Omega) , \tag{5.4}$$

$$\gamma \in W_1^\infty(\Omega) \text{ and } \gamma \geqslant 0 \text{ on } \Omega , \tag{5.5}$$

where $W_1^\infty(\Omega)$, $H^2(\Omega)$ are the usual Sobolev spaces [8], and

$$H(\mathrm{div},\Omega) := \{ \boldsymbol{\tau} \in L^2(\Omega)^2 \mid \mathrm{div}\,\boldsymbol{\tau} \in L^2(\Omega) \} ,$$

with scalar product

$$(\boldsymbol{\sigma},\boldsymbol{\tau})_{H(\mathrm{div},\Omega)} = \int_\Omega \boldsymbol{\sigma}\cdot\boldsymbol{\tau}\, d\mu + \int_\Omega \mathrm{div}\,\boldsymbol{\sigma}\ \mathrm{div}\,\boldsymbol{\tau}\, d\mu ,$$

is a Hilbert space (see also Girault and Raviart, [9] formula 2.15 in section 2.2). We assume, that the equation has a solution and that $f \in L^2(\Omega)$, $g \in H^{3/2}(\partial\Omega)$.

The stationary semiconductor continuity equations take the form (1). Here $\boldsymbol{\beta}$ corresponds to the electric field, the term γu corresponds to a linear approximation to the recombination term and $1/\alpha$ corresponds to the electron or hole mobility. The exact correspondence depends on the choice of scaling [10].

To formulate the weak mixed form of this equation, we use the following bilinear forms

$$(s,t) = \int_\Omega s\, t\, d\mu \quad \forall \ s,t \in L^2(\Omega) ,$$

$$a(\boldsymbol{\sigma},\boldsymbol{\tau}) = \int_\Omega \alpha\boldsymbol{\sigma}\cdot\boldsymbol{\tau}\, d\mu \quad \forall \ \boldsymbol{\sigma},\boldsymbol{\tau} \in H(\mathrm{div},\Omega) ,$$

$$b(\boldsymbol{\sigma},t) = \int_\Omega \boldsymbol{\beta}\cdot\boldsymbol{\sigma}\, t\, d\mu \quad \forall \ \boldsymbol{\sigma} \in H(\mathrm{div},\Omega) ,\ t \in L^2(\Omega) ,$$

$$c(s,t) = \int_\Omega \gamma\, s\, t\, d\mu \quad \forall \ s,t \in L^2(\Omega) ,$$

$$< g,h > \ = \int_{\partial\Omega} g\, h\, d\lambda \quad \forall \ g,h \in L^2(\partial\Omega) .$$

Given these definitions, we see immediately, that any solution $u \in H^2(\Omega)$ of (1) generates a solution $(\boldsymbol{\sigma},u) \in H(\mathrm{div},\Omega) \times L^2(\Omega)$ of

$$a(\boldsymbol{\sigma},\boldsymbol{\tau}) - (\mathrm{div}\,\boldsymbol{\tau},u) + b(\boldsymbol{\tau},u) = \ < g,\boldsymbol{\tau}\cdot\mathbf{n}_{\partial\Omega} > \quad \forall \ \boldsymbol{\tau} \in H(\mathrm{div},\Omega) , \tag{5.6a}$$

$$(\mathrm{div}\,\boldsymbol{\sigma},t) + c(u,t) = (f,t) \quad \forall \ t \in L^2(\Omega) . \tag{5.6b}$$

Where $\boldsymbol{\sigma} = -\frac{1}{\alpha}(\mathbf{grad}\, u + u\boldsymbol{\beta})$.

To simplify the notation, we denote the Cartesian product of a normed linear space E with itself by $\mathbf{E}$ in bold faced type, $\mathbf{E} := E \times E$. We define

$$\| (\mu_1,\mu_2)^T \|_{\mathbf{E}} := (\sum_{i=1}^{2} \| \mu_i \|_E^2)^{1/2} \quad \forall \ (\mu_1,\mu_2)^T \in \mathbf{E} .$$

5.3 Preparations.

We introduce a partition of the domain and we define the adjoint problem of (1), which we use in the derivation of one of our error estimates. Next, we introduce several special projections, that are needed in the definition of our approximation spaces and in the derivation of the error estimates. Finally we give an error estimate for the projections.

5.3.1. The partitioning of the domain.

We assume, that our domain Ω is rectangular. On Ω, we use Cartesian coordinates, with the unit vectors $\mathbf{e}_1$ and $\mathbf{e}_2$ parallel to the edges of Ω. We define $\tau_i := \boldsymbol{\tau}\cdot\mathbf{e}_i$ for $\boldsymbol{\tau} \in \mathbf{L}^2(\Omega)$ and $x_i := \mathbf{x}\cdot\mathbf{e}_i$ for $\mathbf{x} \in \mathbb{R}^2$. Before we treat our discretisation, we define our approximation space. We assume that our partition is the cartesian product of partitions

$$P = \{ \, 0 = p_0 < p_1 < \cdots < p_{N_1} = L_1 \, \} , \tag{5.7}$$

and

$$Q = \{ \, 0 = q_0 < q_1 < \cdots < q_{N_2} = L_2 \, \} \tag{5.8}$$

of the sides of our domain. We define the index set K,

$$K = \{ \, (i+½,j+½) \mid i=0,1,\ldots,N_1-1 \, , j=0,1,\ldots,N_2-1 \, \} ,$$

with the obvious index pair for a given cell,

$$\Omega_{i+½,j+½} = \{ \, \mathbf{x} \mid p_i < x_1 < p_{i+1} \, , q_j < x_2 < p_{j+1} \, \} .$$

We define $\mathbf{x}_k$ to be the centre of Ω_k and $\mathbf{h}_k$ to be the diagonal of Ω_k. We use the notation χ_k for the characteristic function of Ω_k. (The characteristic function of a set is the function that is equal to one in all points of the set and zero elsewhere). The edges of Ω_k are the sets:

$$\Gamma_{k,i,j} = \{ \, \mathbf{x} \in \overline{\Omega}_k \mid \mathbf{x}\cdot\mathbf{e}_i = (\mathbf{x}_k + (j-½)\mathbf{h}_k)\cdot\mathbf{e}_i \, \} \quad \text{for} \quad i=1,2 \, , j = 0,1 \, . \tag{5.9}$$

$\chi_{k,i,j}$ is the characteristic function of edge $\Gamma_{k,i,j}$. So $(i,j)=(1,0),(1,1),(2,0),(2,1)$ denote the left, right, bottom and top edges.

5.3.2. The adjoint problem.

We use the following definition for the adjoint problem of (1) (cf. Douglas and Roberts [7]),

$$w \in \mathrm{H}^2(\Omega) \, , \tag{5.10}$$

$$-\operatorname{div}(\frac{1}{\alpha}\,\mathbf{grad}\,w) + \frac{\boldsymbol{\beta}}{\alpha}\cdot\mathbf{grad}\,w + \gamma w = f \quad \text{on} \quad \Omega \, ,$$

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

The adjoint problem is called regular, if there is a unique solution w for every $f \in \mathrm{L}^2(\Omega)$ and this solution satisfies $\|w\|_{\mathrm{H}^2(\Omega)} \leqslant C \|f\|_{\mathrm{L}^2(\Omega)}$ for every $f \in \mathrm{L}^2(\Omega)$.

Both in the above equation and in the rest of this report, the upper case C, without a subscript, denotes a generic constant. It may have a different value at each appearance.

The weak mixed form of the adjoint problem is:

$$(\boldsymbol{\tau},w) \in \mathrm{H}(\mathrm{div},\Omega)\times \mathrm{L}^2(\Omega)\ , \tag{5.11}$$

$$a(\boldsymbol{\tau},\boldsymbol{\sigma})-(\operatorname{div}\boldsymbol{\sigma},w) = 0 \quad \forall\ \boldsymbol{\sigma}\in \mathrm{H}(\mathrm{div},\Omega)\ \text{ and} \tag{5.11a}$$

$$(\operatorname{div}\boldsymbol{\tau},t)-b(\boldsymbol{\tau},t)+c(w,t) = (f,t) \quad \forall\ t\in \mathrm{L}^2(\Omega)\ . \tag{5.11b}$$

Any solution $w \in \mathrm{H}^2(\Omega)$ of (10) generates a solution $(-\frac{1}{\alpha}\,\mathbf{grad}\,w,w)$ of this problem. If (9) is regular, then this solution satisfies the following regularity condition, $\|w\|_{\mathrm{H}^2(\Omega)} + \|\boldsymbol{\tau}\|_{\mathrm{H}^1(\Omega)} \leqslant C\|f\|_{\mathrm{L}^2(\Omega)}$.

5.3.3. Some projections.

We introduce several local projections, we use these to define four global mappings, P_h, $\mathbf{P}_h$, Π_h and $\tilde{\Pi}_h$ that map function spaces to finite dimensional function spaces. First, we define $P[\Omega_k]$ to be the orthogonal projection from $\mathrm{L}^2(\Omega_k)$ to the space of constant functions on Ω_k, and we define $P[\Gamma_{k,i,j}]$ to be the orthogonal projection from $\mathrm{L}^2(\Gamma_{k,i,j})$ to the space of constant functions on $\Gamma_{k,i,j}$.

We use $P[\Omega_k]$ to create two global mappings, $P_h\colon \mathrm{L}^2(\Omega)\rightarrow \mathrm{L}^2(\Omega)$,

$$P_hf = \sum_{k\in K} \chi_k P[\Omega_k](f) \quad \forall\ f\in \mathrm{L}^2(\Omega)\ , \tag{5.12a}$$

and $\mathbf{P}_h\colon\mathbf{L}^2(\Omega)\rightarrow\mathbf{L}^2(\Omega)$,

$$\mathbf{P_h}\boldsymbol{\beta} = \sum_{k\in K} \chi_k \left[P[\Omega_k](\boldsymbol{\beta}\cdot\mathbf{e}_1)\mathbf{e}_1 + P[\Omega_k](\boldsymbol{\beta}\cdot\mathbf{e}_2)\mathbf{e}_2\right] \quad \forall\ \boldsymbol{\beta}\in\mathbf{L}^2(\Omega)\ .\text{(5.12b)}$$

Next, we introduce two mappings, based on $P[\Gamma_{k,i,j}]$. These mappings have as their domain the space Σ,

$$\Sigma := \{\ \boldsymbol{\tau}\in \mathrm{H}(\mathrm{div},\Omega)\ \mid\ \boldsymbol{\tau}|_{\partial\Omega_k}\cdot\mathbf{n}_{\partial\Omega_k} \in \mathrm{L}^2(\partial\Omega_k) \quad \forall\ k\in K\ \}\ .$$

This space is similar to the one introduced by Roberts and Thomas in formula (1.10) of their report [11].

To simplify the definition of these mappings, we introduce local coordinates on each cell Ω_k,

$$\vec{\xi}_k := \begin{bmatrix} \dfrac{x_1 - x_{k,1}}{h_{k,1}} + \dfrac{1}{2} \\[2ex] \dfrac{x_2 - x_{k,2}}{h_{k,2}} + \dfrac{1}{2} \end{bmatrix}\ . \tag{5.13}$$

The mappings are defined as follows:

$$\Pi_h\boldsymbol{\tau} = \sum_{k\in K}\chi_k\sum_{i=1}^{2}\left[(1-\xi_{k,i})P[\Gamma_{k,i,0}](\boldsymbol{\tau}_i) + \xi_{k,i}P[\Gamma_{k,i,1}](\boldsymbol{\tau}_i)\right]\mathbf{e}_i\ , \tag{5.14}$$

$$\tilde{\Pi}_h \boldsymbol{\tau} = \sum_{k \in K} \chi_k \sum_{i=1}^{2} \Big[(1-\zeta_{k,i})\, P[\Gamma_{k,i,0}](\boldsymbol{\tau}_i) + \zeta_{k,i} P[\Gamma_{k,i,1}](\boldsymbol{\tau}_i) \Big] \mathbf{e}_i \,, \tag{5.15}$$

where

$$\zeta_{k,i} = \begin{cases} \dfrac{\exp(\xi_{k,i} h_{k,i} P[\Omega_k](\beta_i)) - 1}{\exp(h_{k,i}\; P[\Omega_k](\beta_i)) - 1} & \textit{if } P[\Omega_k](\beta_i) \neq 0 \,, \\ \xi_{k,i} & \textit{if } P[\Omega_k](\beta_i) = 0 \,. \end{cases}$$

So, for $\Pi_h \boldsymbol{\tau}$ we get the i^{th} component on Ω_k by linear interpolation between the projections of this component on the two sides orthogonal to $\mathbf{e}_i$. For $\tilde{\Pi}_h \boldsymbol{\tau}$ however, we obtain the same component by using an exponential function to interpolate between the projections of this component on the two sides orthogonal to $\mathbf{e}_i$.

Now we introduce the following finite dimensional function spaces as the ranges of the above projections,

$$V_h = \Pi_h(\Sigma) \,,\; W_h = P_h(\mathrm{L}^2(\Omega)) \;\text{ and }\; X_h = \tilde{\Pi}_h(\Sigma) \,.$$

$V_h \times W_h$ is the lowest order Raviart-Thomas-Nedelec space for rectangles. This space and the above projections were described by Douglas and Roberts, [7] Raviart and Thomas [12] and, for $\Omega \subset \mathbb{R}^3$, by Nedelec [13]. In this chapter we use the usual space, $V_h \times W_h$, as the trial function space and $X_h \times W_h$ as the test function space. In effect, we use exponential test functions instead of the usual linear test functions. Thus, we obtain a Petrov-Galerkin mixed finite element discretisation.

5.3.4. Error estimates for projections.

We prove a lemma on the accuracy of our projections. Considering the number and diversity of articles on error estimates, e.g. the classical projection estimates from Ciarlet and Raviart [14], this may seem superfluous, but we shall see that the relative simplicity of the case under consideration makes it possible to derive sharp error estimates under minimal assumptions.

Lemma 5.1.
If f is a square integrable function with square integrable derivatives on a rectangle $\Omega = [0,h_1] \times [0,h_2]$ with sides $\Gamma_{1,1} = \{\, h_1 \,\} \times [0,h_2]$, $\Gamma_{2,1} = [0,h_1] \times \{\, h_2 \,\}$, $\Gamma_{1,0} = \{\, 0 \,\} \times [0,h_2]$ and $\Gamma_{2,0} = [0,h_1] \times \{\, 0 \,\}$, then the following inequalities hold,

$$\| f - P[A] f \|_{\mathrm{L}^2(\Omega)} \leqslant (2h_1^2 + 2h_2^2)^{1/2} \, \| \mathbf{grad}\, f \|_{\mathrm{L}^2(\Omega)} \,. \tag{5.16a}$$

If s is a continuous function with domain $[0,h_1]$ and range $[0,1]$, then we have,

$$\| f - (1-s)\Pi[\Gamma_{1,0}] f - s\Pi[\Gamma_{1,1}] f \|_{\mathrm{L}^2(\Omega)} \leqslant (2h_1^2 + 2h_2^2)^{1/2} \, \| \mathbf{grad}\, f \|_{\mathrm{L}^2(\Omega)} \,. \tag{5.16b}$$

If $f \in \mathrm{L}^\infty(\Omega)$, $\mathbf{grad} f \in \mathbf{L}^\infty(\Omega)$, then

$$\| f - P[A]f \|_{\mathrm{L}^\infty(\Omega)} \leqslant (h_1 + h_2) \| \mathbf{grad} f \|_{\mathrm{L}^\infty(\Omega)} \, . \qquad (5.16c)$$

Proof.
The inequality (16b) has already been proved in lemma 4.3 of chapter 4. We start by proving the remaining inequalities for $f \in C^1(A)$. We can then extend them by the usual density argument to $\mathrm{H}^1(\Omega)$. To prove the first inequality, we write,

$$\| f - P[A]f \|^2_{\mathrm{L}^2(\Omega)} = \int_{x=0}^{h_1} \int_{y=0}^{h_2} \left(\frac{1}{h_1 h_2} \int_{w=0}^{h_1} \int_{z=0}^{h_2} f(x,y) - f(w,z) dwdz \right]^2 dxdy \, ,$$

by definition,

$$f(x,y) - f(w,z) = \int_{a=w}^{x} \frac{\partial f}{\partial a}(a,z) da + \int_{b=z}^{y} \frac{\partial f}{\partial b}(x,b) da \, .$$

If we substitute this into the above expression, then we find

$$\| f - P[\Omega]f \|^2_{\mathrm{L}^2(\Omega)} =$$

$$\int_{x=0}^{h_1} \int_{y=0}^{h_2} \left(\frac{1}{h_1 h_2} \int_{w=0}^{h_1} \int_{z=0}^{h_2} \left[\int_{a=w}^{x} \frac{\partial f}{\partial a}(a,z) da + \int_{b=z}^{y} \frac{\partial f}{\partial b}(x,b) db \right] dwdz \right]^2 dxdy \, .$$

We apply the Hölder inequality to the inner integrals and extend the integrations over a and b where appropriate,

$$\| f - P[\Omega]f \|^2_{\mathrm{L}^2(\Omega)} \leqslant$$

$$\int_{x=0}^{h_1} \int_{y=0}^{h_2} \left(\frac{h_1^{1/2}}{h_2^{1/2}} \| \partial f / \partial x_1 \|_{\mathrm{L}^2(\Omega)} + h_2^{1/2} \left(\int_{b=0}^{h_2} \left(\frac{\partial f}{\partial b}(x,b) \right]^2 db \right]^{1/2} \right]^2 dxdy \, .$$

We use $(|A| + |B|)^2 \leqslant 2(A^2 + B^2)$ to write this as,

$$\| f - P[\Omega]f \|^2_{\mathrm{L}^2(\Omega)} \leqslant$$

$$2 \int_{x=0}^{h_1} \int_{y=0}^{h_2} \frac{h_1}{h_2} \| \partial f / \partial x_1 \|^2_{\mathrm{L}^2(\Omega)} dxdy + 2 \int_{y=0}^{h_2} h_2 \| \partial f / \partial x_2 \|^2_{\mathrm{L}^2(\Omega)} dy \, .$$

This reduces to,

$$\| f - P[\Omega]f \|^2_{\mathrm{L}^2(\Omega)} \leqslant$$

$$2h_1^2 \| \partial f / \partial x_1 \|^2_{\mathrm{L}^2(\Omega)} + 2h_2^2 \| \partial f / \partial x_2 \|^2_{\mathrm{L}^2(\Omega)} \, .$$

Lastly we verify (16c),

$$f(x,y) - f(w,z) = \int_{a=w}^{x} \frac{\partial f}{\partial a}(a,z) da + \int_{b=z}^{y} \frac{\partial f}{\partial b}(x,b) da \, .$$

So,

$$\frac{1}{h_1 h_2} \int_{x=0}^{h_1} \int_{y=0}^{h_2} f(x,y) - f(w,z) dxdy =$$

$$\frac{1}{h_1 h_2} \int_{x=0}^{h_1} \int_{y=0}^{h_2} \left[\int_{a=w}^{x} \frac{\partial f}{\partial a}(a,z) da + \int_{b=z}^{y} \frac{\partial f}{\partial b}(x,b) da \right] dxdy \leqslant (h_1 + h_2) \| \mathbf{grad} f \|_{\mathbf{L}^{\infty}(\Omega)} .$$

□

Note that the above inequalities imply,

$$\| \boldsymbol{\sigma} - \Pi_h \boldsymbol{\sigma} \|_{\mathbf{L}^2(\Omega)} \leqslant \max_{k \in K} (2h_{k,1}^2 + 2h_{k,2}^2)^{1/2} \| \boldsymbol{\sigma} \|_{\mathbf{H}^1(\Omega)} , \tag{5.17a}$$

$$\| \boldsymbol{\sigma} - \tilde{\Pi}_h \boldsymbol{\sigma} \|_{\mathbf{L}^2(\Omega)} \leqslant \max_{k \in K} (2h_{k,1}^2 + 2h_{k,2}^2)^{1/2} \| \boldsymbol{\sigma} \|_{\mathbf{H}^1(\Omega)} , \tag{5.17b}$$

$$\| u - P_h u \|_{\mathrm{L}^2(\Omega)} \leqslant \max_{k \in K} (2h_{k,1}^2 + 2h_{k,2}^2)^{1/2} \| u \|_{\mathrm{H}^1(\Omega)} , \tag{5.17c}$$

for suitable u and $\boldsymbol{\sigma}$.

5.4 The discretisation.

We describe our discretisation. The basic idea of mixed finite elements with a lowest order Raviart-Thomas trial space and an exponentially fitted test subspace for the vector valued functions is complicated by the use of a quadrature rule, needed to keep the M-matrix property for the system without Lagrange multipliers for non-zero γ. This quadrature rule is discussed in section 5.4.1. Another complication is the approximation of the coefficients by piecewise constant functions, as described below. In section 5.4.2 we give the resulting discretisation.

We replace the coefficients α , $\boldsymbol{\beta}$ and γ by two dimensional step functions. To write our modified problem in weak form, we need to define three new bilinear forms,

$$\bar{a}(\boldsymbol{\sigma},\boldsymbol{\tau}) = \int_{\Omega_k} \boldsymbol{\sigma} \cdot \boldsymbol{\tau} P_h \alpha \, d\mu \quad \forall \ \boldsymbol{\sigma},\boldsymbol{\tau} \in \Sigma ,$$

$$\bar{b}(\boldsymbol{\sigma},t) := \int_{\Omega} t \boldsymbol{\sigma} \cdot \mathbf{P}_h \boldsymbol{\beta} \, d\mu \quad \forall \ \boldsymbol{\sigma} \in \Sigma, t \in \mathrm{L}^2(\Omega) ,$$

$$\bar{c}(s,t) := \int_{\Omega} s t P_h \gamma \, d\mu \quad \forall \ s,t \in \mathrm{L}^2(\Omega) .$$

The bar on the bilinear forms denotes that the coefficients are replaced by their cell-wise averages. We then replace $\bar{a}$ by $\bar{a}_q$, the subscript q indicates that a - not yet specified - quadrature rule will be used in the evaluation of this bilinear form.

5.4.1. The quadrature rule.

We construct a quadrature rule $\overline{a}_{h,1}$ by imposing the condition that, if α, $\boldsymbol{\beta}$ are constant, $\gamma \equiv 0$, $C,K \in \mathbb{R}$, and the solution satisfies $u = C\exp(-\beta_1 x_1 - \beta_2 x_2) + K$, then the discrete solution should satisfy $\boldsymbol{\sigma}_h = \Pi_h \boldsymbol{\sigma}$ and $u_h = P_h u$. We see that for the u given above $\boldsymbol{\sigma} = -K\boldsymbol{\beta}/\alpha$, so $\boldsymbol{\sigma}$ is constant. We define α_h separately for each basis function $\boldsymbol{\eta}_{i,j+½}$ where

$$\boldsymbol{\eta}_{i,j+½} = \begin{cases} \zeta_{(i-½,j+½),1}\mathbf{e}_1 \ \text{ on } \ \Omega_{i-½,j+½} \ , \\ (1-\zeta_{(i+½,j+½),1})\mathbf{e}_1 \ \text{ on } \ \Omega_{i+½,j+½} \ , \\ 0 \ \text{ elsewhere} \ , \end{cases}$$

and $\boldsymbol{\eta}_{i+½,j}$ where

$$\boldsymbol{\eta}_{i+½,j} = \begin{cases} \zeta_{(i+½,j-½),2}\mathbf{e}_2 \ \text{ on } \ \Omega_{i+½,j-½} \ , \\ (1-\zeta_{(i+½,j-½),2})\mathbf{e}_2 \ \text{ on } \ \Omega_{i+½,j+½} \ . \\ 0 \ \text{ elsewhere} \ . \end{cases}$$

We denote the set of all possible indices for the basis functions $\boldsymbol{\eta}$ by

$$E = \{ \ e=(i,j-½) \mid i=0,1,2, \ldots, N_1 \ , j=1,2, \ldots, N_2 \ \} \cup \{ \ e=(i-½,j) \mid i=1,2, \ldots, N_1 \ , j=0,1,2, \ldots, N_2 \ \} \ .$$

Our quadrature rule should satisfy the following condition,

$$\overline{a}_{h,1}(\boldsymbol{\sigma},\boldsymbol{\eta}_r) = \overline{a}(\boldsymbol{\sigma},\boldsymbol{\eta}_r) \ , \tag{A}$$

for all constant $\boldsymbol{\sigma}$ and all $r \in E$. Due to our assumption that the coefficients are constant, we have $a=\overline{a}$ and $b=\overline{b}$. The above condition guarantees that, for constant coefficients and constant $\boldsymbol{\sigma}$,

$$a(\boldsymbol{\sigma},\boldsymbol{\tau}_h) - (u, \operatorname{div} \boldsymbol{\tau}_h) + b(\boldsymbol{\tau}_h,u) = \overline{a}_{h,1}(\Pi_h\boldsymbol{\sigma},\boldsymbol{\tau}_h) - (P_h u, \operatorname{div} \boldsymbol{\tau}_h) + b(\boldsymbol{\tau}_h,P_h u) \quad \forall \ \ \boldsymbol{\tau}_h \in X_h$$

and we also have,

$$(\operatorname{div} \boldsymbol{\sigma},t) = (\operatorname{div} \Pi_h \boldsymbol{\sigma},t) = 0 \quad \forall \ \ t \in L^2(\Omega) \ .$$

So our condition (A) on $\overline{a}_{h,1}$ is sufficient for our purposes. We now select the quadrature rule by taking the following definition for $\overline{a}_{h,1}$,

$$\overline{a}_{h,1}(\boldsymbol{\sigma},\boldsymbol{\tau}) = \tag{5.18a}$$

$$\sum_{k \in K} \sum_{i=1}^{2} \mu(\Omega_k) P[\Omega_k](\alpha) \Big[P[\Omega_k](\zeta_{k,i}) P[\Gamma_{k,i,1}](\sigma_i \tau_i) \ + \ P[\Omega_k](1-\zeta_{k,i}) P[\Gamma_{k,i,0}](\sigma_i \tau_i) \Big] \ .$$

We introduce a new problem dependent norm on X_h

$$\|\boldsymbol{\tau}_h\|_h = \tag{5.18b}$$

$$\sum_{k \in K} \sum_{i=1}^{2} \mu(\Omega_k) \Big[P[\Omega_k](\zeta_{k,i}) P[\Gamma_{k,i,1}](\tau_{h,i}^2) \ + \ P[\Omega_k](1-\zeta_{k,i}) P[\Gamma_{k,i,0}](\tau_{h,i}^2) \Big]^{½} \ .$$

From this point onwards, we take $\overline{a}_q = \overline{a}_{h,1}$.

5.4.2. The discrete system.

We approximate the solution $(\boldsymbol{\sigma},u)$ of (6) by $(\boldsymbol{\sigma}_h,u_h) \in V_h \times W_h$, where

$$\overline{a}_q(\boldsymbol{\sigma}_h,\boldsymbol{\tau}) - (u_h, \operatorname{div}\boldsymbol{\tau}) + \overline{b}(\boldsymbol{\tau},u_h) = \; < \boldsymbol{\tau}\cdot\mathbf{n}_{\partial\Omega}, g > \quad \forall\; \boldsymbol{\tau} \in X_h \;, \tag{5.19a}$$

$$(\operatorname{div}\boldsymbol{\sigma}_h,t) + \overline{c}(u_h,t) = (f,t) \quad \forall\; t \in W_h \;. \tag{5.19b}$$

If we use $\overline{a}$ in stead of $\overline{a}_q$, then that means that our discrete problem does not always yield an M-matrix for u_h. Consider, for instance, the corresponding discretisation on a uniform mesh with mesh width h in one dimension with $\alpha = 1$, $\boldsymbol{\beta} = \vec{0}$ and γ constant. If $\alpha\gamma h^2/6 > 1$, then the off-diagonal elements of the discretisation matrix for u_h after elimination of $\boldsymbol{\sigma}_h$ through static condensation have the same sign as the elements on the diagonal.

The idea of using linear trial functions and exponential test functions was used by Hemker for singularly perturbed two point boundary problems [6]. For the one dimensional case, the introduction of exponential test functions follows from the requirement that it must be possible to approximate the Green's function of the problem by the test functions. For finite elements in one dimension the the singularly perturbed case was studied by O'Riordan and Stynes [15-20] and Reinhardt [21]. For finite element in two dimensions O'Riordan and Stynes derive a uniformly convergent estimate [22] but only for problems with a strictly positive zero-order term.

In the following sections, we prove, that the solution of our discretisation (19) is an $\mathcal{O}(h)$ approximation to the solution of our original problem.

5.5 Several technical results.

This section contains some technical results, collected for later reference.

Lemma 5.2.

$$\tilde{\Pi}_h \circ \Pi_h = \tilde{\Pi}_h \;, \tag{5.20a}$$

$$\Pi_h \circ \tilde{\Pi}_h = \Pi_h \;, \tag{5.20b}$$

$$(\operatorname{div}\sigma, P_h t) = (\operatorname{div}\Pi_h\sigma, t) \quad \forall\; \sigma \in \Sigma \;,\; t \in \mathrm{L}^2(\Omega) \;, \tag{5.20c}$$

$$\Pi_h\tau\cdot\vec{n}_{\partial\Omega} = \tilde{\Pi}_h\tau\cdot\vec{n}_{\partial\Omega} \quad \forall\; \tau \in \Sigma \;. \tag{5.20d}$$

Proof.
Both mappings are based on the same projections $P[\Gamma_{k,i,j}]$, so (20a) and (20b) are trivial.

To prove (20c) we use a special case of Green's theorem:

$$\int_{\Omega_k} \operatorname{div}\sigma \; d\mu = \sum_{i=1}^{2} \frac{\mu(\Omega_k)}{h_{k,i}} \left[P[\Gamma_{k,i,1}](\sigma_i) - P[\Gamma_{k,i,0}](\sigma_i) \right] .$$

If we combine this with the definition of Π_h, the proof of (20c) is complete. Equation (20d) follows immediately from the definitions. □

Lemma 5.3.
If $\boldsymbol{\sigma} \in \Sigma$ and we define $a_{k,i} = P[\Gamma_{k,i,0}](\boldsymbol{\sigma}_i)$ and $b_{k,i} = P[\Gamma_{k,i,1}](\boldsymbol{\sigma}_i)$, then the following inequalities hold for $\|\Pi_h\boldsymbol{\sigma}\|_{\mathbf{L}^2(\Omega_k)}$ and $\|\Pi_h\boldsymbol{\sigma}\|_{\mathbf{L}^2(\Omega_k)}$,

$$\frac{\mu(\Omega_k)}{6}\sum_{i=1}^{2}(a_{k,i}^2+b_{k,i}^2) \leqslant \|\Pi_h\boldsymbol{\sigma}\|_{\mathbf{L}^2(\Omega_k)}^2 \leqslant \frac{\mu(\Omega_k)}{2}\sum_{i=1}^{2}(a_{k,i}^2+b_{k,i}^2)\,. \tag{5.21a}$$

$$\|\tilde{\Pi}_h\boldsymbol{\sigma}\|_{\mathbf{L}^2(\Omega_k)}^2 \leqslant 2\|\tilde{\Pi}_h\boldsymbol{\sigma}\|_h^2 \leqslant 12\|\Pi_h\boldsymbol{\sigma}\|_{\mathbf{L}^2(\Omega)}^2\,. \tag{5.21b}$$

Cf. chapter 2, section 2.5.2, lemma 2.9 and chapter 4, lemma 4.1.
Proof.
Formula (21a) follows immediately from

$$(\Pi_h\boldsymbol{\sigma},\Pi_h\boldsymbol{\sigma}) = \sum_{i=1}^{2}\sum_{k\in K}\int_{\Omega_k}\Big[(1-\xi_{k,i})a_{k,i} + \xi_{k,i}b_{k,i}\Big]^2 d\mu\,.$$

Next, we derive (21b) from,

$$(\tilde{\Pi}_h\boldsymbol{\sigma},\tilde{\Pi}_h\boldsymbol{\sigma}) = \sum_{i=1}^{2}\sum_{k\in K}\int_{\Omega_k}\Big[(1-\zeta_{k,i})a_{k,i} + \zeta_{k,i}b_{k,i}\Big]^2 d\mu\,.$$

We see immediately that

$$\int_{\Omega_k}\Big[(1-\zeta_{k,i})a_{k,i} + \zeta_{k,i}b_{k,i}\Big]^2 d\mu \leqslant \int_{\Omega_k}2(1-\zeta_{k,i})^2a_{k,i}^2 + 2\zeta_{k,i}^2b_{k,i}^2 d\mu \leqslant$$

$$2\int_{\Omega_k}(1-\zeta_{k,i})a_{k,i}^2 + \zeta_{k,i}b_{k,i}^2 d\mu = 2\mu(\Omega_k)\Big[P[\Omega_k](1-\zeta_{k,i})a_{k,i}^2 + P[\Omega_k](\zeta_{k,i})b_{k,i}^2\Big]\,.$$

This implies (21b). □

Lemma 4 shows, that $\bar{a}$ is $\mathbf{L}^2(\Omega)$-bounded and $\mathbf{L}^2(\Omega)$-elliptic.

Lemma 5.4.
Let $\alpha \in \mathrm{W}_1^\infty(\Omega)$, $\alpha \geqslant A > 0$ on Ω and $\bar{a}(\boldsymbol{\sigma},\boldsymbol{\tau}) := \int_\Omega P_h(\alpha)\,\boldsymbol{\sigma}\cdot\boldsymbol{\tau}\,d\mu$ $\forall\ \boldsymbol{\sigma},\boldsymbol{\tau} \in \mathbf{L}^2(\Omega)$, then

$$\bar{a}(\boldsymbol{\sigma},\boldsymbol{\tau}) \leqslant \|\alpha\|_{\mathrm{L}^\infty(\Omega)}\|\boldsymbol{\sigma}\|_{\mathbf{L}^2(\Omega)}\|\boldsymbol{\tau}\|_{\mathbf{L}^2(\Omega)} \quad \forall\ \boldsymbol{\sigma},\boldsymbol{\tau} \in \mathbf{L}^2(\Omega)\,, \tag{5.22a}$$

and

$$\bar{a}(\boldsymbol{\tau},\boldsymbol{\tau}) \geqslant A\|\boldsymbol{\tau}\|_{\mathbf{L}^2(\Omega)}^2 \quad \forall\ \boldsymbol{\tau} \in \mathbf{L}^2(\Omega)\,. \tag{5.22b}$$

Proof.
From (2) it follows that,

$$A \leqslant \frac{\int_{\Omega_k}\alpha\,d\mu}{\mu(\Omega_k)} \leqslant \|\alpha\|_{\mathrm{L}^\infty(\Omega_k)}\,,$$

together with the definitions of P and $\overline{a}$ this implies (22a) and (22b). □.

We introduce the minimum mesh width $h_{\min}$ and the maximum mesh width $h_{\max}$,

$$h_{\min} = \min_{k \in K} \min_{i=1,2} |h_{k,i}| , \tag{5.23a}$$

$$h_{\max} = \max_{k \in K} \max_{i=1,2} |h_{k,i}| . \tag{5.23b}$$

5.5.1. The properties of $\overline{a}_q$.

We discuss the properties of the quadrature rule $\overline{a}_q$. We assume that $\overline{a}_q = \overline{a}_{h,1}$, where $\overline{a}_{h,1}$ is given by (18a). We discuss the interaction between Π, $\tilde{\Pi}$ and $\overline{a}_q$. We show, that $\overline{a}_q$ is $\mathbf{L}^2(\Omega)$-bounded on V_h, and we also show, that $\overline{a}_q$ is $\mathbf{L}^2(\Omega)$-elliptic on V_h and X_h.

Lemma 5.5.
If $\boldsymbol{\sigma},\boldsymbol{\tau} \in \Sigma$, then

$$\begin{aligned} \overline{a}_q(\Pi_h\boldsymbol{\sigma},\Pi_h\boldsymbol{\tau}) = \overline{a}_q(\Pi_h\boldsymbol{\tau},\Pi_h\boldsymbol{\sigma}) = \overline{a}_q(\boldsymbol{\sigma},\Pi_h\boldsymbol{\tau}) = \overline{a}_q(\Pi_h\boldsymbol{\sigma},\boldsymbol{\tau}) = \\ \overline{a}_q(\boldsymbol{\sigma},\tilde{\Pi}_h\boldsymbol{\tau}) = \overline{a}_q(\tilde{\Pi}_h\boldsymbol{\sigma},\boldsymbol{\tau}) = \overline{a}_q(\tilde{\Pi}_h\boldsymbol{\sigma},\tilde{\Pi}_h\boldsymbol{\tau}) , \end{aligned} \tag{5.24a}$$

$$\begin{aligned} \|\alpha\|_{L^\infty(\Omega)} \|\tilde{\Pi}_h\boldsymbol{\sigma}\|_h^2 \geqslant \overline{a}_q(\tilde{\Pi}_h\boldsymbol{\sigma},\tilde{\Pi}_h\boldsymbol{\sigma}) \geqslant \\ ½\overline{a}(\tilde{\Pi}_h\boldsymbol{\sigma},\tilde{\Pi}_h\boldsymbol{\sigma}) \geqslant \frac{A}{2} \|\tilde{\Pi}_h\boldsymbol{\sigma}\|_{\mathbf{L}^2(\Omega)}^2 , \end{aligned} \tag{5.24b}$$

$$\overline{a}_q(\Pi_h\boldsymbol{\sigma},\Pi_h\boldsymbol{\tau}) \leqslant 6\|\alpha\|_{L^\infty(\Omega)} \|\Pi_h\boldsymbol{\sigma}\|_{\mathbf{L}^2(\Omega)} \|\tilde{\Pi}_h\boldsymbol{\tau}\|_h , \tag{5.24c}$$

$$A\|\tilde{\Pi}_h\boldsymbol{\tau}\|_h^2 \leqslant \overline{a}_q(\boldsymbol{\tau},\tilde{\Pi}_h\boldsymbol{\tau}) \leqslant \|\alpha\|_{L^\infty(\Omega)} \|\tilde{\Pi}_h\boldsymbol{\tau}\|_h^2 . \tag{5.24d}$$

Proof.
The definitions of Π_h, $\tilde{\Pi}_h$ and $\overline{a}_q$ imply (24a). Inequality (24b) follows immediately from (18a), (18b) and (21b). To prove (24c), we need some auxiliary variables, $a_{k,i} = P[\Gamma_{k,i,0}](\boldsymbol{\sigma})$, $b_{k,i} = P[\Gamma_{k,i,1}](\boldsymbol{\sigma})$, $c_{k,i} = P[\Gamma_{k,i,0}](\boldsymbol{\tau})$ and $d_{k,i} = P[\Gamma_{k,i,1}](\boldsymbol{\tau})$. We use Cauchy-Schwartz twice to obtain

$$\overline{a}_q(\Pi\boldsymbol{\sigma},\tilde{\Pi}\boldsymbol{\tau}) = \sum_{k \in K} \int_{\Omega_k} \alpha \, d\mu \sum_{i=1}^{2} (P[\Omega_k](1-\zeta_{k,i})a_{k,i}c_{k,i} + P[\Omega_k](\zeta_{k,i})b_{k,i}d_{k,i}) \leqslant$$

$$\sum_{k \in K} \int_{\Omega_k} \alpha \, d\mu \left[\sum_{i=1}^{2} (a_{k,i}^2 + b_{k,i}^2) \right]^{½} \left[\sum_{i=1}^{2} [P[\Omega_k](1-\zeta_{k,i})^2 c_{k,i}^2 + P[\Omega_k](\zeta_{k,i})^2 d_{k,i}^2] \right]^{½} .$$

We use

$$P[\Omega_k](f)^2 \leqslant P[\Omega_k](f^2)$$

to rewrite the term in c and d and we use (21a) to replace the term in a and b by $\|\Pi_h\boldsymbol{\sigma}\|_{\mathbf{L}^2(\Omega_k)}$,

$$\overline{a}_q(\Pi_h\boldsymbol{\sigma},\tilde{\Pi}_h\boldsymbol{\tau}) \leqslant$$

$$6 \int_{\Omega_k} \alpha \, d\mu \, \frac{\|\Pi_h \boldsymbol{\sigma}\|_{\mathbf{L}^2(\Omega_k)}}{\mu(\Omega_k)^{1/2}} \left[\sum_{i=1}^{2} (P[\Omega_k]((1-\zeta_{k,i})^2) c_{k,i}^2 + P[\Omega_k]((\zeta_{k,i})^2) d_{k,i}^2) \right]^{1/2} .$$

We see immediately that this implies,

$$\overline{a}_q(\Pi_h \boldsymbol{\sigma}, \tilde{\Pi}_h \boldsymbol{\tau}) \leqslant$$

$$6 \|\alpha\|_{\mathrm{L}^\infty(\Omega)} \; \|\Pi_h \boldsymbol{\sigma}\|_{\mathbf{L}^2(\Omega)} \; \|\tilde{\Pi}_h \boldsymbol{\tau}\|_h \, .$$

This proves (24c). Inequality (24d) follows immediately from (18).

□

5.5.2. The difference between $\overline{a}$ and $\overline{a}_q$.

For our error estimates, we need an upper bound for the difference between the value of $a(\boldsymbol{\sigma}_h, \boldsymbol{\tau})$ and that of $\overline{a}_q(\boldsymbol{\sigma}_h, \boldsymbol{\tau})$ for $\boldsymbol{\sigma}_h \in V_h$, $\boldsymbol{\tau} \in \mathbf{H}^1(\Omega)$. As we already know from (16c) (see also Lemmas 8 and 9) that,

$$| a(\boldsymbol{\sigma}, \boldsymbol{\tau}_h) - \overline{a}(\boldsymbol{\sigma}, \boldsymbol{\tau}_h) | \leqslant 2 h_{\max} \|\alpha\|_{\mathrm{W}_1^\infty(\Omega)} \|\boldsymbol{\sigma}\|_{\mathbf{L}^2(\Omega)} \|\boldsymbol{\tau}_h\|_{\mathbf{L}^2(\Omega)} \, ,$$

an estimate for $|\overline{a}(\boldsymbol{\sigma}, \boldsymbol{\tau}_h) - \overline{a}_q(\boldsymbol{\sigma}, \boldsymbol{\tau}_h)|$ suffices. Such an estimate is derived in lemma 6.

Lemma 5.6.
Let $\boldsymbol{\tau}_h \in X_h$ and $\boldsymbol{\sigma} \in \mathbf{H}^1(\Omega)$, then

$$| \overline{a}(\boldsymbol{\sigma}, \boldsymbol{\tau}_h) - \overline{a}_q(\boldsymbol{\sigma}, \boldsymbol{\tau}_h) | \leqslant 2 \|\alpha\|_{\mathrm{L}^\infty(\Omega)} \, h_{\max} \, \|\boldsymbol{\tau}_h\|_h \|\boldsymbol{\sigma}\|_{\mathbf{H}^1(\Omega)} \, . \tag{5.25}$$

Proof.
To simplify our notation, we introduce $a_{k,i} = P[\Gamma_{k,i,0}](\boldsymbol{\tau}_h)$, $b_{k,i} = P[\Gamma_{k,i,1}](\boldsymbol{\tau}_h)$, $\sigma_{k,i,0} = P[\Gamma_{\underline{k},i,0}](\sigma_i)$ and $\sigma_{k,i,1} = P[\Gamma_{k,i,1}](\sigma_i)$. We prove the lemma for $\boldsymbol{\sigma}$ with $\sigma_1, \sigma_2 \in C^1(\Omega)$, and extend by density.

We consider the difference between the two forms on one subdomain Ω_k with $P[\Omega_k](\alpha) = 1$.

$$| \int_{\Omega_k} \boldsymbol{\sigma} \cdot \boldsymbol{\tau}_h - \sum_{i=1}^{2} \left[P[\Omega_k](1-\zeta_{k,i}) P[\Gamma_{k,i,0}](\sigma_i \tau_{h,i}) + P[\Omega_k](\zeta_{k,i}) P[\Gamma_{k,i,1}](\sigma_i \tau_{h,i}) \right] d\mu | =$$

$$| \int_{\Omega_k} \sum_{i=1}^{2} \left[(1-\zeta_{k,i}) a_{k,i} + \zeta_{k,i} b_{k,i} \right] \sigma_i \, d\mu -$$

$$\mu(\Omega_k) \sum_{i=1}^{2} \left[P[\Omega_k](1-\zeta_{k,i}) P[\Gamma_{k,i,0}](a_{k,i} \sigma_i) + P[\Omega_k](\zeta_{k,i}) P[\Gamma_{k,i,1}](b_{k,i} \sigma_i) \right] | =$$

$$| \int_{\Omega_k} \sum_{i=1}^{2} \left[(1-\zeta_{k,i}) a_{k,i} \sigma_i + \zeta_{k,i} b_{k,i} \sigma_i - P[\Omega_k](1-\zeta_{k,i}) a_{k,i} \sigma_{k,i,0} - P[\Omega_k](\zeta_{k,i}) b_{k,i} \sigma_{k,i,1} \right] d\mu | =$$

$$| \int_{\Omega_k} \sum_{i=1}^{2} \left[(1-\zeta_{k,i}) a_{k,i} (\sigma_i - \sigma_{k,i,0}) + \zeta_{k,i} b_{k,i} (\sigma_i - \sigma_{k,i,1}) \right] d\mu | \, ,$$

The application of the Cauchy-Schwartz inequality to this last term and insertion of α yields the following result,

$$|\bar{a}(\boldsymbol{\sigma},\boldsymbol{\tau}_h) - \bar{a}_q(\boldsymbol{\sigma},\boldsymbol{\tau}_h)| \leqslant$$
$$h_{\max} \,||\,\alpha\,||_{\,\mathrm{L}^\infty(\Omega)} \,||\,\boldsymbol{\tau}_h\,||_{\,h} \left[\sum_{k \in K} \sum_{i=1}^{2} \sum_{j=0}^{1} ||\,\sigma_i - \sigma_{i,k,j}\,||^2_{\,\mathrm{L}^2(\Omega_k)} \right]^{1/2} .$$

If we take $s \equiv j$ in (16b) then this implies,

$$|\bar{a}(\boldsymbol{\sigma},\boldsymbol{\tau}_h) - \bar{a}_q(\boldsymbol{\sigma},\boldsymbol{\tau}_h)| \leqslant ||\,\alpha\,||_{\,\mathrm{L}^\infty(\Omega)} \,||\,\boldsymbol{\tau}_h\,||_{\,h} \left[\sum_{i=1}^{2} 4h^2_{\max} \,||\,\mathbf{grad}\,\sigma_i\,||^2_{\,\mathbf{L}^2(\Omega)} \right]^{1/2} \leqslant$$
$$2h_{\max} \,||\,\alpha\,||_{\,\mathrm{L}^\infty(\Omega)} \,||\,\boldsymbol{\tau}_h\,||_{\,h} \,||\,\boldsymbol{\sigma}\,||_{\,\mathbf{H}^1(\Omega)} \,.$$

Because $C^1(\bar{\Omega})$ is dense in $\mathrm{H}^1(\Omega)$, the formula also holds for $\sigma_1,\sigma_2 \in \mathrm{H}^1(\Omega)$.

□

5.6 The error estimates.

We use the standard estimates for $||\,\boldsymbol{\sigma} - \Pi_h \boldsymbol{\sigma}\,||_{\,\mathbf{L}^2(\Omega)}$ and $||\,u - P_h u\,||_{\,\mathrm{L}^2(\Omega)}$, as described in section 5.3.4, to reduce the problem to deriving bounds for $||\,P_h u - u_h\,||_{\,\mathrm{L}^2(\Omega)}$ and $||\,\Pi_h \boldsymbol{\sigma} - \boldsymbol{\sigma}_h\,||_{\,\mathbf{L}^2(\Omega)}$. We discuss two possible derivations of an $\mathcal{O}(h)$ error bound. The first derivation needs the assumption, that $h_{\max}$ is "small enough", the second derivation places a condition on an approximation of the discrete version of the adjoint problem.

5.6.1. Errors due to approximation of the bilinear forms.

As preparation for the derivation of a priori error estimates, we derive some upper bounds on the errors caused by the piecewise constant approximation of the coefficients $\alpha, \boldsymbol{\beta}$ and γ. We use the following well-known notation. If V and W are normed linear spaces, then $\mathscr{L}(V,W;\mathbb{R})$ is the space of bounded bilinear forms on V and W. the standard norm of an element $b \in \mathscr{L}(V,W;\mathbb{R})$ is given by

$$||\,b\,||_{\,\mathscr{L}(V,W;\mathbb{R})} = \sup_{v \in V} \sup_{w \in W} \frac{|\,b(v,w)\,|}{||\,v\,||_{\,V} \,||\,w\,||_{\,W}} \,.$$

Lemma 5.7.
If $\alpha \in \mathrm{W}_1^\infty(\Omega)$ then

$$||\,a - \bar{a}_q\,||_{\,\mathscr{L}(\mathbf{H}^1(\Omega),(X_h,\,||.||_h);\mathbb{R})} \leqslant 6h_{\max} \,||\,\alpha\,||_{\,\mathrm{W}_1^\infty(\Omega)} \,,$$

where $(X_h, ||\,.\,||_{\,h})$ is a normed linear space with as elements the elements of X_h but with $||\,.\,||_{\,h}$ as norm.
Proof.
From equation (16c) and (21b) it follows that,

$$|a(\boldsymbol{\sigma},\boldsymbol{\tau}_h) - \bar{a}(\boldsymbol{\sigma},\boldsymbol{\tau}_h)| \leqslant 4h_{\max} \|\alpha\|_{W_1^\infty(\Omega)} \|\boldsymbol{\sigma}\|_{\mathbf{L}^2(\Omega)} \|\boldsymbol{\tau}_h\|_h \ .$$

When combined with lemma 6, this implies

$$\|a - \bar{a}_q\|_{\mathscr{L}(\mathbf{H}^1(\Omega),(X_h,\|.\|_h);\mathbb{R})} \leqslant 6h_{\max} \|\alpha\|_{W_1^\infty(\Omega)} \ .$$

□

Lemma 5.8.
If $\boldsymbol{\beta} \in \mathbf{W}_1^\infty(\Omega)$ then

$$\|b - \bar{b}\|_{\mathscr{L}(\mathbf{L}^2(\Omega),\, \mathrm{L}^2(\Omega)\,;\mathbb{R})} \leqslant 4h_{\max} \|\boldsymbol{\beta}\|_{\mathbf{W}_1^\infty(\Omega)} \ .$$

Proof.
This follows immediately from (16c).

□

Lemma 5.9.
If $\gamma \in \mathrm{W}_1^\infty(\Omega)$ then

$$\|c - \bar{c}\|_{\mathscr{L}(\mathrm{L}^2(\Omega),\, \mathrm{L}^2(\Omega)\,;\mathbb{R})} \leqslant 2h_{\max} \|\gamma\|_{\mathrm{W}_1^\infty(\Omega)} \ .$$

Proof.
This follows immediately from (16c).

□

5.6.2. An a priori error estimate.

The following two lemmas show nice properties of our discretisation. We need these properties to derive the error bound.

Lemma 5.10.
Let $\boldsymbol{\tau} \in \Sigma$, $t \in \mathrm{L}^2(\Omega)$, then

$$\bar{b}(\tilde{\Pi}_h\boldsymbol{\tau}, t - P_h t) - (\operatorname{div} \tilde{\Pi}_h\boldsymbol{\tau}, t - P_h t) = 0 \ . \tag{5.26}$$

Proof.
A straightforward calculation shows that $\mathbf{P}_h(\boldsymbol{\beta})\cdot\tilde{\Pi}_h\boldsymbol{\tau} - \operatorname{div}\tilde{\Pi}_h\boldsymbol{\tau}$ is constant on Ω_k. From this (26) easily follows. □

Lemma 5.11.
If $(\boldsymbol{\sigma},u)$ is a solution of (6) and $(\boldsymbol{\sigma}_h,u_h)$ is a solution of (19), then

$$(\operatorname{div}(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h), P_h t) + c(u-u_h, P_h t) = 0 \quad \forall \ t \in \mathrm{L}^2(\Omega) \ . \tag{5.27}$$

Proof.
We take (19b),

$$(\operatorname{div}\boldsymbol{\sigma}_h,P_ht) + \overline{c}(u_h,P_ht) = (f,P_ht)\ ,$$

$\overline{c}$ is derived by orthogonal $L^2(\Omega_k)$ projection, so this implies

$$(\operatorname{div}\boldsymbol{\sigma}_h,P_ht) + c(u_h,P_ht) = (f,P_ht)\ .$$

If we subtract this from (6b), $(\operatorname{div}\boldsymbol{\sigma},P_ht) + c(u,P_ht) = (f,P_ht)$, then we find (27). □

We are now ready to give an estimate for $\| \Pi_h\boldsymbol{\sigma}-\boldsymbol{\sigma}_h \|_h$.

Theorem 5.1.
If $(\boldsymbol{\sigma},u)$ is the solution of (6), $(\boldsymbol{\sigma}_h,u_h)$ is the solution of (19) and $(\boldsymbol{\sigma},u) \in \mathbf{H}^1(\Omega)\times H^2(\Omega)$, then there exist positive real numbers C and D such that

$$C < \tag{5.28}$$

$$\frac{12}{A}\max(1, \| \alpha \|_{W_1^\infty(\Omega)}, \| \boldsymbol{\beta} \|_{\mathbf{W}_1^\infty(\Omega)}, \| \gamma \|_{W_1^\infty(\Omega)}) \max(1, \| \boldsymbol{\sigma} \|_{H^1(\Omega)}, \| u \|_{L^2(\Omega)})\ ,$$

$$D < 2\frac{\| \boldsymbol{\beta} \|_{L^\infty(\Omega)}}{A}\ ,\ :$$

$$\| \Pi_h\boldsymbol{\sigma} - \boldsymbol{\sigma}_h \|_h^2 \leqslant Ch_{\max}(\| \Pi_h\boldsymbol{\sigma} - \boldsymbol{\sigma}_h \|_h + \| P_hu - u_h \|_{L^2(\Omega)}) +$$
$$D \| \Pi_h\boldsymbol{\sigma} - \boldsymbol{\sigma}_h \|_h \| P_hu - u_h \|_{L^2(\Omega)}\ .$$

Proof.
According to (24d), $A \| \tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h) \|_h^2 \leqslant \overline{a}_q(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h, \tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h))$. This is the starting point for the derivation of our error bound. Equations (6a) and (19a) imply, that

$$\overline{a}_q(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h,\tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h)) =$$
$$(\overline{a}_q-a)(\boldsymbol{\sigma},\tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h)) + a(\boldsymbol{\sigma},\tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h)) - \overline{a}_q(\boldsymbol{\sigma}_h,\tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h)) =$$
$$(\overline{a}_q-a)(\boldsymbol{\sigma},\tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h)) + (\operatorname{div}\tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h),u) -$$
$$b(\tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h),u) + \ < g,\mathbf{n}_{\partial\Omega}\cdot\tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h) > \ +$$
$$\overline{b}(\tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h),u_h) - (\operatorname{div}\tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h),u_h) - \ < g,\mathbf{n}_{\partial\Omega}\cdot\tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h) > \ =$$
$$(\overline{a}_q-a)(\boldsymbol{\sigma},\tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h)) + (\operatorname{div}\tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h),u)_{L^2(\Omega)} - (b-\overline{b})(\tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h),u) -$$
$$\overline{b}(\tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h),u) + \overline{b}(\tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h),u_h) - (\operatorname{div}\tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h),u_h)\ .$$

Where we give $b-\overline{b}$, $\overline{a}_q-a$ etc. their obvious meaning. If we use lemma 10, we find:

$$A \| \tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h) \|_h^2 \leqslant (\overline{a}_q-a)(\boldsymbol{\sigma},\tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h)) - (b-\overline{b})(\tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h),u) +$$
$$(\operatorname{div}\tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h),P_hu-u_h)_{L^2(\Omega)} - \overline{b}(\tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h),P_hu-u_h)\ .$$

If we use (20b) and (20c) to prepare the way, then the application of lemma 11 to this expression results in:

$$A \,\|\, \tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h) \,\|^2_{\mathrm{L}^2(\Omega)} \leqslant (\overline{a}_q - a)(\boldsymbol{\sigma},\tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h)) - (b-\overline{b})(\tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h),u) -$$

$$c(u-u_h,P_hu-u_h) - \overline{b}(\tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h),P_hu-u_h)\ .$$

As γ is non-negative according to (5), we may add $c(P_hu-u_h,P_hu-u_h)$ on both sides of the inequality, we find,

$$A \,\|\, \tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h) \,\|^2_{\mathrm{L}^2(\Omega)} + c(P_hu-u_h,P_hu-u_h) \leqslant$$

$$(\overline{a}_q - a)(\boldsymbol{\sigma},\tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h)) - (b-\overline{b})(\tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h),u) -$$

$$(c-\overline{c})(u-P_hu,P_hu-u_h) - \overline{b}(\tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h),Phu-u_h)\ .$$

We use lemmas 7, 8 and 9 to reduce this to,

$$A \,\|\, \tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h) \,\|^2_h \leqslant$$

$$h_{\max} \left[6 \,\|\, \alpha \,\|_{\mathrm{W}_1^\infty(\Omega)} \,\|\, \boldsymbol{\sigma} \,\|_{\mathrm{H}^1(\Omega)} + 4 \,\|\, \boldsymbol{\beta} \,\|_{\mathbf{W}_1^\infty(\Omega)} \,\|\, u \,\|_{\mathrm{L}^2(\Omega)} \right] \|\, \tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h) \,\|_h +$$

$$2h_{\max} \,\|\, \gamma \,\|_{\mathrm{W}_1^\infty(\Omega)} \,\|\, u-P_hu \,\|_{\mathrm{L}^2(\Omega)} \,\|\, P_hu-u_h \,\|_{\mathrm{L}^2(\Omega)} +$$

$$2 \,\|\, \boldsymbol{\beta} \,\|_{\mathbf{L}^\infty(\Omega)} \,\|\, \tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h) \,\|_{\mathbf{L}^2(\Omega)} \,\|\, Pu-u_h \,\|_{\mathrm{L}^2(\Omega)}\ .$$

Note that for all $u \in \mathrm{L}^2(\Omega)$, $\|\, u-P_hu \,\|_{\mathrm{L}^2(\Omega)} \leqslant \|\, u \,\|_{\mathrm{L}^2(\Omega)}$ and $\|\, \Pi_h\boldsymbol{\sigma} \,\|_h = \|\, \tilde{\Pi}_h\boldsymbol{\sigma} \,\|_h$.

□

Next, we prepare for the second part of our error estimate.

Lemma 5.12.
If $(\boldsymbol{\sigma},u)$ is the solution of (6), $(\boldsymbol{\sigma}_h,u_h)$ is a solution of (19) and $(\boldsymbol{\tau},q)$ is the solution of the adjoint problem for an arbitrary right hand side $p \in \mathrm{L}^2(\Omega)$, then

$$(\operatorname{div}\boldsymbol{\tau},P_hu-u_h) - b(\boldsymbol{\tau},P_hu-u_h) =$$

$$a(\boldsymbol{\sigma},\tilde{\Pi}_h\boldsymbol{\tau}) - \overline{a}_q(\boldsymbol{\sigma}_h,\tilde{\Pi}_h\boldsymbol{\tau}) + (b-\overline{b})(\tilde{\Pi}_h\boldsymbol{\tau},u) + \overline{b}(\tilde{\Pi}_h\boldsymbol{\tau}-\boldsymbol{\tau},P_hu-u_h) + (\overline{b}-b)(\boldsymbol{\tau},P_hu-u_h)\ .$$

Proof.
We start by replacing b by $\overline{b}$,

$$(\operatorname{div}\boldsymbol{\tau},P_hu-u_h) - b(\boldsymbol{\tau},P_hu-u_h) =$$

$$(\operatorname{div}\boldsymbol{\tau},P_hu-u_h) - \overline{b}(\boldsymbol{\tau},P_hu-u_h) + (\overline{b}-b)(\boldsymbol{\tau},P_hu-u_h)\ .$$

We use (20a) and (20c) to get,

$$(\operatorname{div}\boldsymbol{\tau},P_hu-u_h) - b(\boldsymbol{\tau},P_hu-u_h) =$$

$$(\operatorname{div}\tilde{\Pi}_h\boldsymbol{\tau},P_hu-u_h) - \overline{b}(\tilde{\Pi}_h\boldsymbol{\tau},P_hu-u_h) + \overline{b}(\tilde{\Pi}_h\boldsymbol{\tau}-\boldsymbol{\tau},P_hu-u_h) + (\overline{b}-b)(\boldsymbol{\tau},P_hu-u_h)\ .$$

We use lemma 10 to find,

$$(\operatorname{div}\boldsymbol{\tau},P_hu-u_h) - b(\boldsymbol{\tau},P_hu-u_h) =$$
$$(\operatorname{div}\tilde{\Pi}_h\boldsymbol{\tau},u-u_h) - \bar{b}(\tilde{\Pi}_h\boldsymbol{\tau},u-u_h) + \bar{b}(\tilde{\Pi}_h\boldsymbol{\tau}-\boldsymbol{\tau},P_hu-u_h) + (\bar{b}-b)(\boldsymbol{\tau},P_hu-u_h)\,.$$

We use equation (6a) and equation (19a),

$$(\operatorname{div}\boldsymbol{\tau},P_hu-u_h) - b(\boldsymbol{\tau},P_hu-u_h) =$$
$$a(\boldsymbol{\sigma},\tilde{\Pi}_h\boldsymbol{\tau}) - \;< g,\tilde{\Pi}_h\boldsymbol{\tau}\cdot\mathbf{n}_{\partial\Omega} >\; - \bar{a}_q(\boldsymbol{\sigma}_h,\tilde{\Pi}_h\boldsymbol{\tau}) +$$
$$< g,\tilde{\Pi}_h\boldsymbol{\tau}\cdot\mathbf{n}_{\partial\Omega} >\; + (b-\bar{b})(\tilde{\Pi}_h\tau,u) +$$
$$\bar{b}(\tilde{\Pi}_h\boldsymbol{\tau}-\boldsymbol{\tau},P_hu-u_h) + (\bar{b}-b)(\boldsymbol{\tau},P_hu-u_h)\,.$$

□

Lemma 5.13.
If $(\boldsymbol{\sigma},u)$ is the solution of (6), $(\boldsymbol{\sigma}_h,u_h)$ is a solution of (19) and $(\boldsymbol{\tau},w)$ is the solution of the adjoint problem for an arbitrary right hand side $p \in \mathrm{L}^2(\Omega)$, then

$$c(P_hw,u-u_h) = -a(\tau,\tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h)) + \bar{b}(\tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h),w-P_hw)\,.$$

Proof.
According to lemma 11,

$$c(P_hw,u-u_h) = -(\operatorname{div}(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h),P_hw)\,,$$

according to (20b) and (20c) we can rewrite the right hand side,

$$c(P_hw,u-u_h) = -(\operatorname{div}\tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h),P_hw)\,.$$

We wish to use equation (26) from lemma 10 to remove P_h. To do this we must add and subtract a term $\bar{b}(\tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h),P_hw)$ on the right hand side of our equation. We apply lemma 10 and gather terms in $\bar{b}$ together,

$$c(P_hw,u-u_h) = -(\operatorname{div}\tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h),w) + \bar{b}(\tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h),w-P_hw)\,.$$

Finally, we use (11a),

$$c(P_hw,u-u_h) = -a(\tau,\tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h)) + \bar{b}(\tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h),w-P_hw)\,.$$

□

Theorem 5.2.
Assume the adjoint problem (11) has a unique solution for all square integrable right hand sides and assume that there is a constant C_r such that, if $(\boldsymbol{\tau},w)$ is the solution of (11) for a given right hand side f, then

$$\|\boldsymbol{\tau}\|_{\mathbf{H}^1(\Omega)} + \|w\|_{\mathrm{H}^1(\Omega)} \leqslant C_r\|f\|_{\mathrm{L}^2(\Omega)}\,.$$

Now, if $(\boldsymbol{\sigma},u) \in \mathbf{H}^1(\Omega)\times \mathrm{H}^2(\Omega)$ is the solution of (6), and $(\boldsymbol{\sigma}_h,u_h)$ is a solution of (19) then there are constants

$$C,D,E \in (0,4C_r(1+2h_{\max})\max(\|\alpha\|_{W_1^\infty(\Omega)}, \|\boldsymbol{\beta}\|_{\mathbf{W}_1^\infty(\Omega)}, \|\gamma\|_{W_1^\infty(\Omega)})\,] ,$$

such that

$$\|P_h u - u_h\|_{L^2(\Omega)} \leqslant$$
$$C h_{\max}(\|u\|_{L^2(\Omega)} + \|\boldsymbol{\sigma}_{H^1(\Omega)}) + D h_{\max}\|\tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h)\|_h + E h_{\max}\|P_h u - u_h\|_{L^2(\Omega)} .$$

Proof.
If we have an estimate for $(P_h u - u_h, p)$ for all $p \in L^2(\Omega)$, then we can use

$$\|t\|_{L^2(\Omega)} = \sup_{p \in L^2(\Omega), p \neq 0} \frac{(p,t)}{\|p\|_{L^2(\Omega)}} ,$$

to find $\|P_h u - u_h\|_{L^2(\Omega)}$. We use the regularity of the adjoint problem (11) to find a solution $(\boldsymbol{\tau},w) \in \mathbf{H}^1(\Omega) \times L^2(\Omega)$ of (11) for a given right hand side $p \in L^2(\Omega)$. We may write,

$$(p, P_h u - u_h) = (\operatorname{div}\boldsymbol{\tau}, P_h u - u_h) - b(\boldsymbol{\tau}, P_h u - u_h) + c(w, P_h u - u_h) .$$

If we apply lemma 12, we find,

$$(p, P_h u - u_h) =$$
$$a(\boldsymbol{\sigma},\tilde{\Pi}_h\boldsymbol{\tau}) - \bar{a}_q(\boldsymbol{\sigma}_h,\tilde{\Pi}_h\boldsymbol{\tau}) + (b-\bar{b})(\tilde{\Pi}_h\boldsymbol{\tau},u) + \bar{b}(\tilde{\Pi}_h\boldsymbol{\tau}-\boldsymbol{\tau}, P_h u - u_h) +$$
$$(\bar{b}-b)(\boldsymbol{\tau}, P_h u - u_h) + c(w - P_h w, P_h u - u_h) + c(P_h w, P_h u - u_h) .$$

We use lemma 13,

$$(p, P_h u - u_h) =$$
$$a(\boldsymbol{\sigma},\tilde{\Pi}_h\boldsymbol{\tau}) - \bar{a}_q(\boldsymbol{\sigma}_h,\tilde{\Pi}_h\boldsymbol{\tau}) +$$
$$(b-\bar{b})(\tilde{\Pi}_h\tau,u) + \bar{b}(\tilde{\Pi}_h\boldsymbol{\tau}-\boldsymbol{\tau}, P_h u - u_h) + (\bar{b}-b)(\boldsymbol{\tau}, P_h u - u_h) +$$
$$c(w - P_h w, P_h u - u_h) - a(\boldsymbol{\tau},\tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h)) + \bar{b}(\tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h), w - P_h w) .$$

We can write this as follows,

$$(p, P_h u - u_h) =$$
$$(a-\bar{a}_q)(\boldsymbol{\sigma},\tilde{\Pi}_h\boldsymbol{\tau}) + \bar{a}_q(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h,\tilde{\Pi}_h\boldsymbol{\tau}) +$$
$$(b-\bar{b})(\tilde{\Pi}_h\boldsymbol{\tau},u) + \bar{b}(\tilde{\Pi}_h\boldsymbol{\tau}-\boldsymbol{\tau}, P_h u - u_h) + (\bar{b}-b)(\boldsymbol{\tau}, P_h u - u_h) +$$
$$c(w - P_h w, P_h u - u_h) - a(\boldsymbol{\tau},\tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h)) + \bar{b}(\tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h), w - P_h w) .$$

We use (24a) to write this as,

$$(p, P_h u - u_h) =$$
$$(a-\bar{a}_q)(\boldsymbol{\sigma},\tilde{\Pi}_h\boldsymbol{\tau}) - (a-\bar{a}_q)(\boldsymbol{\tau},\tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h)) +$$
$$(b-\bar{b})(\tilde{\Pi}_h\boldsymbol{\tau},u) + \bar{b}(\tilde{\Pi}_h\boldsymbol{\tau}-\boldsymbol{\tau}, P_h u - u_h) + (\bar{b}-b)(\boldsymbol{\tau}, P_h u - u_h) +$$
$$c(w - P_h w, P_h u - u_h) + \bar{b}(\tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h), w - P_h w) .$$

We can use the regularity of the adjoint problem (11), lemma 7, 8 and 9 and the projection error estimates (16a,b,c), to obtain

$$\| P_h u - u_h \|_{L^2(\Omega)} \leqslant$$

$$C_r(1+2h_{\max})2h_{\max} \| \alpha \|_{W_1^\infty(\Omega)} \left[\| \boldsymbol{\sigma} \|_{\mathbf{H}^1(\Omega)} + \| \tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h) \|_h \right] +$$

$$4C_r h_{\max} \| \boldsymbol{\beta} \|_{\mathbf{W}_1^\infty(\Omega)}(1+2h_{\max}) \| u \|_{L^2(\Omega)} + 2C_r h_{\max} \| \boldsymbol{\beta} \|_{\mathbf{L}^\infty(\Omega)} \| P_h u - u_h \|_{L^2(\Omega)} +$$

$$2C_r h_{\max} \left[h_{\max} \| \gamma \|_{W_1^\infty(\Omega)} \| P_h u - u_h \|_{L^2(\Omega)} + \| \boldsymbol{\beta} \|_{L^\infty(\Omega)} \| \tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h) \|_{L^2(\Omega)} \right] .$$

This can be written as,

$$\| P_h u - u_h \|_{L^2(\Omega)} \leqslant$$

$$C h_{\max} + D h_{\max} \| \tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h) \|_h + E h_{\max} \| P_h u - u_h \|_{L^2(\Omega)} .$$

□

If $h_{\max}$ is small enough, theorem 1 and theorem 2 together give an $\mathcal{O}(h_{\max})$ error estimate. An important limit on $h_{\max}$ is implied by the form of the estimates in theorem 1 and 2. The main problem is, that large values of $\| \alpha \|_{W_1^\infty(\Omega)}$, $\| \boldsymbol{\beta} \|_{\mathbf{W}_1^\infty(\Omega)}$ and $\| \gamma \|_{W_1^\infty(\Omega)}$ decrease the range of $h_{\max}$ for which the estimate is valid. This problem can be avoided if we make an extra assumption. We discuss this in the next section.

5.6.3. A different approach.

To improve our estimate of $\| P_h u - u_h \|_{L^2(\Omega)}$, we consider the adjoint of the discrete problem. This means, that we look for $(\boldsymbol{\tau}_h, v_h) \in X_h \times W_h$, such that

$$\bar{a}_q(\boldsymbol{\tau}_h, \boldsymbol{\sigma}_h) - (\operatorname{div} \boldsymbol{\sigma}_h, v_h) = 0 \quad \forall \ \boldsymbol{\sigma}_h \in V_h , \tag{5.29a}$$

$$(\operatorname{div} \boldsymbol{\tau}_h, t_h) - \bar{b}(\boldsymbol{\tau}_h, t_h) + \bar{c}(v_h, t_h) = (f, t_h) \quad \forall \ t_h \in W_h . \tag{5.29b}$$

We call this system regular, if there is at least one solution for each $f \in P_h(\mathrm{L}^2(\Omega))$, and that all solutions for a particular f satisfy

$$\| \boldsymbol{\tau}_h \|_h + \| v_h \|_{L^2(\Omega)} \leqslant C \| P_h f \|_{L^2(\Omega)} , \tag{5.29c}$$

with C independent of the mesh size. This is a somewhat less stringent regularity condition than that given for the continuous adjoint problem (10). Note, that $\boldsymbol{\tau}_h \in X_h$, so $\boldsymbol{\tau}_{h,i}$ is a piecewise exponential function on Ω_k for $i = 1,2$.

An example of a general condition under which this system is regular is the following:

$$\alpha \geqslant A > 0 , \gamma \geqslant C_0 > 0 \text{ and } AC_0 - \| \boldsymbol{\beta} \|^2_{\mathbf{L}^\infty(\Omega)} \geqslant C_1 > 0 . \tag{5.30}$$

To show this, we need the following relations,

$$\int_\Omega \frac{P_h(\alpha)}{4} \boldsymbol{\tau}_h \cdot \boldsymbol{\tau}_h - \mathbf{P}_h(\boldsymbol{\beta}) \cdot \boldsymbol{\tau}_h v_h + P_h(\gamma) v_h v_h \, d\mu = \tag{5.31}$$

$$\int_\Omega \frac{P_h(\alpha)}{4}\left[\boldsymbol{\tau}_h - \frac{2\mathbf{P}_h(\boldsymbol{\beta})}{P_h(\alpha)}v_h\right]^2 + \left[P_h(\gamma) - \frac{\mathbf{P}_h(\boldsymbol{\beta})^2}{P_h(\alpha)}\right]v_h v_h \, d\mu \geqslant \tag{5.31a}$$

$$\int_\Omega P_h(\gamma)\left[v_h - \frac{\mathbf{P}_h(\boldsymbol{\beta})\cdot\boldsymbol{\tau}_h}{2P_h(\gamma)}\right]^2 + \left[P_h(\alpha) - \frac{\mathbf{P}_h(\boldsymbol{\beta})^2}{P_h(\gamma)}\right]\frac{\boldsymbol{\tau}_h\cdot\boldsymbol{\tau}_h}{4}\, d\mu\,. \tag{5.31b}$$

We know, that $(\operatorname{div}\tilde{\Pi}_h\boldsymbol{\sigma},P_h t) = (\operatorname{div}\Pi_h\boldsymbol{\sigma},P_h t)$, so, if we take the sum of (29a) and (29b) with $\boldsymbol{\sigma} = \Pi_h\boldsymbol{\tau}_h$ and $t = v_h$, we find

$$\bar{a}_q(\boldsymbol{\tau}_h,\Pi_h\boldsymbol{\tau}_h) - \bar{b}(\boldsymbol{\tau}_h,v_h) + \bar{c}(v_h,v_h) = (f,v_h)\,. \tag{5.32}$$

According to (24a), $\bar{a}_q(\boldsymbol{\tau}_h,\Pi_h\boldsymbol{\tau}_h) = \bar{a}_q(\tilde{\Pi}_h\boldsymbol{\tau}_h,\tilde{\Pi}_h\boldsymbol{\tau}_h)$, and by (24b) we have

$$\frac{1}{4}\bar{a}(\tilde{\Pi}_h\boldsymbol{\sigma},\tilde{\Pi}_h\boldsymbol{\sigma}) \leqslant \bar{a}_q(\tilde{\Pi}_h\boldsymbol{\sigma},\tilde{\Pi}_h\boldsymbol{\sigma})\,.$$

Hence

$$\int_\Omega \frac{P_h(\alpha)}{4}\boldsymbol{\tau}_h\cdot\boldsymbol{\tau}_h - \mathbf{P}_h(\boldsymbol{\beta})\cdot\boldsymbol{\tau}_h v_h + P_h(\gamma)v_h v_h \, d\mu \leqslant \int_\Omega P_h(f)v_h\, d\mu\,. \tag{5.33}$$

This expression is identical to (31), so (31a) is smaller than (f,v_h), combined with (30) this implies

$$\frac{C_1}{A}\|v_h\|_{L^2(\Omega)} \leqslant \|f\|_{L^2(\Omega)}\,. \tag{5.34a}$$

In the same way, we find, that (31b) is smaller than (f,v_h), together with (30) and (34a), this implies

$$\frac{C_1}{(AC_0)^{1/2}}\|\boldsymbol{\tau}_h\|_{\mathbf{L}^2(\Omega)} \leqslant \|f\|_{L^2(\Omega)}\,. \tag{5.34b}$$

From (32) we see that this implies,

$$A\|\boldsymbol{\tau}_h\|_h^2 \leqslant a_q(\boldsymbol{\tau}_h,\boldsymbol{\tau}_h) \leqslant$$

$$\|f\|_{L^2(\Omega)}\|v_h\|_{L^2(\Omega)} + \|\boldsymbol{\beta}\|_{\mathbf{L}^\infty(\Omega)}\|\boldsymbol{\tau}_h\|_{\mathbf{L}^2(\Omega)}\|v_h\|_{L^2(\Omega)} + \|\gamma\|_{L^\infty(\Omega)}\|v_h\|^2_{L^2(\Omega)}\,,$$

this implies that there is a C such that

$$\|\boldsymbol{\tau}_h\|_h \leqslant C\|f\|_{L^2(\Omega)}\,.$$

Theorem 5.3.
If we assume, that (29c) holds, then

$$\|P_h u - u_h\|_{L^2(\Omega)} \leqslant \tag{5.35}$$

$$h_{\max}\left[6\|\alpha\|_{W_1^\infty(\Omega)}\|\boldsymbol{\sigma}\|_{\mathbf{H}^1(\Omega)} + \; + 2(\|\boldsymbol{\beta}\|_{\mathbf{W}_1^\infty(\Omega)} + \|\gamma\|_{W_1^\infty(\Omega)})\|u\|_{L^2(\Omega)}\right]\,.$$

Proof.
We use (29b),

$$(P_h u - u_h,P_h f) = (\operatorname{div}\boldsymbol{\tau}_h,P_h u - u_h) - \bar{b}(\boldsymbol{\tau}_h,P_h u - u_h) + \bar{c}(P_h u - u_h,v_h)\,.$$

Hence, according to lemma 10 and the definition of $\overline{c}$,

$$(P_h u - u_h, P_h f) = (\operatorname{div}\boldsymbol{\tau}_h, u - u_h) - \overline{b}(\boldsymbol{\tau}_h, u - u_h) + \overline{c}(u - u_h, v_h) .$$

We use (6a) and (19a) to find

$$\begin{gathered}(P_h u - u_h, P_h f) = \\ (\operatorname{div}\boldsymbol{\tau}_h, u - u_h) - (\overline{b} - b)(\boldsymbol{\tau}_h, u) - b(\boldsymbol{\tau}_h, u) + \overline{b}(\boldsymbol{\tau}_h, u_h) + \overline{c}(u - u_h, v_h) = \\ (\overline{b} - b)(\boldsymbol{\tau}_h, u) + a(\boldsymbol{\sigma}, \boldsymbol{\tau}_h) - \overline{a}_q(\boldsymbol{\sigma}_h, \boldsymbol{\tau}_h) + (\overline{c} - c)(u, v_h) + c(u - u_h, v_h) .\end{gathered}$$

According to (24a) and lemma 11, this implies

$$\begin{gathered}(P_h u - u_h, P_h f) = \\ (\overline{b} - b)(\boldsymbol{\tau}_h, u) + (a - \overline{a}_q)(\boldsymbol{\sigma}, \boldsymbol{\tau}_h) + \overline{a}_q(\Pi_h \boldsymbol{\sigma} - \boldsymbol{\sigma}_h, \boldsymbol{\tau}_h) \\ + (\overline{c} - c)(u, v_h) - (\operatorname{div}(\Pi_h \boldsymbol{\sigma} - \boldsymbol{\sigma}_h), v_h) .\end{gathered}$$

Now, (29a) implies,

$$(P_h u - u_h, P_h f) = (\overline{b} - b)(\boldsymbol{\tau}_h, u) + (a - \overline{a}_q)(\boldsymbol{\sigma}, \boldsymbol{\tau}_h) + (\overline{c} - c)(u, v_h) .$$

Finally, we use lemma 7, 8, and 9 and (29c) to obtain our error estimate (35).
□

5.7 A verification of the local maximum principle.

We use the discrete adjoint problem to show that, for this quadrature rule, the matrix after elimination of $\boldsymbol{\sigma}$ by static condensation is an M-matrix. The discrete adjoint problem is defined in (29).

We assume a regular uniform mesh. We denote the matrix corresponding to (29), after elimination of $\boldsymbol{\sigma}_h$, by A. We see, that the matrix A has non-positive off-diagonal elements. We shall show, that A is an M-matrix. To do this, we use theorem 5.12, chapter 5, page 124 of [23]. This theorem states, that, for irreducible matrices with non-positive off-diagonal elements, the M-matrix property is equivalent to the existence of a positive vector with a non-negative image, that is not identically zero. In our case, the vector $(1,1, \ldots, 1)^T$ has a such an image, because all row sums are non-negative, and any row corresponding to an edge or corner has a positive row-sum.

The fact, that the matrix A is irreducible follows from theorem 3.6, [23] which states that, for a square matrix, irreducibility is equivalent to its di-graph being strongly connected. Inspection shows, that the di-graph of the matrix under consideration is indeed strongly connected.

According to theorem 5.6 [23], A^T is an M-matrix too. This implies, that the discrete equations for the original u_h satisfy a local maximum principle.

The M-matrix property implies that the system for u_h has a unique solution. From the form of the equations for $\boldsymbol{\sigma}_h$, we see that a given u_h induces a unique $\boldsymbol{\sigma}_h$. this implies that our system is always uniquely solvable. A quick calculation of the coefficients of u_h in (19a) shows that, for constant coefficients and large $\boldsymbol{\beta}$, i.e. with large convection diffusion ratios, we get a relation between σ_h and u_h where the "upwind" point is weighed more heavily. If $\boldsymbol{\beta}/\alpha$ remains bounded and we go to the limit $|\beta_1|+|\beta_2|\rightarrow\infty$ then we get a first order upwind scheme. This suggests that the scheme, in which the coefficients are continuously dependent on this ratio, remains useful close to such a limit.

5.8 An a-posteriori estimator.

We use a special quadrature rule and obtain a higher order discretisation. We seek an $\overline{a}_{h,3}(.,.)$. that minimises $\overline{a}-\overline{a}_{h,3}$. To do this, we choose a special quadrature rule for each $\overline{a}(.,\boldsymbol{\eta})$, where $\boldsymbol{\eta}$ is one of the basis functions introduced earlier. Due to the nature of our test functions, the quadrature rule is essentially a one-dimensional rule.

5.8.1. The derivation of the quadrature rule.

For $\boldsymbol{\eta}_{i,j+½}$ we proceed as follows. We replace the two dimensional integral by a repeated integral, we integrate exactly in the $\mathbf{e}_2$ direction and then use a three point rule to approximate the remaining integral. As nodes for the last integration we take either the centres of $\Gamma_{i-½,j+½,0}$, $\Gamma_{i+½,j+½,0}$ and $\Gamma_{i+½,j+½,1}$. Or, if we are at a boundary, the edge centre on the boundary and the two next closest edge centres. We choose the weights as follows,

$$\overline{a}_{h,3}(\Pi_h\boldsymbol{\sigma},\boldsymbol{\eta}_{i,j+½}) = \overline{a}(\boldsymbol{\sigma},\boldsymbol{\eta}_{i,j+½}) ,$$

for all $\boldsymbol{\sigma}$ with x_1-components that are second order polynomials in x_1. I.e. for all $p,q,r \in \mathbb{R}$, and all $\boldsymbol{\eta}_{i,j+½}$, we have

$$\overline{a}_{h,3}(\Pi_h((px_1^2+qx_1+r)\mathbf{e}_1),\boldsymbol{\eta}_{i,j+½}) = \overline{a}((px_1^2+qx_1+r)\mathbf{e}_1,\boldsymbol{\eta}_{i,j+½}) ,$$

In a similar manner, we define the rule for $\boldsymbol{\eta}_{i+½,j}$.

5.8.2. An estimator for the local discretisation error and a lower bound for the global error.

If we assume that $c=\overline{c}$, $b=\overline{b}$ and $a=\overline{a}$, then we can use this rule to obtain an a-posteriori estimator for the local discretisation error and a lower bound for the global error as follows. It is immediately obvious, that

$$\overline{a}_{h,3}(\boldsymbol{\sigma},\boldsymbol{\eta}_r) - \overline{a}_{h,1}(\boldsymbol{\sigma},\boldsymbol{\eta}_r) \geqslant \mathcal{O}(h_{\max}^2) ,$$

where r is a possible index-tuple. Moreover,

$$\overline{a}(\boldsymbol{\sigma},\boldsymbol{\eta}_r) - \overline{a}_{h,3}(\boldsymbol{\sigma},\boldsymbol{\eta}_r) = \mathcal{O}(h_{\max}^3) ,$$

if $\boldsymbol{\sigma}$ is smooth enough. If

$$|\overline{a}_{h,3}(\boldsymbol{\rho}_h,\boldsymbol{\eta}_r) - (\operatorname{div}\boldsymbol{\eta}_r - \boldsymbol{\beta}\cdot\boldsymbol{\eta}_r, w_h)| \geqslant K ,$$

then we have either

$$\|w_h\|_{L^\infty(\Omega)} \geqslant C_1 K ,$$

or

$$\| \boldsymbol{\rho}_h \|_{L^\infty(\Omega)} \geqslant C_2 K \ ,$$

We see immediately that, if $(\boldsymbol{\sigma}_h, u_h)$ is the solution of (19) with $\overline{a}_q = \overline{a}_{h,1}$ then

$$\overline{a}_{h,1}(\Pi_h \boldsymbol{\sigma} - \boldsymbol{\sigma}_h, \boldsymbol{\eta}_r) - ((\operatorname{div} - \boldsymbol{\beta})\boldsymbol{\eta}_r, P_h u - u_h) = \mathcal{O}(h^k) \ ,$$

with $k = 1$ or 2 depending on the coefficients in (1) and

$$\overline{a}_{h,3}(\Pi_h \boldsymbol{\sigma} - \boldsymbol{\sigma}_h, \boldsymbol{\eta}_r) - ((\operatorname{div} - \boldsymbol{\beta})\boldsymbol{\eta}_r, P_h u - u_h) = \mathcal{O}(h_{\max}^{k+2}) + \overline{a}_{h,1}(\boldsymbol{\sigma}_h, \boldsymbol{\eta}_r) - \overline{a}_{h,3}(\boldsymbol{\sigma}_h, \boldsymbol{\eta}_r).$$

So, $(a_{h,1} - a_{h,3})(\boldsymbol{\sigma}_h, \boldsymbol{\eta}_r)$ is an estimate for the local discretisation error. Moreover this implies, that there is a constant C such that

$$\| \Pi_h \boldsymbol{\sigma} - \boldsymbol{\sigma}_h \|_{L^\infty(\Omega)} + \| P_h u - u_h \|_{L^\infty(\Omega)} \geqslant C \, | \overline{a}_{h,1}(\boldsymbol{\sigma}_h, \boldsymbol{\eta}_r) - \overline{a}_{h,3}(\boldsymbol{\sigma}_h, \boldsymbol{\eta}_r) | + \mathcal{O}(h_{\max}^{k+2}).$$

If we assume that

$$\| \Pi_h \boldsymbol{\sigma} - \boldsymbol{\sigma}_h \|_{L^\infty(\Omega)} + \| P_h u - u_h \|_{L^\infty(\Omega)} = \mathcal{O}(h_{\max}^{k}) \ ,$$

we see that, for $h_{\max}$ small enough,

$$\| \Pi_h \boldsymbol{\sigma} - \boldsymbol{\sigma}_h \|_{L^\infty(\Omega)} + \| P_h u - u_h \|_{L^\infty(\Omega)} \geqslant \tfrac{1}{2} C \, | \overline{a}_{h,1}(\boldsymbol{\sigma}_h, \boldsymbol{\eta}_r) - \overline{a}_{h,3}(\boldsymbol{\sigma}_h, \boldsymbol{\eta}_r) | \ .$$

This provides a lower bound on the global discretisation error. We expect the solution for $\overline{a}_{h,3}$ to be two orders of magnitude more accurate than the solution for $\overline{a}_{h,1}$.

5.9 Numerical results.

We consider problem (1) with

$$u = \tanh(8(x_1 - x_2)) \ ,$$

$$\alpha = 100 \ , \ \beta_1 = \beta_2 = 100 \ ,$$

$$\Gamma_1 = \partial\Omega \ , \ g = u \,|_{\partial\Omega} \ ,$$

$$f = -\frac{\operatorname{div}(\operatorname{\mathbf{grad}} u + u\boldsymbol{\beta})}{\alpha} \ .$$

We find the following results for the two discretisations. The two components of the error vectors for the fluxes were identical up to the accuracy given. We use the norm described in section 4.8.2 of chapter 4.

the $\log_2$ of the errors for $\overline{a}_q = \overline{a}_{h,1}$,		
meshwidth	$\log_2 \| P_h u - u_h \|$	$\log_2 \| (\Pi_h \boldsymbol{\sigma} - \boldsymbol{\sigma}_h) \cdot \mathbf{e}_1 \|$
1 / 4	-1.5	-1.3
1 / 8	-1.9	-1.9
1 / 16	-2.6	-2.6
1 / 32	-3.8	-3.7
1 / 64	-5.4	-5.4

the $\log_2$ **of the errors for** $\overline{a}_q=\overline{a}_{h,3}$,		
meshwidth	$\log_2 \Vert P_h u - u_h \Vert$	$\log_2 \Vert (\Pi_h \boldsymbol{\sigma} - \boldsymbol{\sigma}_h)\cdot \mathbf{e}_1 \Vert$
1 / 4	-3.0	-3.1
1 / 8	-3.9	-4.5
1 / 16	-6.0	-6.7
1 / 32	-8.5	-10.0
1 / 64	-10.9	-13.7

We see that the order of convergence is indeed higher for the second method. we also see that the difference in order for the fluxes approaches 2. Deviations from the expected order may be caused by the steepness of the solution relative to the mesh.

5.10 Conclusions.

The Petrov Galerkin mixed finite element method with exponentially fitted test functions for the fluxes has several nice properties. For instance, just as for a finite volume method, if the true solution $\boldsymbol{\sigma}$ is divergence-free, then the same holds for $\boldsymbol{\sigma}_h$. Furthermore we have a formal a-priori error estimate, and after elimination of $\boldsymbol{\sigma}_h$ by static condensation the two dimensional discretisation results in an M-matrix for u_h. We can extend the method to three dimensions without additional difficulties. Section 5.9 suggest that the scheme with the three point quadrature rule $\overline{a}_{h,3}$ can serve as a source for a-posteriori error estimates. To judge the effectiveness of the method for singularly perturbed problems is very difficult. However the fact that it incorporates exponential fitting, copes well with the exponential solution of the constant coefficient case and approaches a two-dimensional upwind scheme if the convection goes to infinity suggests the method based on $\overline{a}_{h,1}$ can be applied to such problems.

References

1. D. L. Scharfetter and H. K. Gummel, “Large-Signal Analysis of a Silicon Read Diode Oscillator,” *IEEE Transactions on Electron Devices*, vol. ED-16, no. 1, pp. 64-77, 1969.
2. Siegfried Selberherr, *Analysis and simulation of semiconductor devices,* Springer-verlag, Wien New York, 1984.
3. Peter A. Markowich, *The Stationary Semiconductor Device Equations,* Springer-Verlag, Wien New York, 1986.
4. Wolfgang Fichtner, Donald J. Rose, and Randolph E. Bank, “Numerical methods for semiconductor device simulation,” *SIAM journal of scientific and statistical computing*, vol. 4, no. 3, pp. 416-435, 1983.
5. F. Brezzi, L. D. Marini, and P. Pietra, “Mixed exponential fitting schemes for current continuity equations,” in *Proceedings of the sixth international NASECODE conference*, ed. J. J. H. Miller, Boole Press Ltd,

1989.

6. P. W. Hemker, *A Numerical Study of Stiff Two-Point Boundary Problems,* Mathematical Centre Tracts, 80, Mathematical Centre, Amsterdam, 1977.

7. J. Douglas, Jr. and J. E. Roberts, "Global estimates for mixed methods for second order elliptic equations," *Mathematics of computation*, vol. 44, no. 169, pp. 39-52, 1985.

8. Robert A. Adams, *Sobolev Spaces,* Academic Press, 1975.

9. V. Girault and P. Raviart, *Finite Element Methods for Navier-Stokes Equations,* Springer series in computational mathematics, 5, Springer-Verlag, 1986.

10. S. J. Polak, C. den Heijer, H. A. Schilders, and P. Markowich, "Semiconductor device modelling from the numerical point of view," *International Journal for Numerical Methods in Engineering*, vol. 24, pp. 763-838, 1987.

11. Jean E. Roberts and Jean-Marie Thomas, "Mixed and Hybrid Finite Element Methods," RR 737, INRIA, Rocquencourt, October 1987.

12. P. A. Raviart and J. M. Thomas, "A mixed finite element method for 2-nd order elliptic problems," in *Mathematical aspects of the finite element method,* Lecture Notes in Mathematics, vol. 606, pp. 292-315, Springer, 1977.

13. J. C. Nedelec, "Mixed Finite Elements in $\mathbf{R}^3$," *Numer. Math.*, vol. 35, pp. 315-341, 1980.

14. P. G. Ciarlet and P. A. Raviart, "General Lagrange and Hermite Interpolation in $\mathbf{R}^n$ with Applications to Finite Element Methods.," *Arch. Rational Mech. Anal.*, vol. 46, pp. 177-199, 1972.

15. Eugene O'Riordan, "Singularly Perturbed Finite Element methods," *Numerische Mathematik*, vol. 44, pp. 425-434, Springer-Verlag, 1984.

16. Eugene O'Riordan and Martin Stynes, "A finite element method for a singularly perturbed boundary value problem in conservative form," in *BAIL III*, ed. J. J. H. Miller, pp. 271-275, Boole Press, Dublin, 1984.

17. Martin Stynes and Eugene O'Riordan, "A Finite Element Method for a Singular Perturbed Boundary Value Problem," *Numerische Mathematik*, vol. 50, pp. 1-15, Springer-Verlag, 1986.

18. Eugene O'Riordan and Martin Stynes, "A Uniformly Accurate Finite-Element Method for a Singularly Perturbed One-Dimensional Reaction-Diffusion Problem," *Mathematics of Computation*, vol. 47, no. 176, pp. 555-570, 1986.

19. Eugene O'Riordan and Martin Stynes, "An analysis of a Superconvergence Result for a Singularly Perturbed Boundary Value Problem," *Mathematics of Computation*, vol. 46, no. 173, pp. 81-92, 1986.

20. Eugene O'Riordan and Martin Stynes, "A uniform finite element method for a conservative singularly perturbed problem," *Journal of*

Computational and Applied Mathematics, vol. 18, pp. 163-174, 1987.

21. H.-J. Reinhardt, "A-Posteriori Error Analysis and Adaptive Finite Element Methods for Singularly Perturbed Convection-Diffusion Equations," *Math. Meth. in the Appl. Sci.*, vol. 4, pp. 529-548, 1982.

22. Eugene O'Riordan and Martin Stynes, "A globally uniformly convergent finite element method for a singularly perturbed elliptic problem in two dimensions," *Mathematics of Computation*, vol. 57, no. 195, pp. 47-62, 1991.

23. M. Fiedler, *Special matrices and their applications in numerical mathematics,* Martinus Nijhoff, Dordrecht, 1986.

Samenvatting

Enige aspecten van gemengde eindige elementen methodes voor halfgeleider simulatie.

Als uitgangspunt voor dit proefschrift dient de discretisatie van het stationaire drift-diffusie model voor de halfgeleider. Dit is het eenvoudigste model voor het gedrag van electronen en gaten in een al dan niet gedoteerde halfgeleider en het wordt beschreven door de Van Roosbroeck vergelijkingen. Dit proefschrift bestudeert de discretisering van de afzonderlijke vergelijkingen en de nauwkeurigheid van de discretisatie. De oplossingen worden benaderd in de laagste orde Raviart-Thomas ruimte voor rechthoeken. Hoofdstuk één bevat een korte inleiding over halfgeleiders en halfgeleider-modellering.

In hoofdstuk twee wordt een nieuwe variant op de gemengde eindige elementen methode voor een tweede orde elliptisch probleem besproken. Door het gebruik van een gepaste kwadratuur-regel voor de berekening van de coëfficiënten-matrix levert de methode een betere orde van benadering voor locale gemiddelden. Het verschil in orde tussen de nieuwe variant en de oorspronkelijke methode kan gebruikt worden om een a-posteriori foutschatting voor de oorspronkelijke methode te construeren.

Hoofdstuk drie levert foutschattingen voor een klasse van Petrov-Galerkin gemengde eindige elementen methodes voor de één-dimensionale convectie-diffusie vergelijking. We vinden een uniforme foutschatting voor de flux van de oplossing. Voor het verschil tussen de discrete benadering enerzijds, en een probleemafhankelijke projectie van de continue oplossing anderzijds, kan ook een uniforme afschatting worden afgeleid. De genoemde projectie is bijna gelijk aan de standaard $L^2(\Omega)$-projectie voor alle roostercellen waar de convectie en diffusie van dezelfde orde van grootte zijn, hetgeen aantoont dat een gelocaliseerde singuliere verstoring geen globale gevolgen heeft. Als hulpmiddel bij de hierboven beschreven analyse worden enige stellingen afgeleid over de regulariteit van de oplossing van het continue probleem.

Hoofdstuk vier heeft als doelstelling het afleiden van een a-posteriori foutschatting voor de Scharfetter-Gummel discretisatie van de continuïteits-vergelijkingen in het halfgeleider probleem. We gebruiken de methode van uitgestelde correcties (deferred corrections) om een a-posteriori foutschatting af te leiden. Het discretisatie schema blijkt stabiel en consistent te zijn.

In hoofdstuk vijf wordt een Petrov-Galerkin gemengde eindige elementen formulering van de continuïteits-vergelijkingen gegeven. Het hoofdstuk bevat foutschattingen voor de discretisering. De foutschattingen zijn in principe niet geldig voor het singulier gestoorde geval. Er zijn echter argumenten die aantonen dat de discretisatie voor het singulier gestoorde geval toch bruikbaar is. We ontwikkelen ook een discretisatie die een hogere orde van nauwkeurigheid biedt. Deze discretisatie wordt gebruikt om een a-posteriori schatter voor de locale discretisatie fout af te leiden. Tevens wordt een ondergrens afgeleid voor de globale discretisatie fout.

List of Symbols

$\|\| \cdot \|\|_{L^p(\Omega)}$	35
$\|\| \cdot \|\|_{Wkp}$	35
$\|\| \cdot \|\|_{k,p,\Omega}$	36
$\|\| \cdot \|\|_{p,\Omega}$	36
$A \Delta B$	42
$C^{\infty}(\Omega)$	58
$C_0^{\infty}(\Omega)$	58
χ_A	30,69
D_p , D_n	18
δ_{rs}	76
$\mathbf{e}_1$, $\mathbf{e}_2$	30,75
ϵ (for chapter one only)	18
$\mathbf{E}$	18f
$\mathscr{E}$	80
$\mathscr{E}_{\partial\Omega}$	80
ess. inf.	55
$\mathbf{\Phi}_n$, $\mathbf{\Phi}_p$	19,20
ϕ_n , ϕ_p	19
$\mathbf{J}_n$, $\mathbf{J}_p$	18ff
H(div;Ω)	31,74,103
$H^k(\Omega)$	36
$H^1(\Omega)$	31
$H^2(\Omega)$	103
$H^{3/2}(\Omega)$	103
k_B	15
$\mathscr{L}(V,W;\mathbb{R})$	114
$L^2(\Omega) = W^{0,2}(\Omega)$	
λ_D	18
λ (for chapter one only)	18
λ Lebesgue measure on $\mathbb{R}$	
μ Lebesgue measure on $\mathbb{R}^2$	
μ_n , μ_p	18
N_a , N_d (for chapter one only)	17
n_i (for chapter one only)	18
n , p	17ff
$\overline{n}$, $\overline{p}$	19
$f = \mathcal{O}(h)$ read: f is of order h	
P_h	30,63,105
Π_h	30,63,105
$\tilde{\Pi}_h$	106
$P[A]$	30,78,105
$\hat{P}_h$	64
$\hat{\Pi}_h$	64
ψ	18f
ψ	55
Q_h	79
q (for chapter one only)	18
R	18f
Σ	105
$S(x,y)$	5
Span	30,79
Supp.	42
T (for chapter one only)	15
U_T	18
$W^{k,p}(\Omega)$	35,56
$W_1^{\infty} = W^{1,\infty}(\Omega)$	103
$W_0^{k,p}(\Omega)$	58
$(X_h, \|\| \cdot \|\|_h)$	114

Name Index

Ashcroft 13,14
Bank 13,23,72f,87
Blakemore 13,14
Boltzmann 15
Bramble 35ff
Brezzi 22,54,102
Cartesian 29,74f,103f
Cauchy 38,44,58,112,114
Ciarlet 106

Couperus 28,47
Debye 18
Dirichlet 29,60,70,73,91ff
Douglas 29,54,78,102,104,106
Early 12
Einstein 18
Engl 13,73
Euclidean 39f,49,81
Fair 12
Fermi 15
Fortin 29,63
Fox 92
Fubini 35f
Galerkin 21,22,23,63f,106,125
Gerschgorin 46
Gilbarg 37,58
Girault 74,103
Green 31,56,59ff,66,110
Grisvard 84,95
Gummel 11,13,20,23,72ff,80,87,98,100,102
den Heijer 17
Hemker 102,110
Hilbert 21,35ff,74,78
Hölder 59,107
Il'in 20,23,72
Jobs 11
Jones 12
Kelvin 15
Kronecker 76
Laplace 35
Lebesgue 30,55f,78,80
MacNeal 87
Marini 54
Markowich 13,17,73,102
Mayers 92
Mermin 13,14
Meyers 12
Navier 21,100
Nedelec 106
Neumann 74,76,78
Nicolaides 22,23,72,82f,100
O'Riordan 54,110
Pauli 14
Petritz 12
Petrov 21,22,23,63f,106,125
Pietra 54
Polak 13,17,28,47,73
Poisson 21,83f
Raviart 22,28ff,48,51,54,63f,74,78f,103,106,108
Read 20
Reinhardt 54,110
Riesz 69
Roach 60f
Roberts 29,54,78,102,104ff
Robin 73
Scharfetter 11,13,20,23,72ff,80,87,98,100,102
Schechter 95
Schilders 17,28,47
Schrödinger 14

Schwartz 38,44,112,114
Selberherr 13,17,73,102
Shockley 16
Slotboom 19,20,54
Sobolev 35ff,57,103
Stokes 21,100
Stynes 54,110
Sze 12
Taylor 90,93
Thomas 22,28ff,48,51,54,63f,78f,105ff,108
Trudinger 37,58
Verfürth 21,100
Wang 13,14,16,17
Yosida 60f

Subject Index

A

absolute temperature 15
absolutely continuous 57ff
acceptor 15
adjoint problem 54,61,104,118f
approximation spaces 30

B

band model 14
Boltzmann constant 15
boundary conditions 21
 Dirichlet 29,60,73,95,98
 homogeneous 22,54,60
 Neumann 74,76
 mixed 73
 Robin 73

C

CAD-CAM 12
carrier transport equations 18
Cartesian coordinates 75,103
Cartesian product 29,74,103
Cauchy-Schwartz (inequality) 38,114
characteristic function 30,69,104
charge carrier densities 18,73
choice of variables 19
continuity equation 17,18,22,102
convection-diffusion equation 54
 symmetrised 72
convection-diffusion problems 23
crystal rectifier 12
currents 18

D

- Debye length 18
- deferred correction 23,72,92
- diode 16
- discretisation error 22,89
  - global 23,43,123
  - local 22,72,123
- discretisation scheme 22,31,54
  - Box 21
  - Finite Volume 21,72,74,80
  - Galerkin 22
  - Mixed Finite Elements 21,28,63,108
  - Petrov-Galerkin 21f,63,64,125
  - Scharfetter-Gummel 11,20,23,72,74,87,100,102
- donor 15
- doped 22,54
- doping 17,70
- drift-diffusion model 11
- dual
  - cell 75
  - mesh 74

E

- Einstein relations 18
- electric field 17,22,54,70,103
- electric potential (see also potential) 73
- electron 103
- elements
  - Brezzi-Douglas-Marini 54
  - Raviart-Thomas 27,28,30,54,78,108
- error estimates
  - a-posteriori 11,21,23,50f,72f,92,102,123
  - a-priori 11,21,56,67,115
  - for projections 23,106
  - for quadrature rules 35
- Euclidean norm 39f,49,81
- exponentially fitted 20
- exponential fitting 72,102

F

- Fermi energy 15
- fermion 14
- flux 21,56,59
- Fubini's theorem 36

G

Green's
formula 31
function 56,59ff
theorem 36,110

H

Hölder (inequality) 86,107
holes 15

I

IC 12
Il'in
finite difference scheme 23,72
indefinite integral 57
insulator 14
integrated circuit 12
integration formula 33,35
(see also:quadrature rule)
irreducible matrix 122

J

junction 16,22,54

K

Kelvin 15
Kronecker-δ 76

L

$\mathbf{L}^2(\Omega)$-bounded 111
$\mathbf{L}^2(\Omega)$-elliptic 111
Lagrange multipliers 54,108
Lebesgue integrable 56
local maximum principle 28,122
local mesh refinement 13,21

M

majority carrier 15
metal 14
MFEM 21
M-matrix 35,48,52,108,110,122
mobility 18
multigrid 21

N

norm		35,39,54,60

P

partition		29,74,104
Pauli exclusion principle		14
permittivity		18
potential		17,73
projection		30f,63ff,72,77ff,105ff
	error	23,106
	operator	21

Q

quadrature		21,22,28,32,73,79,86,108f,112,123
(see also integration formula)		
	one point	22f
	three point	22f,32
	trapezium	28,34
	trapezoidal	49f,52
quasi-Fermi potential		19

R

recombination		17,103
regular		104
regularity		120
Riesz representation theorem		69

S

saddle-point problem		72,79,89
scaling		18,103
Schrödinger wve equation		14
second order elliptic equation		22,28
semiconductor		11,14
	continuity equation	102
semiconductor device		11ff
	crystal rectifier	12
	diode	16
	MOSFET	17
	transistor	16
singular perturbation		54,70
	localised	54
singular perturbed		23
Slotboom variables		19
Sobolev embedding theorem		35f

Sobolev space 35,56
Span 30,79
stability 72
subdivision 29

T

Taylor expansion 90,93
Taylor series
(see Taylor expansion)
test space 21,64,108
thermal voltage 18,73
transistor 16
trial space 12,21,64,72,108

U

upwind scheme 23

V

variational formulation 21
VLSI 12

W

weak differentiability 56
weak mixed form 103,105

Z

zero order term 47,52