Application of finite element methods to the simulation of semiconductor devices

1999 Rep. Prog. Phys. 62 277

(http://iopscience.iop.org/0034-4885/62/3/001)

Rep. Prog. Phys. **62** (1999) 277–353. Printed in the UK PII: S0034-4885(99)54707-6

Application of finite element methods to the simulation of semiconductor devices

J J H Miller†, W H A Schilders‡ and S Wang§

† Department of Mathematics, Trinity College, Dublin, Republic of Ireland
‡ Philips Research Laboratories, Eindhoven, The Netherlands
§ School of Mathematics and Statistics, Curtin University of Technology, Perth, Australia

Received 3 February 1998

Abstract

In this paper a survey is presented of the use of finite element methods for the simulation of the behaviour of semiconductor devices. Both ordinary and mixed finite element methods are considered. We indicate how the various mathematical models of semiconductor device behaviour can be obtained from the Boltzmann transport equation and the appropriate closing relations. The drift-diffusion and hydrodynamic models are discussed in more detail. Some mathematical properties of the resulting nonlinear systems of partial differential equations are identified, and general considerations regarding their numerical approximations are discussed. Ordinary finite element methods of standard and non-standard type are introduced by means of one-dimensional illustrative examples. Both types of finite element method are then extended to two-dimensional problems and some practical issues regarding the corresponding discrete linear systems are discussed. The possibility of using special non-uniform fitted meshes is noted. Mixed finite element methods of standard and non-standard type are described for both one- and two-dimensional problems. The coefficient matrices of the linear systems corresponding to some methods of non-standard type are monotone. Ordinary and mixed finite element methods of both types are applied to the equations of the stationary drift-diffusion model in two dimensions. Some promising directions for future research are described.

0034-4885/99/030277+77$59.50

Contents

Glossary of symbols

Physical symbols

p	hole concentration
n	electron concentration
ψ	electrostatic potential
$\boldsymbol{J}_p$, $\boldsymbol{J}_n$	current densities
$\boldsymbol{E}$	electric field
$\boldsymbol{E}_p$, $\boldsymbol{E}_n$	effective fields
n_{int}	intrinsic carrier concentration
T, T_p, T_n	temperatures
T_0	lattice temperature
q	electronic charge
k	Boltzmann constant
R	recombination/generation rate
D	doping profile
D_p, D_n	diffusion constants
μ_p, μ_n	mobilities
Φ_p, Φ_n	Slotboom variables

Mathematical symbols

I	identity matrix
$\lvert\cdot\rvert$	absolute value, length or area
∇, ∇_x, ∇_v	gradient operators
div or $\nabla\cdot$	divergence operator
Δ	Laplace operator
$\mathbb{R}$	set of real numbers
Ω	solution domain
$\overline{\Omega}$	closure of Ω
Γ	boundary of Ω
Γ_1	part of Γ where Dirichlet boundary conditions are imposed
Γ_2	part of Γ where Neumann boundary conditions are imposed
$C^0(\overline{\Omega})$	space of continuous functions on $\overline{\Omega}$
$C^1(\overline{\Omega})$	space of continuously differentiable functions on $\overline{\Omega}$
$L^2(\Omega)$	space of square integrable functions on Ω
$H^1(\Omega)$	space of functions with square integrable derivatives on Ω
$H_{\text{div}}(\Omega)$	space of functions with square integrable divergences on Ω
$(\cdot,\cdot)$	inner product on $L^2(\Omega)$
$\lVert\cdot\rVert$	norm on $L^2(\Omega)$
$\lVert\cdot\rVert_1$	norm on $H^1(\Omega)$
ε	singular perturbation parameter
h	mesh parameter/maximum mesh size
U_h, V_h	scalar finite element spaces
S_h, T_h	vector valued finite element spaces

$RT(E)$	Raviart–Thomas finite element space
E_h	a decomposition of $\overline{\Omega}$
Ed_h	set of edges in E_h
E_i	the ith element of E_h
∂E_i	boundary of E_i
$\boldsymbol{x}_i$	the ith vertex in E_h
d_j	Dirichlet tile associated with x_j
∂d_j	boundary of d_j
γ_{ji}	undirected line segment $\partial d_i \cap \partial d_j$
b_{ji}	box associated with the edge connecting x_j and x_i
I_j	index set for the neighbours of $\boldsymbol{x}_j$
$B(\cdot)$	Bernoulli function
δ_{ij}	Kronecker delta

1. Introduction

In this paper a review is presented of the use of finite element methods for the modelling of semiconductor devices. The current state of the art is described and also some of the more promising research directions are outlined. The numerical modelling of the electromagnetic and thermal behaviour of semiconductor devices is a challenging task, which requires the solution of many difficult numerical problems. The mathematical models are coupled non-stationary systems of parabolic and elliptic partial differential equations in two and three space dimensions which involve exponential nonlinearities. The current flows are convection-dominated, which means that singular perturbation phenomena are present. The doping profiles are essentially discontinuous and can jump from large positive values to large negative values across an extremely thin layer.

The Van Roosbroek equations and the initial and boundary conditions for typical semiconductor device models are interesting examples of singular perturbation problems. The solutions of such problems exhibit both boundary and interior layers. These are small regions of the device in which some physical quantity, such as the electric field or current density, has a large gradient. Some work has been carried out on the singularly perturbed nature of these problems, for example [1], but much still remains to be done.

It is well known that standard numerical methods do not produce satisfactory approximations to the solutions of singular perturbation problems. Therefore, the standard finite element packages, commonly used for structural mechanics and other applications, are of no use for semiconductor device simulation and special techniques are necessary. One of the most popular ways to obtain approximate solutions to singular perturbation problems is to use upwind discretizations of the differential equations. This review concentrates on the non-standard methods that are required for the adequate solution of these problems. Both finite element methods and mixed fixed element methods are discussed. Such an approach leads to numerical solutions that are adequate in a number of cases, but recent fundamental results of Shishkin (see, for example, [2, ch 14]), show that to obtain numerical solutions of guaranteed accuracy it will be necessary to design the meshes much more carefully than in current practice. Indeed, for numerical solutions of guaranteed accuracy at all positions in a device, it is likely that it will be necessary to use both fitted operators and fitted meshes.

The structure of the paper is as follows. In section 1 the basic features of semiconductor devices are discussed and the fundamental role of dopants is explained. Different mathematical models are then derived, depending on the number of moments of the Boltzmann transport

equation that are used. The section ends with a summary of the resulting systems of equations. The mathematical features of these equations relevant to semiconductor device modelling are then discussed in section 2, and a review of appropriate numerical methods is given. The main ideas of such methods are introduced in sections 3 and 4. In section 5 these finite element methods are applied to some of the semiconductor device models discussed earlier. The paper ends with a summary in section 6.

1.1. Semiconductor devices

In an isolated atom, electrons can only occupy discrete energy levels, but if atoms are organized in a lattice, each of these discrete energy levels splits into a continuous band of energy levels due to atomic interactions. Separate energy bands may merge into a single band when the spacing between the atoms decreases. If the spacing is further reduced, single bands may split again into disjoint bands. These bands are then separated by a forbidden energy region in which there are no allowed energy levels. Examples of such disjoint bands are the valence band and the conduction band. The difference in the energy of these bands is termed the bandgap, and is denoted by E_g. The bandgap is an important parameter in solid state physics, since it allows us to classify solids into three important groups: insulators, semiconductors, and conductors. These are defined as follows:

- in an insulator, all energy levels in the valence band are occupied by electrons, whereas all energy levels in the conduction band are empty. The bandgap is rather large, meaning that a large amount of thermal energy or a strong electric field is needed to transfer electrons from the valence band to the conduction band. The valence electrons form strong bonds between neighbouring atoms, and there are no free electrons to participate in the conduction of current.
- in a conductor, the conduction band is either partially filled with electrons, or it overlaps the valence band so that there is no bandgap. In both cases electrons in the partially filled conduction band or electrons at the top of the valence band can be raised to a higher energy level by supplying thermal energy or by applying an electric field. Hence, current conduction is possible in conductors.
- in a semiconductor the energy levels in the valence band are completely filled with electrons, while the energy levels in the conduction band are empty. This is similar to the situation for an insulator, the difference being that the bandgap is much smaller for a semiconductor. In this situation the interatomic bonds are of moderate strength. At low temperatures the electrons are bound in their lattice, and are not available for conduction, but when the temperature is raised, thermal vibrations may break the interatomic bonds and free electrons result. This may also happen when an external electric field is applied. As a consequence, semiconductors may conduct electrical current.

To describe the behaviour of semiconductors, it is convenient to introduce the concept of a positively charged hole. When a negatively charged electron is raised from the valence band towards the conduction band it leaves a positively charged hole in the valence band. Such holes can contribute to the conduction of current. Conduction which results from an interchange of holes and electrons in the nearly filled valence band is referred to as hole conduction. If an electric field is applied, then both the electrons in the conduction band and the holes in the valence band gain kinetic energy, and current conduction results. In the models describing the behaviour of semiconductor devices, equations for both electron and hole current are used (cf section 2).

An intrinsic semiconductor (pure material) contains a negligible number of impurities compared with the number of thermally generated electrons and holes. For an intrinsic

semiconductor in equilibrium the electron concentration in the conduction band equals the hole concentration in the valence band. This concentration is termed the intrinsic carrier concentration, and is denoted by n_{int}. At room temperature, the electrical conductivity of a semiconductor is significantly higher than the conductivity of an insulator, and significantly lower than the conductivity of a conductor. These differences can be explained by comparing the intrinsic concentrations of the electrons in the conduction band of the three groups of solids. The most important semiconductor material is silicon, which has an intrinsic carrier concentration of order 10^{10} cm^{-3}. For a metal, the intrinsic carrier concentration is of order 10^{22} cm^{-3}, while for insulators it is of order 10^{3} cm^{-3}.

When the temperature is raised more electrons are forced into the conduction band, and more holes are left behind in the valence band. As a result the concentration of mobile carriers is increased, which causes higher conductivity at higher temperatures. This is in contrast to the situation for conductors, where the mobility of conduction electrons decreases with increasing temperature due to increasing lattice vibrations.

Conductivity properties may also be influenced by introducing a small number of impurity atoms into the semiconductor lattice. This process is referred to as doping, and the resulting material is said to be extrinsic. Two kinds of dopant atoms can be implanted, namely those producing one or more excess conduction electrons and those accepting electrons and thus producing holes. In the former case, the dopant is called a donor, whereas in the latter case it is called an acceptor. The resulting semiconductor material is said to be of n-type or p-type, respectively. By introducing impurity concentrations several orders of magnitude larger than the intrinsic carrier concentration n_{int}, the extrinsic material can be turned into a good conductor even at room temperature.

Semiconductor devices are formed by combining a number of different types of extrinsic semiconductor material. The precise configuration is determined by the doping process, which in current technology consists of hundreds of steps. Each of these steps is one of the following: epitaxy and crystal growth, ion implantation, diffusion, oxidation, lithography, etching, deposition. The order in which the steps are executed is crucial for the end product, and because of the sequential nature of the process, there is a strong interrelation between the various steps. This implies that an adequate physical description of these processing steps is needed so that the result of a simulation has a practical value. Fortunately, within certain limitations, such a description is available for most of the steps. The goal of process modelling is to provide such a description.

Nowadays many different types of semiconductor devices are in use. The most important basic element is the p–n junction, which is obtained when a piece of p-type material is brought into contact with a piece of n-type material. Under thermal equilibrium conditions a transition region is then formed, which is devoid of mobile carriers since the electrons and holes recombine on both sides of the junction. This transfer of charge across the junction destroys the neutrality of the p-type and n-type regions, since the fixed charges due to the ionized impurities are no longer balanced by the mobile charge carriers. The transition region, also called the space charge region, contains a double layer of charge which creates an electric field that eventually inhibits any further diffusion of carriers. Because of this double layer, there is a jump in the electrostatic potential across the transition region. The resulting potential at equilibrium is referred to as the built-in potential. For silicon, this potential varies between approximately -0.5 and 0.5 V. Additional carriers can cross the transition region only if an external potential is applied. If this applied potential has the opposite sign to the built-in potential, then current can flow again. In this case, the junction is said to be biased in the forward direction, or forward-biased. On the other hand, if the applied potential has the same sign as the built-in potential, then the potential difference is increased even further than in the

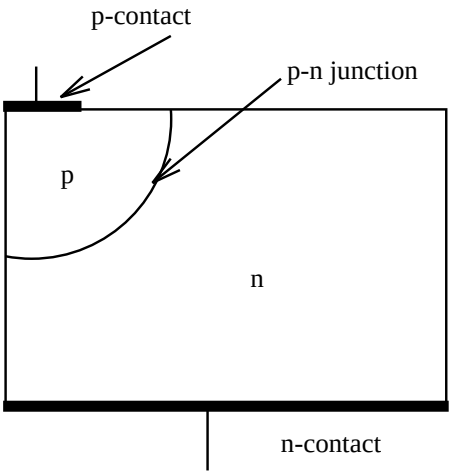

Figure 1. Configuration of a diode.

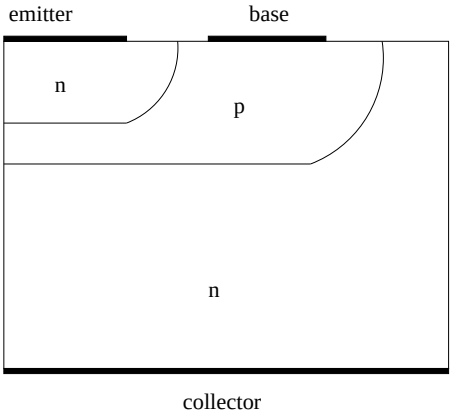

Figure 2. Configuration of a bipolar transistor.

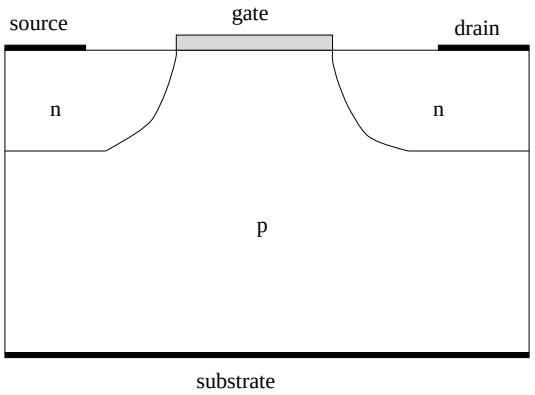

Figure 3. Configuration of an MOS transistor.

equilibrium situation. In this mode, which is referred to as reverse-biased, it is even more difficult for carriers to cross the transition region, and no current flows.

Another important element of semiconductor devices is the metallic contact. Just as in the case of the p–n junction, where two semiconductor materials are in contact with each other, a transition layer devoid of charge carriers is formed at the interface of the metallic contact and the semiconductor material. These metal–semiconductor interfaces must be taken into account when simulating the behaviour of semiconductor devices. Often, this is done by an appropriate choice of the boundary conditions at the contacts.

Most semiconductor devices consist of a number of p–n junctions and metallic contacts. The simplest such device is the diode in figure 1, which consists of a p-type region and an n-type region, with metallic contacts at both ends.

We can also make a structure with two p–n junctions, as shown in figure 2. The resulting device is termed a bipolar transistor. Another type of device consisting of two p–n junctions is the MOS transistor, shown schematically in figure 3. From this figure it is clear that the abbreviation MOS stands for metal, oxide and semiconductor. A thin insulating oxide layer,

usually SiO_2, separates the gate contact from the silicon surface.

Besides the three devices mentioned above, many others exist such as, for example, thyristors and charge-coupled devices. A useful reference for readers not familiar with semiconductor devices is the book by Frederiksen [3], which contains an easy-to-read introduction. The books by Sze [4, 5] are more advanced.

1.2. Derivation of the differential equations

To model the behaviour of semiconductor devices, it is necessary to have a physical or mathematical model. Most models in use are derived from the Boltzmann transport equation

$$\frac{\partial f}{\partial t} + \boldsymbol{v} \cdot \nabla_{\boldsymbol{x}} f + \frac{1}{m} \boldsymbol{F}_{\text{eff}}(\boldsymbol{x}, t) \cdot \nabla_{\boldsymbol{v}} f = Q(f, f). \tag{1}$$

Here ∇_x and ∇_v denote the gradient operators $(\partial/\partial x_1, \partial/\partial x_2, \partial/\partial x_3)$ and $(\partial/\partial v_1, \partial/\partial v_2, \partial/\partial v_3)$, $\boldsymbol{F}_{\text{eff}}$ is the effective force, f denotes the probability density function for one particle, and Q is the collision operator. Q has a rather complicated structure. Cercignani [6] gives a detailed discussion of this term in the gas dynamics case, whereas in [7] alternative forms are presented for the case of charged particles moving in a semiconductor. Roughly speaking, we have that

$$Q(f, f) = \int_{\mathbb{R}^3} \{P[(\boldsymbol{x}, \boldsymbol{v}', t) \to (\boldsymbol{x}, \boldsymbol{v}, t)] - P[(\boldsymbol{x}, \boldsymbol{v}, t) \to (\boldsymbol{x}, \boldsymbol{v}', t)]\} \, \mathrm{d}\boldsymbol{v}'. \tag{2}$$

In this expression, $P[(\boldsymbol{x}, \boldsymbol{v}_1, t) \to (\boldsymbol{x}, \boldsymbol{v}_2, t)]$ denotes the probability that the particle is scattered from the state $(\boldsymbol{x}, \boldsymbol{v}_1, t)$ into the state $(\boldsymbol{x}, \boldsymbol{v}_2, t)$. Stated differently, it denotes the probability that the particle velocity changes instantaneously from $\boldsymbol{v}_1$ to $\boldsymbol{v}_2$.

The Boltzmann transport equation (1) was originally derived in the context of gas dynamics. If we wish to use it to describe the movement of electrons and holes in semiconductors, several modifications have to be undertaken. In particular, the quantum behaviour of electrons and holes has to be taken into account. This leads to modifications of the differential operator of the Boltzmann transport equation and in the form of the collision term. Furthermore, transport equations for each type of charge carrier must be used. The interaction between these different types of particle can be included by introducing additional coupling terms. If these changes are made, the resulting system of equations describing the transport of electrons and holes in a semiconductor device is

$$\frac{\partial f_p}{\partial t} + \frac{1}{\hbar} \nabla_{\boldsymbol{k}} E_p \cdot \nabla_{\boldsymbol{x}} f_p + \frac{1}{\hbar} \boldsymbol{F}_{\text{eff,p}}(\boldsymbol{x}, t) \cdot \nabla_{\boldsymbol{k}} f_p = Q_p(f_p, f_p) + R_p(f_p, f_n), \tag{3}$$

$$\frac{\partial f_n}{\partial t} + \frac{1}{\hbar} \nabla_{\boldsymbol{k}} E_n \cdot \nabla_{\boldsymbol{x}} f_n + \frac{1}{\hbar} \boldsymbol{F}_{\text{eff,n}}(\boldsymbol{x}, t) \cdot \nabla_{\boldsymbol{k}} f_n = Q_n(f_n, f_n) + R_n(f_p, f_n), \tag{4}$$

where $\boldsymbol{k}$ denotes the wavevector and f_p and f_n are the probability density functions for the holes and electrons respectively. These equations serve as the basis for most models used in semiconductor device simulation. For a detailed derivation of these equations, the reader is referred to [7, 8]. In the first reference similar equations are given for the full quantum mechanical case.

A wide variety of mathematical methods has been developed to obtain approximate solutions of the Boltzmann transport equation. Some of these methods attempt a direct solution, a common feature being their time-consuming character. This is the case especially for Monte Carlo techniques. For this reason, most device modelling programs to date still use the classical method of moments to obtain approximate solutions. Its use is fairly widespread due to its relative simplicity and its utility for extracting useful information from the Boltzmann transport equation. This is a consequence of the fact that the moments of the

distribution function are related to physical quantities. The simplest example is the scalar zeroth-order moment, defined as

$$M^0(\boldsymbol{x}, t) = \int f(\boldsymbol{x}, \boldsymbol{k}, t)\, \mathrm{d}k_1\, \mathrm{d}k_2\, \mathrm{d}k_3, \tag{5}$$

which is easily recognized to be the concentration of the charged particles under consideration. The first-order moment, defined as the vector

$$\boldsymbol{M}^1(\boldsymbol{x}, t) = \int \boldsymbol{k} f(\boldsymbol{x}, \boldsymbol{k}, t)\, \mathrm{d}k_1\, \mathrm{d}k_2\, \mathrm{d}k_3, \tag{6}$$

is essentially the (electron or hole) current density $\boldsymbol{J}$, where

$$\boldsymbol{J}(\boldsymbol{x}, t) = -\frac{q\overline{h}}{m}\boldsymbol{M}^1(\boldsymbol{x}, t).$$

In a similar way physical interpretations can be assigned to higher-order moments. A general definition of the ith-order moment, for $i \geqslant 1$, is the tensor

$$\begin{aligned} \boldsymbol{M}^i_{q_1,\ldots,q_i}(\boldsymbol{x}, t) &= \int k_{q_1} \ldots k_{q_i} f(\boldsymbol{x}, \boldsymbol{k}, t)\, \mathrm{d}k_1\, \mathrm{d}k_2\, \mathrm{d}k_3, \\ & \quad q_j \in \{1, 2, 3\} \quad \text{for all} \quad 1 \leqslant j \leqslant i. \end{aligned} \tag{7}$$

Here, the integral extends over momentum space.

On multiplying the Boltzmann transport equation by $k_{q_1} \ldots k_{q_i}$ and integrating over momentum space the following equations are obtained for the components of the moments of different order

$$\begin{aligned} &\frac{\partial}{\partial t}\boldsymbol{M}^i_{q_1,\ldots,q_i} + \frac{\overline{h}}{m}\nabla_{\boldsymbol{x}} \cdot (\boldsymbol{M}^{i+1}_{q_1,\ldots,q_i,1}, \boldsymbol{M}^{i+1}_{q_1,\ldots,q_i,2}, \boldsymbol{M}^{i+1}_{q_1,\ldots,q_i,3})^T - \frac{1}{\overline{h}}\sum_{j=1}^{i} F_{q_j}\boldsymbol{M}^{i-1}_{(q_1,\ldots,q_i)|q_j} \\ &\qquad = \int_{\mathbb{R}^3} k_{q_1} \ldots k_{q_i} C(f). \end{aligned} \tag{8}$$

In this way, an infinite hierarchy is constructed in which the ith moment is directly related to the moments $\boldsymbol{M}^{i-1}$ and $\boldsymbol{M}^{i+1}$. That the latter higher-order moment plays a role in the evolution of $\boldsymbol{M}^i$ is unfortunate, and means that a closing relation has to be specified in order to be able to generate a finite number of equations of the form (8). This can be done in various ways and the complexity of the resulting model depends on the number of moments taken into account. A reasonable first step is to formulate a model based on a minimum number of moments. The resulting model, which is known as the drift-diffusion model, provides a useful description of the behaviour of semiconductor devices. The usefulness of the model is greatly enhanced by using sophisticated expressions for the recombination, mobility and bandgap narrowing parameters occurring in the equations.

The derivation of the drift-diffusion model requires the use of the first two moments of the Boltzmann transport equation (assuming a parabolic band structure). The resulting equations are the equation expressing the conservation of mass, often referred to as the continuity equation (for electrons),

$$\frac{\partial n}{\partial t} - \frac{1}{q}\nabla_{\boldsymbol{x}} \cdot \boldsymbol{J} = C_n, \tag{9}$$

and the equation expressing the conservation of momentum

$$\frac{\partial \boldsymbol{P}}{\partial t} + \boldsymbol{v}^M \nabla_{\boldsymbol{x}} \cdot \boldsymbol{P} + (\boldsymbol{P} \cdot \nabla_{\boldsymbol{x}})\boldsymbol{v}^M + \nabla_{\boldsymbol{x}} \cdot (nk\overline{\overline{\boldsymbol{T}}}) - n\boldsymbol{F} = \boldsymbol{C}_{\boldsymbol{P}}. \tag{10}$$

Here, $\boldsymbol{C}_{\boldsymbol{P}}$ represents the time rate of change of $\boldsymbol{P}$ due to collisions.

Equation (10) can be simplified by making use of the equation (9). It follows that

$$\frac{\partial \boldsymbol{P}}{\partial t} = \frac{\partial (mn\boldsymbol{v}^M)}{\partial t} = mn\frac{\partial \boldsymbol{v}^M}{\partial t} + m\boldsymbol{v}^M C_n - \boldsymbol{v}^M \nabla_{\boldsymbol{x}} \cdot \boldsymbol{P}.$$

Hence,

$$mn\frac{\partial \boldsymbol{v}^M}{\partial t} + (\boldsymbol{P} \cdot \nabla_{\boldsymbol{x}})\boldsymbol{v}^M + \nabla_{\boldsymbol{x}} \cdot (nk\overline{\overline{\boldsymbol{T}}}) - n\boldsymbol{F} = \boldsymbol{C}_{\boldsymbol{P}} - \frac{1}{n}\boldsymbol{P}C_n. \tag{11}$$

Since

$$C_n = \left(\frac{\partial n}{\partial t}\right)_{\text{coll}},$$

$$\frac{1}{m}\boldsymbol{C}_{\boldsymbol{P}} = n\left(\frac{\partial \boldsymbol{v}^M}{\partial t}\right)_{\text{coll}} + \boldsymbol{v}^M\left(\frac{\partial n}{\partial t}\right)_{\text{coll}},$$

the right-hand side of (11) reduces to

$$mn\left(\frac{\partial \boldsymbol{v}^M}{\partial t}\right)_{\text{coll}}.$$

At this point, it is common to make further assumptions. First, the temperature tensor is assumed to be scalar. As a consequence,

$$\nabla_{\boldsymbol{x}} \cdot (nk\overline{\overline{\boldsymbol{T}}}) = kT\nabla_{\boldsymbol{x}} n + n\nabla_{\boldsymbol{x}}(kT).$$

Secondly, the time rate of change of $\boldsymbol{v}^M$ due to collisions is assumed to be proportional to $\boldsymbol{v}^M$

$$\left(\frac{\partial \boldsymbol{v}^M}{\partial t}\right)_{\text{coll}} = -\frac{1}{\tau_{\text{mom}}}\boldsymbol{v}^M.$$

Here, τ_{mom} is the momentum relaxation time and it is related to the mobility of the charge carriers through the equation

$$\mu = \frac{q}{m}\tau_{\text{mom}}.$$

With these definitions and assumptions, (11) can be written as

$$mn\frac{\partial \boldsymbol{v}^M}{\partial t} + m(n\boldsymbol{v}^M \cdot \nabla_{\boldsymbol{x}})\boldsymbol{v}^M + kT\nabla_{\boldsymbol{x}} n + n\nabla_{\boldsymbol{x}}(kT) - n\boldsymbol{F} = -\frac{q}{\mu}n\boldsymbol{v}^M. \tag{12}$$

The drift-diffusion model is now closed by using (12) in the following way as the defining relation for the current density $\boldsymbol{J}$. Assuming a constant temperature T, and neglecting the convective term (that is the second term in the left-hand side) and the time derivative of $\boldsymbol{v}^M$, it is not hard to show that

$$-\frac{q}{\mu}n\boldsymbol{v}^M = kT\nabla_{\boldsymbol{x}} n - n\boldsymbol{F},$$

or, equivalently,

$$\boldsymbol{J} = kT\mu\nabla_{\boldsymbol{x}} n - \mu n\boldsymbol{F}. \tag{13}$$

This is a special case of the more general expression

$$\boldsymbol{J} = qD\nabla_{\boldsymbol{x}} n - \mu n\boldsymbol{F}, \tag{14}$$

where D denotes the diffusion coefficient. If the Einstein relation

$$D = \frac{kT}{q}\mu,$$

is assumed to hold, then (14) reduces to (13).

The derivation of more advanced models is similar to the above derivation, with different types of closing relations giving different models. A model receiving increasing attention is the hydrodynamic model. Here, the above assumption of a constant temperature is relaxed, and effective hole and electron temperatures are introduced as new variables. The result is a coupled system of five partial differential equations, which are very similar to the Euler equations of gas dynamics. In the following section, the equations for this model are summarized. A complete derivation can be found in [8].

1.3. Formulation of the mathematical models

The drift-diffusion model consists of the following nonlinear system of partial differential equations. First we have Poisson's equation

$$-\nabla \cdot (\varepsilon \nabla \psi) = q(p - n + D), \tag{15}$$

and then the continuity equations for holes and electrons respectively

$$q\frac{\partial p}{\partial t} = -\nabla \cdot \boldsymbol{J}_p - qR, \tag{16}$$

and

$$q\frac{\partial n}{\partial t} = \nabla \cdot \boldsymbol{J}_n - qR. \tag{17}$$

Poisson's equation is solved in the entire simulation domain, whereas the continuity equations are restricted to the parts of the domain with semiconductor material. The right-hand side of the Poisson equation has to be replaced by zero in the parts of the domain with non-semiconductor material.

The current densities $\boldsymbol{J}_p$ and $\boldsymbol{J}_n$ are related to the electrostatic potential and the carrier concentrations by the equations

$$\boldsymbol{J}_p = -kT\mu_p \nabla p + q\mu_p p \boldsymbol{E}_p, \tag{18}$$

$$\boldsymbol{J}_n = kT\mu_n \nabla n + q\mu_n n \boldsymbol{E}_n. \tag{19}$$

Here, the effective fields $\boldsymbol{E}_p$ and $\boldsymbol{E}_n$ reflect the effects of band gap narrowing and are given by

$$\boldsymbol{E}_p = \boldsymbol{E} + \frac{kT}{q}\nabla \log n_{\text{int,eff}}, \tag{20}$$

$$\boldsymbol{E}_n = \boldsymbol{E} - \frac{kT}{q}\nabla \log n_{\text{int,eff}}. \tag{21}$$

With the on-going miniaturization of devices, phenomena which are neglected in the drift-diffusion model are becoming increasingly important. The additional physical effects are taken account of in the hydrodynamic model which is described below. If the convective term is neglected, the resulting model is also referred to as the energy balance model.

The steady-state hydrodynamic model consists of the following system of nonlinear partial differential equations. First we have Poisson's equation

$$-\nabla \cdot (\varepsilon \nabla \psi) = q(p - n + D). \tag{22}$$

Next we have the continuity equations for holes and electrons, respectively

$$\nabla \cdot \boldsymbol{J}_p = -qR, \tag{23}$$

and

$$\nabla \cdot \boldsymbol{J}_n = qR, \tag{24}$$

where the current densities are given by the expressions

$$\boldsymbol{J}_p = -q D_p \nabla p + q\mu_p p \boldsymbol{E}_p - q D_p^T p \nabla T_p, \tag{25}$$

$$\boldsymbol{J}_n = q D_n \nabla n + q\mu_n n \boldsymbol{E}_n + q D_n^T n \nabla T_n, \tag{26}$$

in which

$$\boldsymbol{E}_p = -\nabla\psi + \frac{kT_0}{q}\nabla \log\ n_{int,eff}, \tag{27}$$

$$\boldsymbol{E}_n = -\nabla\psi - \frac{kT_0}{q}\nabla \log\ n_{int,eff}. \tag{28}$$

Here, T_p and T_n denote the hole and electron temperatures, respectively, and T_0 is the lattice temperature.

In the hydrodynamic model, we also have the following equations for the energy flux densities

$$\nabla \cdot \boldsymbol{S}_p = \boldsymbol{E} \cdot \boldsymbol{J}_p - R w_p - p\frac{w_p - w_0}{\tau_w^p}, \tag{29}$$

and

$$\nabla \cdot \boldsymbol{S}_n = \boldsymbol{E} \cdot \boldsymbol{J}_n - R w_n - n\frac{w_n - w_0}{\tau_w^n}, \tag{30}$$

where w_p and w_n are the average energies of the carriers and are defined by

$$w_p = \frac{1}{2} m_p \left(\frac{J_p}{pq}\right)^2 + \frac{3}{2}kT_p, \tag{31}$$

$$w_n = \frac{1}{2} m_n \left(\frac{J_n}{nq}\right)^2 + \frac{3}{2}kT_n, \tag{32}$$

and the quantities $\boldsymbol{S}_p$ and $\boldsymbol{S}_n$ are given in terms of the carrier temperatures by

$$\boldsymbol{S}_p = -\kappa_p \nabla T_p + \frac{1}{q}(w_p + kT_p)\boldsymbol{J}_p, \tag{33}$$

and

$$\boldsymbol{S}_n = -\kappa_n \nabla T_n - \frac{1}{q}(w_n + kT_n)\boldsymbol{J}_n. \tag{34}$$

The coefficients κ_p and κ_n are usually modelled using the Wiedemann–Franz relations

$$\kappa_p = C_{\mathrm{WF}} k D_p p,$$

$$\kappa_n = C_{\mathrm{WF}} k D_n n,$$

where C_{WF} is a constant.

Finally we remark that the following relations are often used:

$$D_p = \frac{kT_p}{q}\mu_p, \qquad D_n = \frac{kT_n}{q}\mu_n \tag{35}$$

$$D_p^T = \frac{k}{q}\mu_p, \qquad D_n^T = \frac{k}{q}\mu_n \tag{36}$$

and

$$C_{\mathrm{WF}} = \tfrac{5}{2} + \alpha, \qquad \alpha = 0.8. \tag{37}$$

2. Analysis of the mathematical models

The drift-diffusion model of semiconductor devices in two and three dimensions is defined by the Van Roosbroeck equations (15)–(21). These form a complicated system of nonlinear time-dependent partial differential equations, two of which are parabolic and one of which is elliptic [8, 9]. The geometry of devices in two or three dimensions, and therefore of the domain in which the solution of the above equations is sought, is usually complex. There is no general mathematical theory for such systems, and thus questions of existence, uniqueness and smoothness of the solutions cannot be answered in general. Indeed there are cases where it is known that more than one solution is possible. Thus the situation is somewhat similar to that for the Navier–Stokes equations which govern fluid dynamics.

It is obvious that analytic solutions in closed form are not possible except in rare special cases. Because of the great importance of semiconductor devices in the modern electronics industry, it is desirable that the design and behaviour of a device be predicted to a known accuracy in advance of its actual fabrication. The cost of rectifying a design error at a later stage is enormous. One of the main ways of guarding against such a disaster is to simulate the device using a suitable numerical model. This has given rise to the creation of expensive numerical software both within the larger multinational electronics companies and in small specialist service companies.

In parallel to the lack of a mathematical analysis of the Van Roosbroeck equations, there is also no general theory for the convergence, or otherwise, of the numerical methods used to simulate semiconductor devices. The numerical analysis of such methods has usually been carried out only for some linearized form of the problem, and the system of equations is often assumed to be decoupled, so that the equations can be studied one at a time. Thus, there are no error estimates applicable to the simulations used for practical applications. This gives rise to the need for other criteria to judge whether or not the results of a particular simulation are of scientific value. The current state of the art is based mainly on the comparison of the results of simulations on a given mesh and on a refinement of that mesh.

2.1. Mathematical properties of the models

The drift-diffusion model derived in section 1.3 is frequently used to model the behaviour of semiconductor devices. It is expected that solutions of this system of equations describe the behaviour adequately provided that the models for the physical parameters are sufficiently accurate. On the other hand, it is possible to analyse solutions of the coupled system of equations mathematically, and to extract properties of these solutions without knowing anything about the semiconductor devices. It is hoped, of course, that these properties reflect accurately the typical behaviour of the semiconductor devices. Only a detailed mathematical analysis can confirm this. Although simulations of semiconductor device behaviour using the drift-diffusion model have been performed for a number of decades now, a careful mathematical analysis of the underlying system of equations and its solutions has been carried out only fairly recently. A rigorous mathematical analysis can be found in the work of the two pioneers, Mock [10–13] and Markowich [1, 7, 14]. Other workers in the field are Brezzi [15], Gajewski [16, 17], Jerome [18], Kerkhoven [19, 20] and Seidman [21, 22]. The reader is referred to these papers and references therein for more details.

For the nonlinear Poisson equation, the existence of solutions follows directly from the ellipticity of the partial differential equation. For the complete system of drift-diffusion equations the existence of solutions cannot be concluded directly. However, a change of variables can be used to transform the continuity equations into self-adjoint equations. To this

end the Slotboom variables Φ_p and Φ_n are introduced:

$$\Phi_p = p \exp\left(\frac{q\psi}{kT}\right), \qquad \Phi_n = n \exp\left(-\frac{q\psi}{kT}\right). \tag{38}$$

Using these new variables, the current densities have the form

$$\boldsymbol{J}_p = -kT\mu_p n_{\text{int,eff}} \exp\left(-\frac{q\psi}{kT}\right) \nabla\Phi_p, \tag{39}$$

and

$$\boldsymbol{J}_n = kT\mu_n n_{\text{int,eff}} \exp\left(\frac{q\psi}{kT}\right) \nabla\Phi_n. \tag{40}$$

These expressions are substituted into the continuity equations, so that the steady-state drift-diffusion system in the variables ψ, Φ_p and Φ_n becomes

$$-\nabla \cdot (\varepsilon\nabla\psi) = q\left(n_{\text{int,eff}} \exp\left(-\frac{q\psi}{kT}\right)\Phi_p - n_{\text{int,eff}} \exp\left(\frac{q\psi}{kT}\right)\Phi_n + D\right), \tag{41}$$

$$-\nabla \cdot \left(kT\mu_p n_{\text{int,eff}} \exp\left(-\frac{q\psi}{kT}\right)\nabla\Phi_p\right) = -qR, \tag{42}$$

$$-\nabla \cdot \left(kT\mu_n n_{\text{int,eff}} \exp\left(\frac{q\psi}{kT}\right)\nabla\Phi_n\right) = -qR. \tag{43}$$

If the mobilities μ_p and μ_n are positive, then the three partial differential equations in this system are elliptic for a given potential ψ and a given recombination/generation rate R. This observation can be used in an existence proof if the equations are treated independently. For this reason, most such proofs are based on a decoupling of the equations. To this end the operator T is introduced, which maps a pair (Φ_p^0, Φ_n^0) onto a pair (Φ_p^1, Φ_n^1), in the following way:

(1) For given $(\Phi_p, \Phi_n) = (\Phi_p^0, \Phi_n^0)$, solve the nonlinear Poisson equation (41) subject to the given boundary conditions. The solution is denoted by ψ^1.
(2) Solve (42), with $\psi = \psi^1$, $\Phi_n = \Phi_n^0$, and $R = \hat{R}$, subject to the given boundary conditions. Here, $\hat{R}$ is a suitably linearized (around $\Phi_p = \Phi_p^0$) form of R. Denote the solution by Φ_p^1.
(3) Solve (43), with $\psi = \psi^1$, $\Phi_p = \Phi_p^0$, and $R = \bar{R}$, subject to the given boundary conditions. Here, $\bar{R}$ is a suitably linearized (around $\Phi_n = \Phi_n^1$) form of R. Denote the solution by Φ_n^1.

Clearly, additional assumptions are needed to guarantee that the solutions Φ_p^1 and Φ_n^1 defined in steps (2) and (3) exist. Markowich [14] assumes that the mobilities satisfy the following conditions:

- μ_p and μ_n depend only on the spatial coordinates and on the electric field $\boldsymbol{E} = -\nabla\psi$
- μ_p and μ_n are uniformly bounded away from zero, and are uniformly bounded above
- the functions μ_p and μ_n are Lipschitz-continuous with respect to the electric field.

Unfortunately, these assumptions are quite restrictive. In fact, mobility models depending on the carrier concentrations and the quasi-Fermi levels do not fulfil these requirements. Note also that steps (2) and (3) require a suitable linearization of the recombination/generation term. Finally, it is clear that the solution domain and the boundary conditions also must satisfy certain conditions in order to guarantee the existence of solutions in steps (2) and (3). Again, Markowich [14] states precise conditions.

Having defined T, the question arises whether or not this operator has a fixed point. We assume that this is the case, and denote the components of the solution by Φ_p^* and Φ_n^*.

In addition, we denote the corresponding electric potential by ψ^*. Then $(\psi^*, \Phi_p^*, \Phi_n^*)$ is clearly a solution of the drift-diffusion system (41)–(43). Hence, establishing the existence of a solution is reduced to proving that the operator T has a fixed point. Using the above assumptions, Schauder's fixed point theorem [23] can be used to establish the existence of a fixed point. For a proof, the reader is referred to [14].

In the above, mathematical detail has intentionally been kept to a minimum. Indeed, the purpose was not to give a rigorous proof of the existence of solutions, but rather to highlight some aspects which are important from a numerical point of view. The first of these concerns the conditions imposed on the mobility models. Although these are sometimes restrictive, they should certainly not be discarded as being mathematical restrictions imposed solely for the purpose of making possible a proof of the required result. On the contrary, it is not too difficult to construct models violating the conditions, for which no solution in terms of the Slotboom variables exists. The conclusion is that the results obtained from a mathematical analysis of the equations can be useful in designing appropriate models. Since the ultimate aim is to obtain an accurate description of the behaviour of semiconductor devices, it may even be advisable to let the design of the physical models be guided by the mathematical restrictions.

The second important aspect is the use of the fixed point operator T. It is clear that the solution of the drift-diffusion system can be found by determining the fixed point of T. This equivalence suggests a numerical solution procedure in which steps (1)–(3) in the above are used sequentially to yield new approximations to the potential and the Slotboom variables. Clearly, the fixed point can be found only when this process converges. This is the case whenever T is a contraction mapping. Gummel's sequential method, which will be discussed in section 2.2, is based on these observations. Kerkhoven has been the main contributor to the analysis of the special properties of the fixed point operator T [19, 20]. He investigated the location of the eigenvalues of the Jacobian matrix of the maps which constitute the fixed point mapping T. In particular, he proved that the absolute values of these (complex) eigenvalues are unconditionally smaller than unity for a suitable choice of the linearization in steps (2) and (3). This paves the way for a globally convergent solution strategy. Indeed, experimental evidence indicates that Gummel's algorithm converges from any initial guess.

So far, the discussion has been limited to establishing existence rather than uniqueness of solutions. Clearly, there is a good reason for this limitation: it cannot be expected that uniqueness of steady-state solutions can be proved in general. Potentially several physical mechanisms, such as avalanche generation, yield multiple solutions. On the other hand, uniqueness of the solution under zero bias conditions can be proved rigorously [13]. Because of the continuous dependence of the solution on the boundary conditions, uniqueness can also be demonstrated for sufficiently small values of the applied potentials. When performing numerical simulations, one should certainly keep in mind the possibility of the existence of multiple solutions. Often, stagnation of the convergence or even divergence of the solution process is an indication that the solution is not unique. If this occurs, a slight modification of the model may lead to a problem which again has a unique solution.

2.2. Numerical techniques

Although analytical solutions have been obtained for several devices with a simple geometry and piecewise constant doping function, it is evident from the structure of the system of differential equations obtained in section 1.2 that numerical methods must be used to obtain approximate solutions in most cases. Within the area of semiconductor device simulation, a large variety of methods has been investigated, and it is beyond the scope of this paper to describe all of these in detail. For an extensive overview, the reader is referred to [8, 9]. In this

section, a brief discussion of the most popular numerical techniques is given.

In general, the following main steps can be identified in a typical numerical algorithm for obtaining approximate solutions of a nonlinear system of partial differential equations:

- discretization of the domain (meshing)
- discretization of the differential equations
- solution of the resulting nonlinear algebraic systems
- solution of sequences of linear systems.

The difference between the continuous solution and the approximate solution is mainly determined by the methods used in the first two steps. The methods used in the third and fourth steps influence mainly the computation time. Although the choice of method within each of these steps is not necessarily related to the choice of methods in the other steps, there is usually some degree of interdependence. If, for example, a rectangular grid is chosen in the first step, the matrices in the resulting linear systems have a banded structure, and dedicated methods for such systems can be chosen in the fourth step. Furthermore, the choice of a discretization method also influences the properties of the resulting system matrices, which in turn may be decisive for the choice of the linear solution technique.

The first main step above consists of selecting a finite set of points at which the solution will be determined approximately. Most discretization methods in the second main step require a set of points which is based upon a subdivision of the solution domain into smaller elements. This can be done in several ways. For example, if the solution domain has a rectangular structure, then it is convenient to subdivide it into small rectangular elements. Typically, the edge of the solution domain in each coordinate direction is subdivided into 40 to 100 pieces. This can be done in a uniform way, keeping the length of the subintervals constant, or in a non-uniform way. The latter is particularly appropriate if some *a priori* information about the behaviour of the solution is available. An alternative discretization of the solution domain is obtained if the rectangular elements are subdivided into triangular elements with analogous subdivisions into tetrahedral elements for three-dimensional brick elements. If the solution domain is not of a rectangular structure, it is often more convenient to cover it with triangular or tetrahedral elements. This task should be carried out, preferably automatically, by a mesh generation program. Once a mesh has been generated, the solution can be determined approximately at special points associated with this mesh. A common choice is the set of vertices, but it is also possible to use the midpoints of the element edges or the barycentre of each element.

To obtain equations for the approximate solution values at the selected points, the differential equations and boundary conditions are then discretized. A variety of techniques is available for this task. Some of these techniques require only the coordinates of the selected points, whereas other techniques make use of the subdivision into elements of the solution domain. A well known example of the former is the finite difference method, in which derivatives are approximated by divided differences. An approximate solution on the entire domain can then be obtained by piecewise polynomial interpolation of the values at the selected points. On the other hand, an example of the latter is the finite element method, which makes use of the underlying structure of the mesh. In fact, it generates approximate solution values on the entire solution domain by assuming that the function is a polynomial on each of the elements. For semiconductor device simulation, the finite volume method is the most popular discretization technique. Its use is restricted to differential equations which are in conservative form, that is of the form

$$\nabla \cdot \boldsymbol{f}(u) = g(u),$$

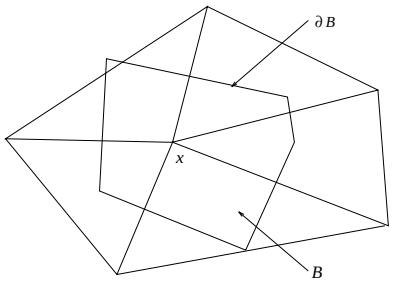

Figure 4. Construction of a finite volume in the case of a triangular mesh.

where $\boldsymbol{f}(u)$ is a flux. Integrating this equation over an arbitrary subdomain B with boundary ∂B, and using Green's formula for the integral of the left-hand side, gives

$$\int_{\partial B} \boldsymbol{f}(u) \cdot \boldsymbol{n} = \int_B g(u).$$

The integral on the left-hand side is then approximated by a finite sum of the form

$$\sum_i \boldsymbol{f}(u) \cdot \boldsymbol{n}_i l_i,$$

where $\boldsymbol{f}(u) \cdot \boldsymbol{n}_i$ is the normal component at some point on an associated small part of the boundary ∂B, and l_i is its length. The integral on the right-hand side can be approximated in a similar way, using only a finite number of values of $g(u)$. Notice that the finite volume is the area of the subdomain in two-dimensional problems and its volume in three-dimensional problems.

The finite volume method can be viewed as a generalization of the finite difference method, but it is also related to the finite element method. The former view arises from the fact that the normal components of the fluxes are approximated by difference quotients of the approximate solution values at the mesh points. On the other hand, the construction of the subdomains of B is based on the underlying mesh. For example, if a triangular mesh is used, the subdomains (boxes or finite volumes) B are constructed using the intersecting mid-edge perpendiculars, as illustrated in figure 4. This construction also applies in the simpler case of a rectangular mesh, resulting in rectangular boxes. It is easy to check that the collection of boxes constructed in this way covers the entire solution domain.

After the meshing and discretization steps, a system of algebraic equations is obtained for the unknown approximate solution values at the mesh points. Since the original problem is nonlinear, the discretized system of equations is also nonlinear. Therefore, iterative solution techniques are usually used to solve this system. Unfortunately, there is no guarantee that a solution of such a system exists. As has been explained in detail in [8, 9], the occurrence of obtuse triangles may lead to the non-existence of solutions, although solutions may exist for some bias conditions. For this reason, it is often advisable to use either a rectangular mesh, or a triangular mesh which satisfies the Delaunay criterion. The latter is equivalent to requiring that the sum of the angles opposite the common edge of two neighbouring triangles does not exceed π. Clearly, this is a condition which is hard to achieve manually. Unfortunately, several commercially available simulation packages still place the burden of designing a suitable mesh on the user.

The nonlinear system of algebraic equations can be solved using for example Newton's method. If the discretized system is denoted by

$$\boldsymbol{F}(\boldsymbol{u}) = \boldsymbol{0},$$

then this amounts to solving a sequence of linear systems of the form

$$\boldsymbol{F}'(\boldsymbol{u}^{(k)})\delta\boldsymbol{u}^{(k)} = -\boldsymbol{F}(\boldsymbol{u}^{(k)}),$$

for the unknown vector $\delta\boldsymbol{u}^{(k)}$. Here $\boldsymbol{u}$ is the vector of the unknown solution values, $\boldsymbol{F}$ is the vector containing the discretized equations and boundary conditions, and $\boldsymbol{F}'$ is the Fréchet derivative (or, in this discrete case, the Jacobian matrix) of $\boldsymbol{F}$. The vectors $\boldsymbol{u}^{(k)}$, $k = 0, 1, 2, \ldots$, are successive approximations to the solution $\boldsymbol{u}$, and are related by

$$\boldsymbol{u}^{(k+1)} = \boldsymbol{u}^{(k)} + \delta\boldsymbol{u}^{(k)}.$$

Hence, the Newton process yields corrections to previously determined approximations. To start the process, an initial estimate $\boldsymbol{u}^{(0)}$ is required.

The main advantage of Newton's method, over other nonlinear solution techniques, is its quadratic convergence property. Unfortunately, relatively accurate initial guesses are required in order for the sequence of iterates to converge in this manner. To avoid the latter difficulty, a different method has become rather popular within the area of semiconductor device simulation. Recognizing the fact that the discretized system of equations is obtained by discretizing three different differential equations, it is natural to make use of this subdivision also in the solution of the nonlinear system. With this is mind the system is written in the form

$$\begin{aligned}
\boldsymbol{F}_1(\boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{u}_3) &= \boldsymbol{0},\\
\boldsymbol{F}_2(\boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{u}_3) &= \boldsymbol{0},\\
\boldsymbol{F}_3(\boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{u}_3) &= \boldsymbol{0},
\end{aligned}$$

where the unknown vectors $\boldsymbol{u}_1$, $\boldsymbol{u}_2$, $\boldsymbol{u}_3$ denote the values of the electrostatic potential, the electron quasi-Fermi level, and the hole quasi-Fermi level, respectively. Similarly, $\boldsymbol{F}_1$, $\boldsymbol{F}_2$, and $\boldsymbol{F}_3$ denote the discretized versions of the corresponding differential equations and boundary conditions. The three nonlinear equations can now be solved in turn for the corresponding unknown vectors, keeping the other unknown vectors constant. It turns out that this nonlinear Gauss–Seidel method, which is known as Gummel's method [9, 24], is quite robust, in the sense that it converges from almost any initial estimate. Note that the solution of the individual nonlinear systems can still be performed using Newton's method.

Both Newton's method and Gummel's method require the solution of sequences of linear systems of equations of the form

$$A\boldsymbol{u} = \boldsymbol{b}. \tag{44}$$

Two distinct classes of methods for accomplishing this task are common. The first is the class of direct methods based on a decomposition of the matrix into simple matrix components, which are then solved by a simple solution algorithm. An example of this technique is Gaussian elimination, which leads to a triangular decomposition of the form $A = LU$ where L is a lower triangular matrix, and U is upper triangular. Solving (44) then amounts to first solving the system

$$L\boldsymbol{y} = b$$

by forward substitution and then solving the system

$$Ua = \boldsymbol{y}.$$

by backward substitution. The second class is that of the iterative methods, in which the solution is obtained by using a sequence of matrix-vector products. For large systems of equations this is more attractive than direct methods, since no decomposition of the matrix is required. A popular example is the conjugate gradient method [25, 26]. Unfortunately, the convergence of these iterative methods is often too slow for practical purposes. For this

reason, the resulting linear systems are often solved using a combination of iterative and direct methods. First the linear system is preconditioned by using an approximate decomposition of the matrix, usually of the form $\hat{L}\hat{U}$. Next, an iterative method is used to solve the system

$$(\hat{L}\hat{U})^{-1}A\boldsymbol{x} = (\hat{L}\hat{U})^{-1}\boldsymbol{b}.$$

It is sometimes possible to give rigorous mathematical reasons for the improved convergence of these preconditioned iterative methods, but it is intuitively clear that the matrix $(\hat{L}\hat{U})^{-1}A$ approximates the identity matrix when the approximate decomposition is close to the exact LU-decomposition of A.

For time-dependent problems, techniques similar to those described above may be used. Because of the unidirectional character of these problems (in the time direction), the mesh is often determined adaptively. Furthermore, the solution techniques can benefit from the fact that the solutions are often not significantly different at consecutive time levels. We will not give details of these methods here, since our primary goal in this paper is to present finite element methods for the discretization of the spatial part of the partial differential equations.

Numerical techniques designed for solving the drift-diffusion model can be extended also to the case of the hydrodynamic model, although there are several changes to be made. In [9] it is demonstrated that a straightforward application of the methods commonly used for solving the drift-diffusion model may lead to unrealistic hole and electron temperatures. The same reference also contains recommendations for overcoming these problems, but it is beyond the scope of this paper to give any further details.

3. Ordinary finite element methods

The finite element method was used by engineers for solving problems in structural mechanics long before it was discovered by mathematicians. The main reason for its use in engineering problems is its geometrical flexibility, combined with the ease with which its equations can be assembled for complicated problems. From a mathematical point of view, one might argue that the method is essentially a weighted residual method. In such methods, a continuous problem is put into variational form. Next, an approximate solution is sought in the form $\sum_i w_i\phi_i$ where the ϕ_i are trial functions. The weights w_i are determined so as to minimize some functional. The origin of these methods is the well known Rayleigh–Ritz–Galerkin technique. The new feature introduced by the finite element method is that the trial functions are piecewise polynomials, with each component polynomial being defined only on a small subdomain or element, rather than functions which extend over the entire domain. It is exactly this feature that has made the finite element method such a tremendous success in many areas.

In this section we consider ordinary finite element methods, while mixed finite element methods are discussed in the next. Finite element methods are said to be of standard type if no special techniques such as upwinding, fitting or inverse averaging are used in their construction. Otherwise they are of non-standard type. The theory of standard finite element methods is well developed, and there is a large amount of literature on the subject. Many French mathematicians, such as Lions, Temam, Raviart and Ciarlet, have published extensively on the theoretical aspects [27–29]. Also recommended for readers are the books by Strang and Fix [30], and Oden and Reddy [31]. More recent books are those by Axelsson and Barker [32], Silvester and Ferrari [33], and Johnson [34], which also discuss some of the computational issues. The theory of non-standard finite element methods is less comprehensive.

In what follows it is convenient to introduce the notation $\Omega = (0, 1)$, $\overline{\Omega} = [0, 1]$, $C^0(\overline{\Omega})$ for the space of continuous functions on $\overline{\Omega}$, $C^1(\overline{\Omega}) = \{v : v, v' \in C^0(\overline{\Omega})\}$ for the space of continuously differentiable functions on $\bar{\Omega}$, $L^2(\Omega)$ for the space of square integrable functions

on Ω and $H^1(\Omega) = \{v : v, v' \in L^2(\Omega)\}$ for the space of functions with square integrable derivatives on Ω.

The standard scalar product for $u, v \in L^2(\Omega)$ is $(u, v) = \int_0^1 u(x)v(x)\,\mathrm{d}x$ and the norm is $\|v\| = \sqrt{(v, v)}$. The norm on any subspace $V \subset H^1(\Omega)$ is defined to be $\|v\|_1 = \sqrt{\|v\|^2 + \|v'\|^2}$.

3.1. A standard ordinary finite element method

When discretizing a problem using the finite element method, two basic steps can be distinguished. In order not to complicate the discussion, the following simple ordinary differential equation in conservation form is used to illustrate these

$$\begin{cases} -(a(x)u'(x))' + b(x)u(x) = f(x), & x \in \Omega \\ u(0) = 0, \qquad u(1) = 0. \end{cases} \tag{45}$$

It is assumed that $a(x) \geqslant \alpha > 0$, for some constant α, and that $b(x) \geqslant 0$ for all $x \in \overline{\Omega}$. If the problem has non-zero boundary conditions $u(0) = u_0$ and $u(1) = u_1$ it is easy to check that the substitution $v(x) = u(x) - [u_0 + (u_1 - u_0)x]$ renders the non-homogeneous boundary conditions homogeneous without changing the form of the above equation. In fact only the source term on the right-hand side changes.

The first step is to replace this classical formulation of the problem by a suitable variational formulation. The latter can be derived by multiplying the differential equation by a smooth test function $v(x)$ and integrating over the solution domain $\overline{\Omega}$ giving

$$\int_0^1 \{-(a(x)u'(x))'v(x) + b(x)u(x)v(x)\}\,\mathrm{d}x = \int_0^1 f(x)v(x)\,\mathrm{d}x.$$

Then, integrating the first term on the left-hand side by parts yields

$$-[a(x)u'(x)v(x)]_0^1 + \int_0^1 \{a(x)u'(x)v'(x) + b(x)u(x)v(x)\}\,\mathrm{d}x = \int_0^1 f(x)v(x)\,\mathrm{d}x.$$

Note that the differentiability of the function v has been used. Furthermore, it would be convenient if $v(0) = v(1) = 0$, since this would eliminate the first term on the left-hand side. This suggests introducing the space of admissible trial and test functions

$$V = \{v \in C^1(\overline{\Omega}) : v(0) = v(1) = 0\}. \tag{46}$$

With this choice of V the original problem (45) has been transformed to the desired variational or weak form

$$\begin{cases} \text{Find } u \in V \qquad \text{such that for all} \quad v \in V, \\ \displaystyle\int_0^1 \{a(x)u'(x)v'(x) + b(x)u(x)v(x)\}\,\mathrm{d}x = \int_0^1 f(x)v(x)\,\mathrm{d}x. \end{cases} \tag{47}$$

It might be expected that solutions of this problem and the original problem coincide. However, this may not always be the case, since (47) may have a solution which does not have a second derivative, and hence cannot satisfy the original differential equation (45). Such a solution is called a weak solution. If a weak solution does satisfy the original differential equation then it is called a strong or classical solution.

The problem of finding a suitable weak formulation for a given problem is discussed extensively in several books on finite element methods. An important issue is the choice of the space V of the admissible test functions v and the possible solutions u. Note that the weak problem (47) can be posed for a larger class of functions V than given in (46), since the

problem is well posed whenever the integrals have a finite value. Thus, the functions u and v may be in the space $H_0^1(\Omega)$, defined as

$$H_0^1(\Omega) = \{v \in H^1(\Omega) : v(0) = v(1) = 0\}. \tag{48}$$

The variational problem can then be posed on this space by taking V in (47) to be $H_0^1(\Omega)$. Note that its solutions are not always strong solutions for the same reasons as before.

Before discussing our finite element method we first show that the above variational problem is uniquely solvable. By the Lax–Milgram theorem (see [35] for example) we know that if the bilinear form (integral) corresponding to the left-hand side of (47) is V-elliptic or coercive, that is if

$$\int_0^1 \{a(x)(v'(x))^2 + b(x)v^2(x)\}\,\mathrm{d}x \geqslant C\|v\|_1, \qquad \forall v \in V \tag{49}$$

for a constant $C > 0$, independent of V and u, then the variational problem has a unique solution. We now verify that this is indeed the case. We let $C > 0$ be a generic constant, independent of V and u. Choosing $u = v$ and using the assumption that $a(x) \geqslant \alpha$ and $b(x) \geqslant 0$ we have

$$\int_0^1 \{a(x)(v'(x))^2 + b(x)v^2(x)\}\,\mathrm{d}x \geqslant \alpha \int_0^1 (v'(x))^2\,\mathrm{d}x = \alpha\|v'\|^2. \tag{50}$$

To complete the proof it remains to show that $\|v'\|^2 \geqslant C\|v\|_1^2$, or equivalently $\|v'\|^2 \geqslant C\|v\|^2$, because $\|v\|_1^2 = \|v\|^2 + \|v'\|^2$. Integrating by parts we get

$$\begin{aligned}\int_0^1 u^2\,\mathrm{d}x &= \left[u(x)\int_0^x u(t)\,\mathrm{d}t\right]_0^1 - \int_0^1 u'(x)\left(\int_0^x u(t)\,\mathrm{d}t\right)\mathrm{d}x \\ &\leqslant \int_0^1 |u|\,\mathrm{d}x \cdot \int_0^1 |u'|\,\mathrm{d}x \\ &\leqslant \left(\int_0^1 u^2\,\mathrm{d}x\right)^{1/2}\left(\int_0^1 (u')^2\,\mathrm{d}x\right)^{1/2}.\end{aligned}$$

In the above we have used the Cauchy–Schwarz inequality. From this it follows that

$$\|v\| \leqslant \|v'\|. \tag{51}$$

Combining this and (50) we obtain (49) with $C = \alpha$. Thus the problem (47) is uniquely solvable. We comment that the inequality (51) is the Poincaré–Friedrichs inequality in one dimension. It can be extended to multi-dimensions and to cases where the boundary conditions involve function values on just part of the boundary of Ω (cf, for example, [36, p 139]).

The second step in the discretization procedure is to transform the weak problem (47), which is defined on an infinite-dimensional space V, to a problem defined on a finite-dimensional subspace V_h of V. This means that the space V has to be replaced by a finite-dimensional subspace V_h. To this end, the solution domain Ω is partitioned into a large number of small subdomains called elements. Next, the finite-dimensional subspace V_h of the space V is constructed in the following way. In the finite element method these finite-dimensional spaces consist of piecewise polynomials, in the sense that the restriction of any function to an element is a polynomial. The polynomial is in general different on different elements. Sometimes, it is not even required that these functions are continuous at the interfaces between two elements.

For the one-dimensional problem under consideration, it is assumed that a mesh $x_0, \ldots, x_N$ is specified, where $0 = x_0 < x_1 < \cdots < x_{N-1} < x_N = 1$. Then, the elements are the

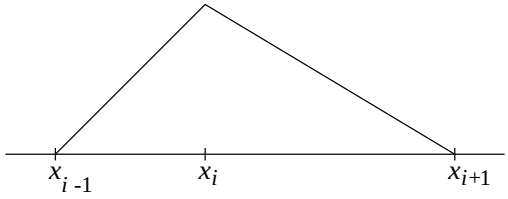

Figure 5. Piecewise linear hat functions ϕ_i.

subintervals $\Omega_i = (x_{i-1}, x_i)$, $i = 1, \ldots, N$. We let $h_i = x_i - x_{i-1}$ for $i = 1, \ldots, N$ and $h = \max_{1 \leqslant i \leqslant N} h_i$. One of the simplest piecewise polynomial spaces is

$$V_h = \{v_h \in C^0(\overline{\Omega}) : v_h|_{\overline{\Omega}_i} \in P^1(\overline{\Omega}_i) \text{ for all } i = 1, 2, \ldots, N-1, \text{ and } v_h(0) = v_h(1) = 0\}, \tag{52}$$

where $P^1(\overline{\Omega}_i)$ is the space of linear polynomials on $\overline{\Omega}_i$ and $v|_{\overline{\Omega}_i}$ denotes the restriction of v to the subinterval $\overline{\Omega}_i$. It is easy to check that $V_h \subset V$. The discrete variational problem corresponding to (47) is then

$$\begin{cases} \text{Find } u_h \in V_h \qquad \text{such that for all} \quad v_h \in V_h \\ \displaystyle\int_0^1 \{a(x)u_h'(x)v_h'(x) + b(x)u_h(x)v_h(x)\}\,\mathrm{d}x = \int_0^1 f(x)v_h(x)\,\mathrm{d}x. \end{cases} \tag{53}$$

Since $V_h \subset V$, a discrete analogue of the coercivity condition (49) is also satisfied by the finite element space V_h, and thus the discrete problem also has a unique solution. An alternative discrete variational formulation is obtained if the coefficient functions $a(x)$, $b(x)$, and $f(x)$ are also replaced by piecewise polynomials. In that case, the integrals in (53) can be evaluated exactly.

In the form (53), the problem is still not suitable for computer implementation, since it is not obviously of a finite nature. The latter can be achieved by constructing a finite-dimensional basis of the space V_h. To this end, the piecewise linear hat functions ϕ_i are introduced. These are defined by

$$\phi_i(x) = \begin{cases} \dfrac{x - x_{i-1}}{x_i - x_{i-1}}, & x \in \overline{\Omega}_i \\ \dfrac{x - x_i}{x_{i+1} - x_i}, & x \in \overline{\Omega}_{i+1} \\ 0, & \text{otherwise} \end{cases} \tag{54}$$

for $i = 1, \ldots, N-1$. It is easy to see that

$$\phi_i(x_j) = \begin{cases} 1 & \text{if} \quad j = i \\ 0 & \text{if} \quad j \neq i \end{cases}$$

and that $\phi_i \in V_h$. In figure 5 one of these functions is shown. Note that ϕ_0 and ϕ_N are only half hats.

Moreover, it is not hard to see that dim $V_h = N-1$ and that the ϕ_i are linearly independent, therefore $\{\varphi_i\}_1^{N-1}$ is a basis for V_h. Thus every function $u_h \in V_h$ can be written in the form

$$u_h(x) = \sum_{i=1}^{N-1} u_h(x_i)\phi_i(x).$$

Also, if the equations in (53) hold for each $v_h = \phi_j$, $j = 1, \ldots, N-1$, then they hold for all functions $v_h \in V_h$. It follows that the problem (53) can be reformulated as follows: find numbers $u_1, \ldots, u_{N-1}$ such that, for all $j \in \{1, \ldots, N-1\}$,

$$\int_0^1 \left\{ a(x) \sum_{i=1}^{N-1} u_i \phi_i'(x) \phi_j'(x) + b(x) \sum_{i=1}^{N-1} u_i \phi_i(x) \phi_j(x) \right\} \mathrm{d}x = \int_0^1 f(x) \phi_j(x) \, \mathrm{d}x, \tag{55}$$

where $u_0 = u_N = 0$. This amounts to solving a system of $N-1$ linear algebraic equations for $N-1$ unknowns, which is a standard computational task.

It is important to note that the integrations to be performed in (55) do not extend over the entire interval $\overline{\Omega}$. Since the basis functions ϕ_i are non-zero on only the two subintervals $\overline{\Omega}_i$ and $\overline{\Omega}_{i+1}$, the integrals have to be evaluated on only these two neighbouring subintervals. The system (55) can therefore be simplified to

$$\begin{aligned} u_{j-1} & \int_{x_{j-1}}^{x_j} \{a(x)\phi_{j-1}'(x)\phi_j'(x) + b(x)\phi_{j-1}(x)\phi_j(x)\} \, \mathrm{d}x \\ & + u_j \int_{x_{j-1}}^{x_{j+1}} \{a(x)(\phi_j'(x))^2 + b(x)(\phi_j(x))^2\} \, \mathrm{d}x \\ & + u_{j+1} \int_{x_j}^{x_{j+1}} \{a(x)\phi_{j+1}'(x)\phi_j'(x) + b(x)\phi_{j+1}(x)\phi_j(x)\} \, \mathrm{d}x \\ = & \int_{x_{j-1}}^{x_{j+1}} f(x)\phi_j(x) \, \mathrm{d}x, \qquad \text{for} \quad j = 1, \ldots, N-1. \end{aligned}$$

In compact form, this can be represented as a system of the form (44), where $\boldsymbol{u} = (u_1, \ldots, u_{N-1})^T$, A is a tridiagonal matrix with non-zero entries for $1 \leqslant j \leqslant N-1$ given by

$$\begin{cases} a_{j,j-1} = \displaystyle\int_{x_{j-1}}^{x_j} \{a(x)\phi_{j-1}'(x)\phi_j'(x) + b(x)\phi_{j-1}(x)\phi_j(x)\} \, \mathrm{d}x, \\ a_{j,j} = \displaystyle\int_{x_{j-1}}^{x_{j+1}} \{a(x)(\phi_j'(x))^2 + b(x)(\phi_j(x))^2\} \, \mathrm{d}x, \\ a_{j,j+1} = \displaystyle\int_{x_j}^{x_{j+1}} \{a(x)\phi_{j+1}'(x)\phi_j'(x) + b(x)\phi_{j+1}(x)\phi_j(x)\} \, \mathrm{d}x, \end{cases} \tag{56}$$

and $\boldsymbol{b}$ is the vector of the right-hand sides with the entries

$$b_j = \int_{x_{j-1}}^{x_{j+1}} f(x)\phi_j(x) \, \mathrm{d}x. \tag{57}$$

Note that if non-zero values happen to be specified for u_0 and u_N, then b_1 and b_{N-1} must be corrected by the terms

$$-u_0 \int_{x_0}^{x_1} \{a(x)\phi_0'(x)\phi_1'(x) + b(x)\phi_0(x)\phi_1(x)\} \, \mathrm{d}x,$$

and

$$-u_N \int_{x_{N-1}}^{x_N} \{a(x)\phi_N'(x)\phi_{N-1}'(x) + b(x)\phi_N(x)\phi_{N-1}(x)\} \, \mathrm{d}x,$$

respectively.

The linear system can be solved provided that the coefficient matrix A is not singular. In section 3.4 this issue will be analysed in more detail, where it will be shown what the implications are for practical problems.

As in the case of finite difference methods, the crucial question is whether the discrete solution u_h converges to the solution u of the original problem in the form either (45) or (47)). It is easy to verify that the error $u - u_h$ satisfies, for all $v_h \in V_h$,

$$\int_0^1 \{a(x)(u-u_h)'(x)v_h'(x) + b(x)(u-u_h)(x)v_h(x)\}\,\mathrm{d}x = 0.$$

We assume that both a and b are bounded above in Ω by a positive constant and we let $u_I = \sum_{i=1}^{N-1} u(x_i)\phi_i(x)$ be the V_h-interpolant of the exact solution u. For any $v_h \in V_h$, subtracting $\int_0^1 (u_I' v_h' + b u_I v_h)\mathrm{d}x$ from both sides of (53) and using the above equality we have

$$\int_0^1 \{a(u_h - u_I)'v_h' + b(u_h - u_I)v_h\}\,\mathrm{d}x = \int_0^1 \{a(u-u_I)'v_h' + b(u-u_I)v_h\}\,\mathrm{d}x.$$

Let $C > 0$ be a generic constant depending only on Ω. Note that $u_h - u_I \in V_h$. Then, choosing $v_h = u_h - u_I$ in the above and using (49) ($C = \alpha$), the boundedness of a and b and the Cauchy–Schwarz inequality we see that

$$\begin{aligned}
\alpha\|u_h - u_I\|_1^2 &\leqslant C\int_0^1 \{a(u'-u_I')(u_h'-u_I') + b(u-u_I)(u_h'-u_I')\}\,\mathrm{d}x \\
&\leqslant C\Bigg[\bigg(\int_0^1 (u'-u_I')^2\,\mathrm{d}x\bigg)^{1/2}\bigg(\int_0^1 (u_h'-u_I')^2\,\mathrm{d}x\bigg)^{1/2} \\
&\quad + \bigg(\int_0^1 (u-u_I)^2\,\mathrm{d}x\bigg)^{1/2}\bigg(\int_0^1 (u_h-u_I)^2\,\mathrm{d}x\bigg)^{1/2}\Bigg] \\
&= C(\|u'-u_I'\|\|u_h'-u_I'\| + \|u-u_I\|\|u_h-u_I\|) \\
&\leqslant C(\|u'-u_I'\| + \|u-u_I\|)(\|u_h'-u_I\| + \|u_h-u_I\|) \\
&\leqslant C\|u-u_I\|_1\|u_h-u_I\|_1.
\end{aligned}$$

This gives us

$$\|u_h - u_I\|_1 \leqslant \frac{C}{\alpha}\|u-u_I\|_1.$$

From this and the triangle inequality we have

$$\|u-u_h\|_1 \leqslant \|u-u_I\|_1 + \|u_h-u_I\|_1 \leqslant \frac{C}{\alpha}\|u-u_I\|_1.$$

Since V_h is the space of all piecewise linear functions on Ω, u_I is the piecewise linear interpolant (that is, a linear approximation on each element) of u. Using a Taylor expansion we can show that

$$|u(x) - u_I(x)| \leqslant \frac{h^2}{8}\max_{0\leqslant y\leqslant 1}|u''(y)|, \qquad \text{for all} \quad x \in \Omega$$

and also that

$$|u'(x) - u_I'(x)| \leqslant h\max_{0\leqslant y\leqslant 1}|u''(y)|,$$

where h is the mesh parameter defined before. A simple exercise then leads to

$$\|u-u_I\|_1 \leqslant \frac{C}{\alpha}h,$$

and so

$$\|u-u_h\|_1 \leqslant \frac{C}{\alpha}h. \tag{58}$$

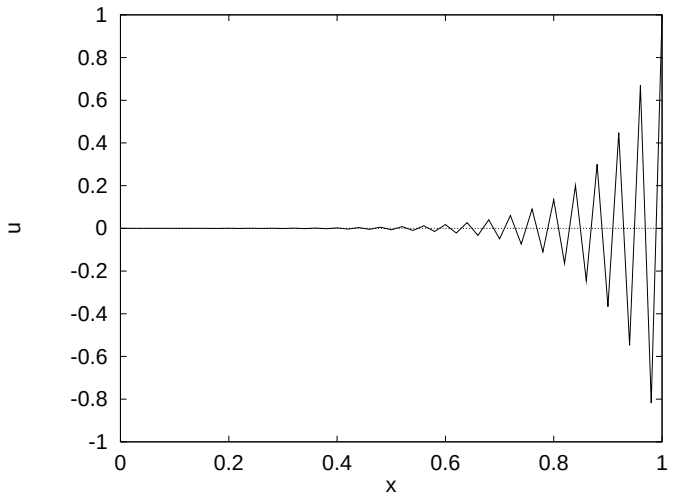

Figure 6. Numerical solution of (59) using the standard finite element method.

With considerable extra effort, estimates in other norms can be obtained.

From the above analysis we see that the constant C in the error bound (58) depends on the second derivative u'' of the exact solution u. For some problems, u'' may be large in some subsets of Ω. For example, solutions to the semiconductor device equations change very rapidly in thin subregions of Ω containing the p–n junctions due to the abrupt change of the doping function $D(x)$ (these subregions are called depletion regions). Thus these solutions have second derivatives of large magnitude. Furthermore, the lower bound α of $a(x)$ can also be close to zero. The combined effect of large u'' and small α will make the right-hand side of (58) large, unless h is impractically small. This shows that the standard finite element method does not work well if the behaviour of u is almost singular. This will be demonstrated further in the next section. Therefore, we need to either reformulate the discrete problem or mesh the solution region Ω judiciously, or do both. We now discuss these options in more detail.

3.2. Some non-standard ordinary finite element methods in one dimension

The standard finite element method has been described for problems with the conservation form (45), but it can also be formulated in an analogous way for other types of problem. Unfortunately, the method turns out to be unsatisfactory for convection-diffusion problems if convection dominates diffusion in the physical problem. In such cases the numerical solution may exhibit unphysical spurious oscillations. An example of this is the linear problem

$$\begin{cases} -\varepsilon u''(x) + a u'(x) = 0, & x \in \Omega \\ u(0) = 0, \qquad u(1) = 1 \end{cases} \tag{59}$$

with constant coefficients ε and a. This problem is singularly perturbed if ε is small compared with a, that is if convection dominates diffusion. In that case, the solution exhibits boundary layer behaviour, meaning that the solution changes rapidly within a thin region of a boundary point. The exact solution of (59) is

$$u(x) = \frac{\exp(ax/\varepsilon) - 1}{\exp(a/\varepsilon) - 1}.$$

It it easy to see that, for small ε, this varies rapidly close to the boundary point $x = 1$. It turns out that this behaviour can be captured by a standard finite element method only if the mesh spacing is ridiculously small. In figure 6 the approximate solution is shown for a uniform mesh spacing $h = 0.02$, in the case that $\varepsilon = 0.001$ and $a = 1$. It is obvious that the numerical solution is not a good approximation of the exact solution.

To derive a variational formulation, first introduce the new dependent variable

$$\hat{u}(x) = u(x) - x,$$

so that the problem for $\hat{u}$ has homogeneous boundary conditions, and is given by

$$\begin{cases} -\varepsilon\hat{u}''(x) + a\hat{u}'(x) = -a, & x \in (0,1), \\ \hat{u}(0) = 0, \qquad \hat{u}(1) = 0, \end{cases} \tag{60}$$

As before we multiply the differential equation in (60) by a smooth test function v, and integrate over the solution domain $\overline{\Omega}$ to get

$$\int_0^1 (-\varepsilon\hat{u}'' + a\hat{u}')v = -\int_0^1 av.$$

Now use integration by parts on the left-hand side to relax the smoothness condition on the function $\hat{u}$. We note that in this case there is a choice, since integration by parts can be applied either to both terms or to just the second-order term. Choosing the latter option leads to

$$[-\varepsilon\hat{u}'v]_0^1 + \int_0^1 (\varepsilon\hat{u}'v' + a\hat{u}'v) = -\int_0^1 av,$$

where, for conciseness, explicit mention of the variable x has been omitted from the notation. Taking $V = H_0^1(\Omega)$ the variational form can then be written in the form

$$\begin{cases} \text{Find } \hat{u} \in V \qquad \text{such that for all} \quad v \in V \\ \displaystyle\int_0^1 (\varepsilon\hat{u}'v' + a\hat{u}'v) = -\int_0^1 av. \end{cases} \tag{61}$$

To obtain a discrete formulation, we choose the same piecewise linear test and trial functions as in section 3.1. The result, after transforming back to the original function u, is

$$\begin{cases} -\dfrac{\varepsilon}{h}(u_{j+1} - 2u_j + u_{j-1}) + \dfrac{a}{2}(u_{j+1} - u_{j-1}) = 0, & 1 \leqslant j \leqslant N-1, \\ u_0 = 0, \qquad u_N = 1. \end{cases} \tag{62}$$

The solution of this discrete problem is

$$u_j = \frac{\left(\frac{1+\beta}{1-\beta}\right)^j - 1}{\left(\frac{1+\beta}{1-\beta}\right)^N - 1},$$

where

$$\beta = \frac{ah}{2\varepsilon} \tag{63}$$

is called the cell Reynolds number. Clearly, the discrete solution is non-oscillatory at all mesh points if and only if the following condition is satisfied:

$$\beta < 1. \tag{64}$$

This is not the case in the example for which the solution is displayed in figure 6, whence the oscillatory behaviour. Note that the condition (64) is extremely restrictive, since it implies that

the mesh width must satisfy $h < \frac{2\varepsilon}{a} = 2 \times 10^{-3}$ for this particular problem. Therefore more than 500 nodes are needed to obtain an acceptable solution.

Just as for finite difference methods, the above shortcoming of standard finite element methods for problem (59) can be overcome by introducing upwinding techniques. These involve the introduction of an additional diffusive term into the discrete equations. In fact, this is equivalent to replacing the central difference term for the first order derivative in (62) by a one-sided difference term. The difference between the two discrete equations is equal to a diffusive term

$$a\frac{u_{i+1} - u_{i-1}}{2} - a(u_i - u_{i-1}) = \frac{a}{2}(u_{i+1} - 2u_i + u_{i-1}) = \beta\frac{\varepsilon}{h}(u_{i+1} - 2u_i + u_{i-1}).$$

This means that the diffusion coefficient in (62) is effectively multiplied by $1 + \beta$. For finite element methods, the upwinding technique must be translated into the choice of different test functions, as is shown in the following (see also [37, ch 7]).

Among the first to describe upwinding in the context of finite element methods were Christie, Griffiths, Heinrich, Huyakorn, Miller, Mitchell and Zienkiewicz [38–40]. The ideas were developed further by many other researchers, see for example [41, 42]. It should be noted that an unsatisfactory feature of upwinding techniques is that they essentially change the original problem, because they introduce an unphysical artificial diffusion term into the equations.

Up to now the trial functions u_h and the test functions v_h have always been chosen from the same space V_h. But it is often better to choose the trial functions to be in a space U_h and the test functions to be in some other space V_h. If $U_h = V_h$ the method is called a Bubnov–Galerkin method, otherwise it is a Petrov–Galerkin method.

To obtain an upwind finite element method for problem (59) consider the weak formulation (61) again, but in this case use a Petrov–Galerkin discretization defined as follows. Take the same piecewise linear trial space as before but use a modified test space of piecewise quadratic polynomials, continuous on the whole of $\overline{\Omega}$. A basis for this test space is $\{\psi_j\}_0^N$ where

$$\psi_j(x) = \phi_j(x) + \alpha\sigma_j(x), \tag{65}$$

with

$$\sigma_j(x) = \begin{cases} \dfrac{3(x - x_{j-1})(x_j - x)}{(x_j - x_{j-1})^2}, & x \in \overline{\Omega}_j \\ -\dfrac{3(x - x_j)(x_{j+1} - x)}{(x_{j+1} - x_j)^2}, & x \in \overline{\Omega}_{j+1} \\ 0 & \text{otherwise} \end{cases}$$

and α is a suitable constant. In figure 7 the functions ϕ_j, ψ_j and σ_j are shown schematically.

As the parameter α, which controls the influence of the antisymmetric function σ_j, increases from zero the contribution from the subinterval $\overline{\Omega}_j$ increases and the contribution from $\overline{\Omega}_{j+1}$ decreases. Since σ_j is antisymmetric about the point x_j, there is no contribution to the discretized diffusive term.

Using a uniform grid spacing h, and scaling the discretized equations by an appropriate constant factor, the discretized problem takes the form

$$\begin{cases} -[1 + \alpha\beta](u_{i+1} - 2u_i + u_{i-1}) + \beta(u_{i+1} - u_{i-1}) = 0, & 1 \leqslant i \leqslant N - 1, \\ u_0 = 0, \qquad u_N = 1. \end{cases} \tag{66}$$

The solution of this discrete problem is

$$u_i = \frac{\left(\frac{1+\beta(\alpha+1)}{1+\beta(\alpha-1)}\right)^i - 1}{\left(\frac{1+\beta(\alpha+1)}{1+\beta(\alpha-1)}\right)^N - 1}.$$

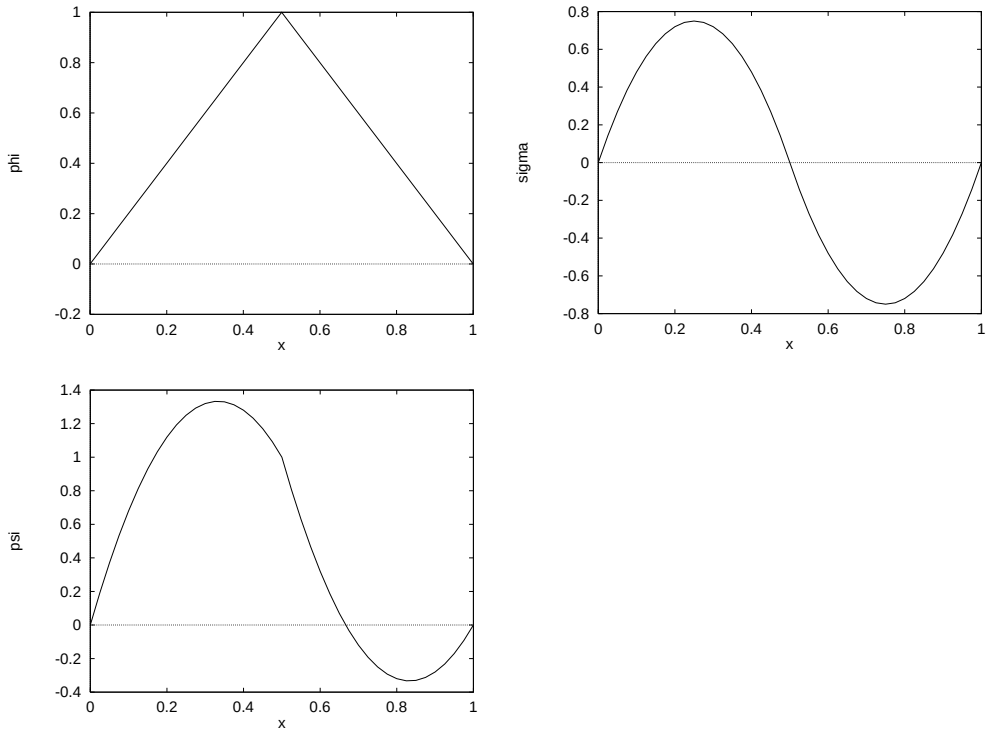

Figure 7. Test functions for the upwind finite element method ($\alpha = 1$).

Clearly, this solution is non-oscillatory if and only if

$$\text{either} \qquad \alpha \geqslant 1 \qquad \text{or} \qquad \alpha < 1, \quad \beta < \frac{1}{1-\alpha}. \tag{67}$$

It is not hard to show that the transformation of the dependent variable

$$u = v \exp\left(\frac{1}{\varepsilon}\int a(x)\,\mathrm{d}x\right)$$

changes the convection-diffusion equation in (59) into a self-adjoint equation in v of the form (45). Thus, the standard finite element method is also not satisfactory for (45) if the coefficient $a(x)$ is singularly behaved. This shows that more sophisticated finite element methods are needed in order to handle such problems with boundary and/or interior layers.

Returning now to problems of the form (45), it should be noted that there are many possible alternatives to the variational formulation of the previous section. One such is obtained using the idea of inverse averages described, for example, in [43]. This is also called harmonic averages in [44] where it is used for approximating discontinuous coefficients in some eigenvalue problems. We associate with each node x_i, $i = 0, \ldots, N$, a Dirichlet subinterval $d_i = [x_{i-1/2}, x_{i+1/2}]$ where for $1 \leqslant i \leqslant N-1$, $x_{i-1/2} = \frac{1}{2}(x_{i-1} + x_i)$, $x_{i+1/2} = \frac{1}{2}(x_i + x_{i+1})$, and $x_{-1/2} = x_0$, $x_{N+1/2} = x_N$. Then use integration by parts on each such subinterval to obtain

$$\int_0^1 (au')'v = \sum_{i=0}^{N}\int_{x_{i-1/2}}^{x_{i+1/2}} (au')'v = \sum_{i=0}^{N}\left([au'v]_{x_{i-1/2}}^{x_{i+1/2}} - \int_{x_{i-1/2}}^{x_{i+1/2}} au'v'\right).$$

A possible variational form of the differential equation in (45) is thus

$$\sum_{i=0}^{N}\left(-[au'v]_{x_{i-1/2}}^{x_{i+1/2}}+\int_{x_{i-1/2}}^{x_{i+1/2}}(au'v'+buv)\right)=\sum_{i=0}^{N}\int_{x_{i-1/2}}^{x_{i+1/2}}fv.$$

Assuming that the discrete test functions v_h are constant on each d_i, this suggests the following Petrov–Galerkin discretization

$$\begin{cases}\text{Find } u_h\in U_h & \text{such that for all} \quad v_h\in V_h\\ \displaystyle\sum_{i=0}^{N}\left((au_h'v_h)(x_{i-1/2})-(au_h'v_h)(x_{i+1/2})+\int_{x_{i-1/2}}^{x_{i+1/2}}bu_hv_h\right)=\sum_{i=0}^{N}\int_{x_{i-1/2}}^{x_{i+1/2}}fv_h.\end{cases}\tag{68}$$

The test space V_h is now chosen to be the space of piecewise constant functions (piecewise polynomials of degree zero) on the Dirichlet subintervals d_i with the basis $\{\psi_i\}_1^{N-1}$, where ψ_i is the characteristic function of d_i, that is

$$\psi_i(x)=\begin{cases}1 & \text{if} \quad x\in d_i\\ 0 & \text{otherwise.}\end{cases}$$

Then the equation in (68) becomes

$$\sum_{i=0}^{N}\left((au_h')(x_{i-1/2})-(au_h')(x_{i+1/2})+\int_{x_{i-1/2}}^{x_{i+1/2}}bu_h\right)v_h(x_i)=\sum_{i=0}^{N}\left(\int_{x_{i-1/2}}^{x_{i+1/2}}f\right)v_h(x_i).$$

Now approximating the integrals by the midpoint quadrature rule gives

$$\sum_{i=0}^{N}((au_h')(x_{i-1/2})-(au_h')(x_{i+1/2})+bu_h(x_i)|d_i|)v_h(x_i)=\sum_{i=0}^{N}f(x_i)|d_i|v_h(x_i),\tag{69}$$

where the length of d_i is $|d_i|=x_{i+1/2}-x_{i-1/2}=(x_{i+1}-x_{i-1})/2$. In (69) choose $v_h=\psi_j$ for $j=1,\ldots,N-1$. Since $\psi_j(x_i)=\delta_{ji}$ this yields the system of equations

$$-2\frac{(au_h')(x_{j+1/2})-(au_h')(x_{j-1/2})}{x_{j+1}-x_{j-1}}+(bu_h)(x_j)=f(x_j)\qquad j=1,\ldots,N-1.\tag{70}$$

Now choose the trial space U_h to be a non-piecewise polynomial space with the basis $\{\varphi_i\}_1^{N-1}$, where for each fixed i, $1\leqslant i\leqslant N-1$, the function ϕ_i is defined on each subinterval Ω_i, $1\leqslant j\leqslant N$, to be the solution of

$$(a(x)\phi_i'(x))'=0,\qquad x\in\Omega_j$$

with the boundary conditions

$$\phi_i(x_k)=\delta_{ik},\qquad 0\leqslant k\leqslant N.$$

Solving these two-point boundary value problems on each Ω_j explicitly gives the expressions

$$a(x)\phi_i'(x)=\begin{cases}\dfrac{\overline{a}_{i-1/2}}{x_i-x_{i-1}} & \text{if } x\in\overline{\Omega}_i\\ -\dfrac{\overline{a}_{i+1/2}}{x_{i+1}-x_i} & \text{if } x\in\overline{\Omega}_{i+1}\\ 0 & \text{otherwise}\end{cases}$$

where the inverse averages

$$\overline{a}_{k-1/2}=\left(\frac{1}{x_k-x_{k-1}}\int_{x_{k-1}}^{x_k}a^{-1}(s)\,\mathrm{d}s\right)^{-1},\qquad 1\leqslant k\leqslant N$$

have been introduced. Since $u_h(x) = \sum_{j=1}^{N-1} u_i \phi_i(x)$, where $u_i = u_h(x_i)$, it follows from these expressions that

$$\begin{aligned}(au_h')(x_{j-1/2}) &= u_{j-1}(a\phi_{j-1}')(x_{j-1/2}) + u_j(a\phi_j')(x_{j-1/2}) \\ &= \overline{a}_{j-1/2}\frac{u_j - u_{j-1}}{x_j - x_{j-1}}\end{aligned}$$

and similarly that

$$(au_h')(x_{j+1/2}) = \overline{a}_{j+1/2}\frac{u_{j+1} - u_j}{x_{j+1} - x_{j-1}}.$$

Substituting these expressions into (70) gives

$$-\frac{2}{x_{j+1} - x_{j-1}}\left(\overline{a}_{j+1/2}\frac{u_{j+1} - u_j}{x_{j+1} - x_j} - \overline{a}_{j-1/2}\frac{u_j - u_{j-1}}{x_j - x_{j-1}}\right) + b_j u_j = f_j, \qquad (71)$$

for $j = 1, \ldots, N-1$. Some properties of this system of equations will be discussed later in section 3.4.

Hemker [45] also has developed fitted finite element methods, but in a different way than that described in the foregoing. He uses the Green function of the problem (59) to develop a system of trial and test functions having an exponential character. This is natural in view of the nature of the solution of the continuous problem. The approach is not easily extended to the two-dimensional case, since it is difficult to find a suitable set of trial and test functions.

3.3. Extension to higher dimensions

In two dimensions the analogous problem to the two-point boundary value problem (45) is

$$\begin{cases} -\nabla \cdot (a(\boldsymbol{x})\nabla u(\boldsymbol{x})) + b(\boldsymbol{x})u(\boldsymbol{x}) = f(\boldsymbol{x}), & \boldsymbol{x} \in \Omega, \\ u(\boldsymbol{x}) = 0, \quad \boldsymbol{x} \in \Gamma_1, \qquad \nabla u(\boldsymbol{x}) \cdot \boldsymbol{n} = 0, \quad \boldsymbol{x} \in \Gamma_2 \end{cases} \qquad (72)$$

where $\boldsymbol{x} = (x_1, x_2)$, Ω is a bounded domain, $\boldsymbol{n}$ denotes the outward normal direction on the boundary and Γ_1 and Γ_2 are boundary components such that $\Gamma_1 \cap \Gamma_2 = \emptyset$ and $\Gamma = \overline{\Gamma}_1 \cup \overline{\Gamma}_2$ is the boundary of Ω. We comment that although the equation considered here is in self-adjoint form, it can also be regarded as a variant of a singularly perturbed convection-diffusion equation of the form

$$-\nabla \cdot (\varepsilon \nabla u - \boldsymbol{c}u) + bu = f, \qquad (73)$$

provided that $\boldsymbol{c}$ is irrotational ($\boldsymbol{c}$ is said to be irrotational if there exists a scalar function ψ such that $\boldsymbol{c} = \nabla\psi$). In this case, it is easy to show that the substitution

$$u = v \exp(\psi/\varepsilon)$$

transforms this convection-diffusion equation into the self-adjoint form in (72).

To obtain the standard variational formulation of (72), multiply the equations by a smooth test function v and integrate over the whole of Ω

$$-\int_\Omega \nabla \cdot (a(\boldsymbol{x})\nabla u(\boldsymbol{x}))\,\mathrm{d}x + \int_\Omega b(\boldsymbol{x})u(\boldsymbol{x})v(\boldsymbol{x})\,\mathrm{d}x = \int_\Omega f(\boldsymbol{x})v(\boldsymbol{x})\,\mathrm{d}x.$$

Applying Green's formula to the first integral gives

$$\begin{aligned}-\int_\Gamma a(\boldsymbol{x})\frac{\mathrm{d}u}{\mathrm{d}n}(\boldsymbol{x})v(\boldsymbol{x})\,\mathrm{d}x + \int_\Omega a(\boldsymbol{x})\nabla u(\boldsymbol{x}) \cdot \nabla v(\boldsymbol{x})\,\mathrm{d}x + \int_\Omega b(\boldsymbol{x})u(\boldsymbol{x})v(\boldsymbol{x})\,\mathrm{d}x \\ = \int_\Omega f(\boldsymbol{x})v(\boldsymbol{x})\,\mathrm{d}x.\end{aligned}$$

Assuming now that the test functions all vanish on Γ_1 we then obtain

$$\int_\Omega (a(\boldsymbol{x})\nabla u(\boldsymbol{x}) \cdot \nabla v(\boldsymbol{x}) + b(\boldsymbol{x})u(\boldsymbol{x})v(\boldsymbol{x}))\,\mathrm{d}x = \int_\Omega f(\boldsymbol{x})v(\boldsymbol{x})\,\mathrm{d}x.$$

Choosing suitable spaces of trial functions U and test functions V then gives the variational formulation

$$\begin{cases} \text{Find } u \in U \qquad \text{such that for all} \quad v \in V \\ \displaystyle\int_\Omega (a\nabla u \cdot \nabla v + buv) = \int_\Omega fv. \end{cases} \tag{74}$$

If we assume that $a(\boldsymbol{x}) \geqslant \alpha > 0$ and $b(\boldsymbol{x}) \geqslant 0$, then the same argument as that for the proof of (49) and the two-dimensional analogue of the Poincaré–Friedrichs inequality (51) shows that the integral on the left-hand side of (74) is V-elliptic. It follows that this variational problem has a unique solution.

This problem can now be discretized in the routine way by choosing subspaces U_h and V_h of U and V, and defining the discrete problem

$$\begin{cases} \text{Find } u_h \in U_h \qquad \text{such that for all} \quad v_h \in V_h \\ \displaystyle\int_\Omega (a\nabla u_h \cdot \nabla v_h + bu_h v_h) = \int_\Omega f v_h. \end{cases} \tag{75}$$

For simplicity it is assumed here that Ω is a rectangle, which has been subdivided into elements which may be either rectangles or triangles. A standard finite element method is obtained in the former case by choosing both U_h and V_h to be the space of piecewise polynomials which are bilinear on each rectangle and continuous on all of Ω, while in the latter case both U_h and V_h can be chosen to be the space of piecewise polynomials which are linear on each triangle and continuous on Ω. A linear system equivalent to the Bubnov–Galerkin problem (75) is then obtained by introducing a suitable basis for $U_h = V_h$, choosing v_h to be each of the basis elements successively and writing u_h as a linear combination of the basis elements. This leads to a linear system of the form (44), where $\boldsymbol{u} = (u_1, \ldots, u_N)^T$ and $\dim U_h = \dim V_h = N$.

The upwinding technique for the convection-diffusion problem, described above in one dimension, can be extended in some cases to two-dimensional problems (see, for example, [46, 47]). However, it is not obvious how to do this unless the elements are rectangles. Furthermore, in two dimensions this approach generates algebraic equations, each of which involves values of the unknown at nine mesh points in contrast to the five occurring in upwind finite difference methods. Just as for standard finite elements on rectangles, the trial space U_h is chosen to be the space of piecewise polynomials which are bilinear on each rectangle and continuous on all of Ω. A basis for this space is composed of the set of all pyramid functions $\phi_{ij}(x_1, x_2) = \phi_i(x_1)\phi_j(x_2)$ where $\phi_i(x_1)$ is the one-dimensional hat function for the ith node in the x_1-direction and $\phi_j(x_2)$ is the corresponding function for the jth node in the x_2-direction. Thus, there is a basis function corresponding to each node of the rectangular mesh. The upwinding is achieved by choosing a different test space V_h, and so the method is a Petrov–Galerkin discretization. In analogy to the one-dimensional case and the above construction of U_h, the basis function ψ_{ij} in V_h for the (i, j)th node is defined as the tensor product $\psi_{ij}(x_1, x_2) = \psi_i(x_1)\psi_j(x_2)$ where $\psi_i(x_1)$, $\psi_j(x_2)$ are the piecewise quadratic functions defined in (65) for one-dimensional problems. Hughes and Brooks [48] also proposed a triangular streamline diffusion or streamline upwind Petrov–Galerkin method for (73). In their method, the solution space is chosen to be the conventional piecewise polynomial finite element space, but the equation

$$\sum_t \delta_t \int_t [-\nabla \cdot (\varepsilon \nabla u_h - \boldsymbol{c}u) + bu_h]\boldsymbol{c} \cdot \nabla v_h = \sum_t \delta_t \int_t f\boldsymbol{c} \cdot \nabla v_h$$

is added to the discrete variational form associated with (73), where the sums are taken for all elements of a mesh and δ_t is called the streamline diffusion parameter. For some choices of δ_t, the method is stable and convergent.

On the other hand, a non-standard variational formulation involving inverse averages, analogous to that discussed in section 3.2 may be obtained as follows. Let E_h be a decomposition of Ω into elements, which may be either rectangles or triangles, and with each vertex x_i of E_h associate the Dirichlet tile d_i. This is defined to be the set of all points of $\overline{\Omega}$ which are closer to x_i than to any other vertex x_j of E_h. If the elements consist of rectangles then the Dirichlet tiles are also rectangles, whereas if they are triangles then the Dirichlet tiles are polygons.

The corresponding non-standard variational formulation is now obtained by writing

$$\int_\Omega \nabla \cdot (a\nabla u)v = \sum_{i=0}^{N} \int_{d_i} \nabla \cdot (a\nabla u)v,$$

where N is the number of Dirichlet tiles covering $\overline{\Omega}$. Using Green's formula separately on each integral on the right-hand side gives

$$\int_{d_i} \nabla \cdot (a\nabla u)v = \int_{\partial d_i} a\nabla u \cdot \boldsymbol{n}_j v - \int_{d_i} a\nabla u \cdot \nabla v,$$

where ∂d_i is the boundary of the tile d_i and $\boldsymbol{n}_j$ is the unit normal to ∂d_j. This suggests the following variational formulation

$$\begin{cases} \text{Find } u \in U \qquad \text{such that for all} \quad v \in V \\ \displaystyle\sum_{i=0}^{N} \left[-\int_{\partial d_i} a\nabla u \cdot \boldsymbol{n}_j v + \int_{d_i} a\nabla u \cdot \nabla v \right] + \int_\Omega buv = \int_\Omega fv. \end{cases} \tag{76}$$

The discrete version of this is

$$\begin{cases} \text{Find } u_h \in U_h \qquad \text{such that for all} \quad v_h \in V_h \\ \displaystyle\sum_{i=0}^{N} \left[-\int_{\partial d_i} a\nabla u_h \cdot \boldsymbol{n}_j v_h + \int_{d_i} a\nabla u_h \cdot \nabla v_h \right] + \int_\Omega bu_h v_h = \int_\Omega f v_h. \end{cases} \tag{77}$$

A Petrov–Galerkin discretization is then obtained as follows. The test space V_h is defined to be the space of piecewise constant functions, constant on each Dirichlet tile d_i. Thus V_h is spanned by the basis functions $\{\psi_i\}_1^N$ where

$$\psi_i(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \boldsymbol{x} \in d_i \\ 0 & \text{otherwise} \end{cases} \tag{78}$$

is the characteristic function of d_i. Then, in (77) v_h is chosen as ψ_j successively for each $j = 1, \ldots, N$, which gives the N equations

$$-\int_{\partial d_j} a\nabla u_h \cdot \boldsymbol{n}_j + \int_{d_j} bu_h = \int_{d_j} f, \qquad j = 1, \ldots, N.$$

The integrals on each tile are approximated by the one-point quadrature rule giving the simpler equations

$$-\int_{\partial d_j} a\nabla u_h \cdot \boldsymbol{n}_j + b_j u_j |d_j| = f_j |d_j|, \qquad j = 1, \ldots, N, \tag{79}$$

where u_j, f_j denote the function values at the vertex $\boldsymbol{x}_j$ and $|d_j|$ is the area of the tile d_j.

So far nothing has been said about the trial space U_h. This is now defined to be a non-piecewise polynomial space as follows. Let γ_{ij} denote the undirected line segment forming

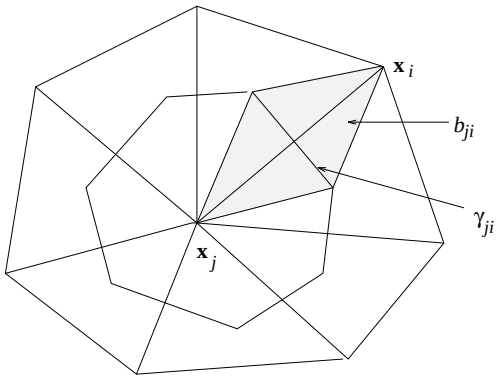

Figure 8. The box b_{ji}.

the common edge of any two tiles d_i, d_j with a common edge. In this case the vertices $\boldsymbol{x}_j$ and $\boldsymbol{x}_i$ are said to be neighbours. Let I_j denote the index set for the neighbours of $\boldsymbol{x}_j$, that is the set of all vertices $\boldsymbol{x}_i$ such that the intersection of the tile d_i with d_j is an edge γ_{ij}. When γ_{ij} is viewed as part of the boundary ∂d_i it is denoted by ∂d_{ij}, and as part of ∂d_j by ∂d_{ji}.

Let b_{ji} denote the diamond-shaped box obtained by joining the two end points of γ_{ij} to both $\boldsymbol{x}_j$ and $\boldsymbol{x}_i$, as shown in figure 8.

A function ϕ_{ji} corresponding to the vertex $\boldsymbol{x}_j$ is now defined on b_{ji}. On the line joining $\boldsymbol{x}_j$ to $\boldsymbol{x}_i$ define ϕ_{ji} to be the solution of the two-point boundary value problem

$$\frac{\mathrm{d}}{\mathrm{d}s}\left(a\frac{\mathrm{d}\phi}{\mathrm{d}s}\right) = 0,$$
$$\phi(\boldsymbol{x}_j) = 1, \qquad \phi(\boldsymbol{x}_i) = 0,$$

where $\frac{\mathrm{d}}{\mathrm{d}s}$ denotes differentiation in the direction from $\boldsymbol{x}_j$ to $\boldsymbol{x}_i$. On this line an explicit formula for ϕ_{ji} is clearly

$$\phi_{ji}(\boldsymbol{x}) = \frac{\overline{a}_{ji}}{|\boldsymbol{x}_i - \boldsymbol{x}_j|}\int_{\boldsymbol{x}}^{\boldsymbol{x}_i} a^{-1}(s)\,\mathrm{d}s,$$

where

$$\overline{a}_{ji} = \left(\frac{1}{|\boldsymbol{x}_i - \boldsymbol{x}_j|}\int_{\boldsymbol{x}_j}^{\boldsymbol{x}_i} a^{-1}(s)\,\mathrm{d}s\right)^{-1} \tag{80}$$

is an inverse average and $|\boldsymbol{x}_i - \boldsymbol{x}_j|$ denotes the distance between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. The function ϕ_{ji} is defined on the rest of b_{ji} by taking it to be constant on lines perpendicular to the line joining $\boldsymbol{x}_j$ to $\boldsymbol{x}_i$. The complete basis function ϕ_j, corresponding to the vertex $\boldsymbol{x}_j$, is now defined by

$$\phi_j(\boldsymbol{x}) = \begin{cases} \phi_{ji}(\boldsymbol{x}) & \text{if } \boldsymbol{x} \in b_{ji} \quad \text{and} \quad i \in I_j, \\ 0 & \text{otherwise.} \end{cases}$$

Then U_h is defined to be the non-piecewise polynomial space spanned by the basis functions $\{\phi_j\}_1^N$.

The contour integral in the above equations can now be approximated. Since

$$u_h = \sum_{i=1}^{N} u_i\phi_i$$

it follows that

$$\int_{\partial d_j} a\nabla u_h \cdot \boldsymbol{n}_j = \sum_{i=1}^{N} \left(\int_{\partial d_j} a\nabla \phi_i \cdot \boldsymbol{n}_j \right) u_i. \tag{81}$$

Note now that $\nabla \phi_i \cdot \boldsymbol{n}_j$ is zero everywhere on ∂d_j unless either (a) $i = j$, in which case it is non-zero on the whole of $\partial d_j = \cup_{i \in I_j} \partial d_{ij}$, or (b) $i \in I_j$, in which case it is non-zero only on ∂d_{ji}. This means that the right-hand side of the above equation becomes

$$\left(\int_{\partial d_j} a\nabla \phi_j \cdot \boldsymbol{n}_j \right) u_j + \sum_{i \in I_j} \left(\int_{\partial d_j} a\nabla \phi_i \cdot \boldsymbol{n}_j \right) u_i = \sum_{i \in I_j} \int_{\partial d_{ji}} a(\nabla \phi_j \cdot \boldsymbol{n}_j u_j + \nabla \phi_i \cdot \boldsymbol{n}_j u_i).$$

Recalling the explicit formula for ϕ_{ji} and the fact that $\frac{\mathrm{d}}{\mathrm{d}s}$ is differentiation along the line from $\boldsymbol{x}_j$ to $\boldsymbol{x}_i$, we see that ϕ_j on ∂d_{ji} is ϕ_{ji} and so

$$a\nabla \phi_j \cdot \boldsymbol{n}_j = a\frac{\mathrm{d}\phi_{ji}}{\mathrm{d}s} = -\frac{\overline{a}_{ji}}{|\boldsymbol{x}_i - \boldsymbol{x}_j|} \qquad \text{on} \quad \partial d_{ji}.$$

Similarly ϕ_i on ∂d_{ji} is ϕ_{ij} and so

$$a\nabla \phi_i \cdot \boldsymbol{n}_j = a\frac{\mathrm{d}\phi_{ij}}{\mathrm{d}s} = -a\frac{\mathrm{d}\phi_{ji}}{\mathrm{d}s} = \frac{\overline{a}_{ji}}{|\boldsymbol{x}_i - \boldsymbol{x}_j|} \qquad \text{on} \quad \partial d_{ji}.$$

The above expression (81) for the contour integral then simplifies to

$$\int_{\partial d_j} a\nabla u_h \cdot \boldsymbol{n}_j = \sum_{i \in I_j} \overline{a}_{ji} \frac{(u_i - u_j)}{|\boldsymbol{x}_i - \boldsymbol{x}_j|} |\partial d_{ji}|,$$

where $|\partial d_{ji}|$ is the length of ∂d_{ji}. Using this expression in the equations gives at once the difference equations

$$\sum_{i \in I_j} \frac{\overline{a}_{ji} |\partial d_{ji}|}{|d_j|} \frac{(u_j - u_i)}{|\boldsymbol{x}_j - \boldsymbol{x}_i|} + b_j u_j = f_j, \qquad j = 1, \ldots, N, \tag{82}$$

which is a system of linear algebraic equations of the form (44). It is not hard to verify that the corresponding system matrix A is a positive definite M-matrix.

3.4. Practical considerations

The solutions of the continuous and discrete problems will coincide only in very special cases. In general, there will be an error which can be expressed in terms of the mesh spacing h as, for example, in (58). However the replacement of the original continuous problem by a finite-dimensional problem may also have undesired side effects. For example, in section 2.2 it has already been indicated that the system of discrete equations may not always have a solution. This is not acceptable and hence it is important to find conditions which guarantee at least the existence of a discrete solution. In this section this is done for the standard finite element methods for problem (45) described in section 3.1. Analogous results for the non-standard finite element methods of section 3.2 and other properties of the discrete solutions are also discussed.

The existence of a discrete solution is usually demonstrated by establishing the non-singularity of the system matrix A of the discrete problem. This, in turn, can be verified in a number of ways. For finite element methods, if A is symmetric, it is customary to verify its positive definiteness. This is equivalent to the condition

$$\boldsymbol{x}^T A \boldsymbol{x} \geqslant \alpha \boldsymbol{x}^T \boldsymbol{x} \qquad \text{for all} \quad \boldsymbol{x} \quad \text{and for some} \quad \alpha > 0$$

and it implies that all of the eigenvalues of A are positive.

Verifying the nonsingularity of A directly is usually difficult, and therefore it is often done in an indirect way. To this end, the concepts of an L-matrix and an M-matrix are introduced as follows.

Definition 1. *An $n \times n$ matrix A is an L-matrix if, for all $1 \leqslant i, j \leqslant n$,*

$$a_{ii} > 0 \qquad \text{and} \qquad a_{ij} \leqslant 0.$$

It is an M-matrix if, in addition, A^{-1} exists and all of its elements are non-negative, that is

$$A^{-1} \geqslant 0.$$

It can be shown that an irreducibly diagonally dominant L-matrix is an M-matrix (see [49, p 85]). In other words, if A is an irreducible L-matrix and if also, for all $i = 1, \ldots, n$,

$$a_{i,i} \geqslant -\sum_{j=1, j\neq i}^{n} a_{i,j},$$

with strict inequality for at least one value of i, then the matrix A^{-1} exists, contains only non-negative elements and therefore A is an M-matrix.

It follows that if the matrix A of the discrete system is an M-matrix then not only does the discrete system have a unique solution, but also that this discrete solution is non-negative whenever the term $\boldsymbol{b}$ on the right-hand side of the discrete system is non-negative. To see this, note that under the latter assumption $b_j \geqslant 0$ (see (44)) for all $1 \leqslant j \leqslant N-1$, and since all of the entries of A^{-1} are non-negative, we have

$$\boldsymbol{u} = A^{-1}\boldsymbol{b} \geqslant 0.$$

An analogous property can be demonstrated for the solution of the continuous problem (45) (see [50, p 112, theorem 8.1] or [51]) and the associated problem is said to satisfy a maximum principle. Thus, when the coefficient matrix is an M-matrix, the discrete problem inherits this property from the original problem. It is then said that the discrete problem satisfies a discrete maximum principle.

If the system matrix A is not an M-matrix, it may still be nonsingular, and the discrete solution may also be non-negative. However, in this case it is possible to construct a non-negative function $\boldsymbol{F}$ for which this does not hold. Thus, despite the fact that a discrete solution exists, it does not necessarily share the non-negativity of the continuous solution. If the solution to be found is a temperature or a concentration, for example, then this is clearly an undesirable situation. For this reason, it is wise to design discretization methods for which the system matrix is an M-matrix. This applies not only to problem (45), but also to other problems for which a maximum principle holds.

The discrete problem obtained in section 3.1 for the problem (45) is now investigated in more detail. The system of discretized equations has the compact form (44), with matrix elements given in (56). Due to the assumptions on the functions a and b, the diagonal elements $a_{j,j}$ can easily be shown to be positive. The off-diagonal elements $a_{j,j-1}$ are

$$a_{j,j-1} = \frac{1}{(x_j - x_{j-1})^2} \int_{x_{j-1}}^{x_j} (-a(x) + (x_j - x)(x - x_{j-1})b(x))\,\mathrm{d}x$$

and since $(x_j - x)(x - x_{j-1}) < 0$, they are negative. Hence, the L-character of the system matrix is established. Establishing the M-character is now straightforward. From (56) and the precise form of the basis functions ϕ_j it follows that

$$a_{j,j-1} + a_{j,j} + a_{j,j+1} = \int_{x_{j-1}}^{x_j} \{a(x)\phi_j'(x)(\phi_{j-1}'(x) + \phi_j'(x))$$

$$+b(x)\phi_j(x)(\phi_{j-1}(x)+\phi_j(x))\}\,\mathrm{d}x$$
$$+\int_{x_j}^{x_{j+1}}\{a(x)\phi_j'(x)(\phi_j'(x)+\phi_{j+1}'(x))+b(x)\phi_j(x)(\phi_j(x)+\phi_{j+1}(x))\}\,\mathrm{d}x$$
$$=\int_{x_{j-1}}^{x_{j+1}} b(x)\phi_j(x)\,\mathrm{d}x$$
$$\geqslant 0.$$

Strict inequality is obtained in the first and the last rows, since these contain only two non-zero entries. Hence, the coefficient matrix is irreducibly diagonally dominant, and it follows that $A^{-1}\geqslant 0$, as required.

As has been shown at the beginning of section 3.2, discrete solutions may not always be non-negative. For convection-diffusion problems, oscillatory discrete solutions may occur, depending on the size of the cell Reynolds number (see condition (64)). The same condition on the mesh spacing is found by examining the system matrix corresponding to (62), since

$$a_{j,j+1}=-\frac{\varepsilon}{h}+\frac{a}{2}$$

which is negative if and only if $\beta=\frac{ah}{2\varepsilon}<1$. Hence, the system matrix is an L-matrix if and only if (64) is satisfied. This is a severe restriction, which makes the standard finite element method quite unsuitable for singularly perturbed convection-diffusion problems.

A similar analysis can be given for the non-standard finite element methods discussed in section 3.2. For the upwind method it is easily verified that the condition (67) is satisfied if and only if the matrix is an L-matrix. Since the sum of the elements in each row is non-negative, the system matrix is also an M-matrix if (67) is satisfied.

For the one-dimensional method using inverse averages, the discrete equations are given in (71), and it is rather simple to demonstrate that the coefficient matrix is an M-matrix. Clearly, the coefficients of u_{j-1} and u_{j+1} are negative, whereas the coefficient of u_j is positive. Furthermore, the jth row sum is equal to b_j, $j=2,\ldots,N-1$, and it is positive for the first and last rows. Thus, the system matrix is an M-matrix and so the non-standard finite element method using inverse averages produces non-negative discrete solutions. It is not hard to verify that the same is true in two dimensions.

So far in this section only the non-negativity of discrete solutions has been discussed. Since the quantities to be approximated usually have a physical significance, this is an important issue. From a mathematical point of view, however, accuracy is of equal importance. Unfortunately, an acceptable discrete solution is not necessarily of sufficient accuracy. It is well known, for example, that discrete solutions obtained with upwind finite element methods are often too smooth. This is due to the presence of unphysical artificial diffusion mentioned in section 3.2. This has the effect of smearing out the boundary layer behaviour, thus causing errors in the discrete solution. In figure 9, the upwind solution corresponding to that in figure 6 is displayed, where it is seen that the boundary layers are not resolved well.

It is possible to remedy this deficiency of upwind finite element methods in various ways. The easiest approach is to modify the upwind method described in section 3.2. For the specific choice

$$\gamma=\coth\beta-\frac{1}{\beta},$$

the exponentially fitted scheme of Allen and Southwell [52] and Il'in [53] is recovered. This scheme has the property that the nodal values of the discrete solution coincide with the nodal values of the solution of the continuous problem:

$$u_i=u(x_i).$$

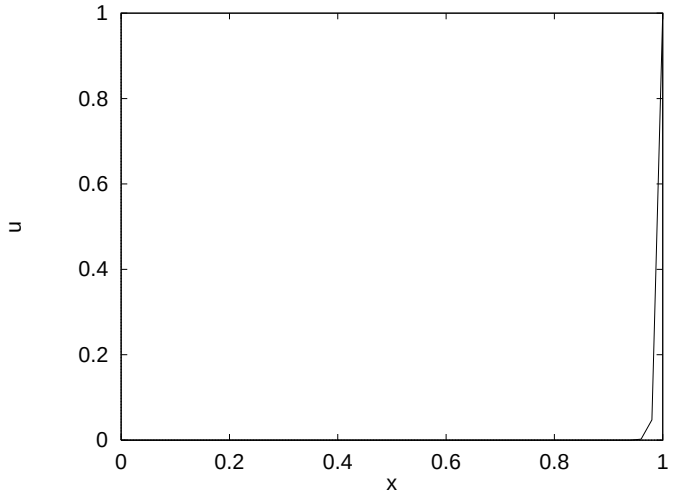

Figure 9. Numerical solution of (59) using an upwind finite element method.

This property is valid only when the coefficient a is constant, and when the right-hand side of (59) is constant. However, as has been demonstrated in [50] for comparable finite difference schemes, solutions of problems with non-constant coefficients also have the special property that the discrete solutions converge uniformly in ε to the solution of the continuous problem. This means that, in error estimates of the form (58), the error constant C is independent of the small parameter ε. A simple exercise shows that this is not the case for the standard finite element method, or for the upwind finite element method with a different choice of the parameter γ.

The finite element method using inverse averages described in the second part of section 3.2 also has a non-piecewise polynomial fitted character. This is due to the fact that the coefficients $\bar{a}_i$ are inverse averages of the continuous coefficient a. To see this for the problem (59), first rewrite this problem in the form (45). This can be done by multiplying the differential equation in (59) by the integrating factor $\exp(-cx/\varepsilon)$, and introducing the change of variables $v(x) = \exp(-cx/\varepsilon)u(x)$. Then the differential equation for v becomes

$$(av')' = 0,$$

with

$$a(x) = -\varepsilon \exp(cx/\varepsilon).$$

It turns out that the use of the inverse average of this function leads to a discrete scheme which coincides with the Allen–Southwell or Il'in scheme described above. For more details, and a derivation of this fact, the reader is referred to [8, 9].

Although exponentially fitted methods are optimal as far as accuracy is concerned, it is rather difficult to formulate these for two-dimensional problems. If rectangular meshes are used, tensor product schemes may be used, but these often suffer from a considerable amount of cross-wind diffusion. The latter can be minimized by aligning the mesh lines with the field lines and the equipotential lines. Furthermore, as has been demonstrated by Shishkin [2], for a two-dimensional singularly perturbed problem on a *uniform* rectangular mesh it is not possible to construct a finite difference method having an error constant C independent of the small parameter ε if parabolic boundary layers are present. As this is more often than not the case in realistic semiconductor models, the adequacy of exponentially fitted methods is severely

curtailed. In a recent monograph, Miller *et al* [2] have therefore adopted an entirely different approach. Instead of modifying the discrete operator, a careful selection of mesh points is undertaken. Details of this approach and a clear exposition of Shishkin's fundamental negative result can also be found in this monograph. In the following section we give a brief account of the method for a singular perturbation problem which can be regarded as a model problem for the Poisson equation in the semiconductor device model.

3.5. The Shishkin mesh technique

From the physics of semiconductor devices and the mathematical analysis of the device equations, we know that interior layers in a device are caused by the abrupt changes or discontinuities of its doping function on the right-hand side of the Poisson equation. As a simple model of this behaviour we now consider the following one-dimensional Poisson equation

$$\begin{cases} -\varepsilon u''(x) + u(x) = H(x-d), & x \in \Omega \\ u(0) = u_0, \qquad u(1) = u_1 \end{cases} \tag{83}$$

where $\varepsilon > 0$, u_0, u_1 and d are constants, $d \in (0, 1)$ and $H(x)$ denotes the Heaviside function defined as 0 when $x < 0$ and 1 when $x > 0$. The reduced problem corresponding to (83) is obtained by putting $\varepsilon = 0$ in the differential equation. The resulting problem has the discontinuous solution $u(x) = H(x - d)$, which does not in general satisfy the boundary conditions in (83). This shows that when $0 < \varepsilon \ll 1$, the solution of (83) is smooth, but it has boundary layers at $x = 0$ and 1 and an interior layer at $x = d$. Obviously, the derivative of the solution is large in these layers and almost constant elsewhere. So, if the widths of these layers are known, then we can design a mesh in such a way that the mesh nodes are condensed in these layers. In general, it is almost impossible to find the exact width of a layer, but we can estimate its order in terms of the singular perturbation parameter ε. The latter also involves intensive analysis even for a simple problem such as (83). Let us explain this briefly using our present example.

It can be shown that (cf [2, ch 6]), for $0 \leqslant k \leqslant 3$,

$$|u^{(k)}(x)| \leqslant \begin{cases} C(1 + \varepsilon^{-\frac{k}{2}} e_1(x)), & x \in (0, d) \\ C(1 + \varepsilon^{-\frac{k}{2}} e_2(x)), & x \in (d, 1) \end{cases}$$

where C is a constant, independent of ε, and

$$e_1(x) = \mathrm{e}^{-x\sqrt{1/\varepsilon}} + \mathrm{e}^{-(d-x)\sqrt{1/\varepsilon}}$$
$$e_2(x) = \mathrm{e}^{-(x-d)\sqrt{1/\varepsilon}} + \mathrm{e}^{-(1-x)\sqrt{1/\varepsilon}}.$$

From these bounds we see that when x lies in any of the three subdomains $x \leqslant \mathrm{O}(\sqrt{\varepsilon})$, $|x - d| \leqslant \mathrm{O}(\sqrt{\varepsilon})$ or $1 - x \leqslant \mathrm{O}(\sqrt{\varepsilon})$, the singular part of $u^{(k)}(x)$ behaves like $\varepsilon^{-k/2}$ for $k = 1, 2, 3$. Outside these subdomains, the layer functions e_1 and e_2 are negligible, because

$$\mathrm{e}^{-|x-c|\sqrt{1/\varepsilon}} \leqslant \varepsilon^k, \qquad \text{if} \quad |x - c| \geqslant K\sqrt{\varepsilon}\log\varepsilon$$

for any $c \in [0, 1]$ and any constant $K > 0$. This suggests that the widths of the boundary and the interior layers are $\mathrm{O}(\sqrt{\varepsilon})$. When designing a mesh for (83), it is natural to insist that some mesh points are located in these layer regions. This requirement suggests decomposing Ω into six subdomains Ω_i, where

$$\Omega_1 = [0, \sigma_1) \qquad \Omega_2 = [\sigma_1, d - \sigma_1) \qquad \Omega_3 = [d - \sigma_1, d)$$
$$\Omega_4 = [d, d + \sigma_2) \qquad \Omega_5 = [d + \sigma_2, 1 - \sigma_2) \qquad \Omega_6 = [1 - \sigma_2, 1].$$

To obtain a Shishkin mesh we use a uniform mesh in each of the six subdomains with $N/8$ nodes uniformly distributed in each of the four layer subregions Ω_i $(i = 1, 3, 4, 6)$, and $N/4$ nodes uniformly distributed in each of the subregions Ω_2 and Ω_5. The transition parameters σ_1 and σ_2 are chosen to be

$$\sigma_1 = \min\left\{\frac{d}{4}, 2\sqrt{\varepsilon}\log N\right\}, \qquad \sigma_2 = \min\left\{\frac{1-d}{4}, 2\sqrt{\varepsilon}\log N\right\},$$

which ensures that the mesh is appropriately fitted to the layers. This particular choice of the transition parameters is crucial in the numerical method. Note that the transition parameters depend on both the singular perturbation parameter ε and the number of mesh points N. With this distribution, the fitted mesh becomes a uniform mesh in the case that ε or N is sufficiently large that $\sigma_1 = \frac{d}{4}$ and $\sigma_2 = \frac{1-d}{4}$.

The piecewise-uniform fitted mesh defined above can be used essentially for any discretization scheme for solving (83). After a considerable amount of analysis, see [115], it can be shown that the error bound in the maximum norm for a simple classical scheme on this mesh satisfying a discrete maximum principle, such as the central difference scheme, is $CN^{-1}\ln N$ with C a positive constant independent of N and the singular perturbation parameter ε. Thus, this mesh technique guarantees that a discretization scheme will result in accurate results independent of ε.

Although finding the width of a layer involves intensive mathematical analysis, it is fortunate that many layers fall into one or other of two categories—regular layers and parabolic layers. When solving a singularly perturbed problem, we only need to identify the widths of the relevant layers. The locations of the layers are normally known *a priori*. With this information we can design a Shishkin mesh for the problem, as demonstrated above. For a complete discussion of this mesh technique for various layers, we again refer to [2].

4. Mixed finite element methods

In contrast to ordinary finite element methods, mixed finite element methods have first been suggested by mathematicians. These methods differ from ordinary finite element methods only in the sense that the problem to be discretized is split into a number of subproblems, each of which can be discretized using an ordinary finite element method. The latter can be different for the different subproblems, whence the terminology 'mixed'.

One of the first papers on mixed finite element methods, which still serves as a basic reference, is the paper by Brezzi [54]. Since then, the mixed finite element method has been applied to a multitude of problems, including the semiconductor device problem. It is attractive to use it for this problem, in view of the fact that one is often more interested in electric fields and current densities than in potentials and carrier concentrations. A feature of the mixed finite element method is that the fields can be represented as polynomials of a higher order than those representing the potentials. This feature cannot be replicated in ordinary finite element methods.

In this section we describe both standard mixed finite element methods and a number of recently developed non-standard methods. As was done above for ordinary finite element methods, extensions to higher dimensions are discussed, as well as practical issues.

4.1. A standard mixed finite element method

Mixed finite element methods are very similar to finite element methods, the main difference being that the differential equation to be discretized is split into several lower-order differential

equations so that a system of equations is obtained. Applying this to (45), for example, gives

$$a(x)u'(x) - \sigma(x) = 0, \qquad x \in \Omega \tag{84}$$

$$\sigma'(x) - b(x)u(x) = -f(x), \qquad x \in \Omega. \tag{85}$$

We see that the second-order differential equation for the single unknown function u is transformed into a system of first-order differential equations for the two unknown functions u and σ, where σ can be regarded as a flux. This system can be put into variational form using methods similar to those in section 3.1. If the intrinsic relationship between u and σ is ignored for the moment, appropriate function spaces U, V, S, T can be chosen for which the following weak problem is defined:

$$\begin{aligned} &\text{Find } (u, \sigma) \in U \times S \qquad \text{such that for all} \quad \tau \in T \\ &\int_0^1 u(x)(a(x)\tau(x))' + \int_0^1 \sigma(x)\tau(x)\,\mathrm{d}x = 0 \end{aligned} \tag{86}$$

$$\text{and for all} \quad v \in V$$

$$\int_0^1 \sigma'(x)v(x)\,\mathrm{d}x - \int_0^1 b(x)u(x)v(x)\,\mathrm{d}x = -\int_0^1 f(x)v(x)\,\mathrm{d}x. \tag{87}$$

Note that integration by parts and the assumption that $u(0) = u(1) = 0$ have been used to obtain the first of these equations, just as in the derivation of (47). Note also that the occurrence of derivative in (86) and (87) suggests that S and T should be subspaces of $H^1(\Omega)$. The spaces V and T are usually equal to U and S, respectively, but they may also be chosen differently (see section 4.2). In this section, it is assumed that $V = U$ and $T = S$.

Brezzi [54] and Babuška [55] analysed variational systems of the above form, and derived two conditions which guarantee the existence and uniqueness of solutions in the case where $b = 0$. To formulate these conditions, an additional space W, a subspace of T, is introduced, where

$$W = \left\{ \tau \in T : \int_0^1 \tau'(x)v(x)\,\mathrm{d}x = 0 \right\}.$$

The first condition, known as W-ellipticity, is that there exists a constant $\alpha > 0$ such that

$$\int_0^1 \tau^2(x)\,\mathrm{d}x \geqslant \alpha \|\tau\|_1^2 \qquad \text{for all} \quad \tau \in W \tag{88}$$

where the H^1-norm $\|\cdot\|_1$ is defined in section 3.1. The second condition requires the existence of a constant $\beta > 0$ such that

$$\sup_{\tau \in T} \frac{\int_0^1 \tau'(x)v(x)\,\mathrm{d}x}{\|\tau\|_1} \geqslant \beta \|v\|_V, \qquad \text{for all} \quad v \in V, \tag{89}$$

where $\|\cdot\|_V$ is the norm for the space V. The latter condition is known as the inf-sup condition or the Babuška–Brezzi condition. This condition is an extension of the V-ellipticity condition (49) to the mixed variational formulation. Analogous conditions to these for discrete problems are respectively the W_h-ellipticity condition

$$\int_0^1 \tau_h^2(x)\,\mathrm{d}x \geqslant \alpha \|\tau_h\|_1^2 \qquad \text{for all} \quad \tau_h \in W_h \tag{90}$$

where $\alpha > 0$ is a constant and $W_h = \{\tau_h \in T_h : \int_0^1 (x)v_h(x)\,\mathrm{d}x = 0\}$, and the discrete Babuška–Brezzi condition

$$\sup_{\tau_h \in T_h} \frac{\int_0^1 \tau_h'(x)v_h(x)\,\mathrm{d}x}{\|\tau_h\|_1} \geqslant \beta \|v_h\|_{V_h} \qquad \text{for all} \quad v_h \in V_h \tag{91}$$

where $\|\cdot\|_{V_h}$ is the norm for the space V_h.

If conditions (88), (89) are satisfied, the variational problem (86), (87) has a unique solution $(u, \sigma) \in U \times S = V \times T$. The two conditions ensure that the subspaces chosen are rich enough to represent a solution of the variational system. This is necessary, since there is an intrinsic relation between the functions u and σ.

The discretization procedure is analogous to that for the standard finite element method described in section 3.1. We choose finite-dimensional subspaces U_h and S_h and construct a basis for each of them. We then express the trial and test functions in terms of these basis functions and set up the resulting system of algebraic equations.

One of the simplest choices for U_h and S_h satisfying (90), (91) is the following: let U_h be the set of all piecewise constant functions which are constant on each Dirichlet subinterval d_i defined in section 3.2, and let S_h be the set of all continuous piecewise linear functions which are linear on each subinterval Ω_i defined in section 3.1. At first sight this may seem a strange choice, since this implies that the fluxes are approximated by piecewise polynomials of a higher degree than those used for the original function. However, this is one of the attractive features of the mixed finite element method, because it permits a more accurate approximation of the flux, which is often the quantity of most interest. It is left to the reader to verify that conditions (90), (91) are fulfilled.

In order to obtain a system of algebraic equations, a basis must now be chosen for each of the spaces U_h and S_h. Taking account of the boundary conditions $u(0) = u(1) = 0$, it is clear that the set $\{\psi_i\}_1^{N-1}$ of characteristic functions of the Dirichlet subintervals d_i corresponding to interior nodes forms a basis of the space U_h, and that the set of hat functions $\{\phi_i\}_0^N$ introduced in section 3.1 forms a basis of the space S_h. Thus, any $u_h \in U_h$ may be written in the form

$$u_h(x) = \sum_{i=1}^{N-1} u_i \psi_i(x),$$

where $u_i = u_h(x_i)$, and any $\sigma_h \in S_h$ may be written as

$$\sigma_h(x) = \sum_{i=0}^{N} \sigma_i \phi_i(x)$$

where $\sigma_i = \sigma_h(x_i)$. The discrete form of (86)-(87) is now

Find $u_1, \ldots, u_{N-1}$ and $\sigma_0, \ldots, \sigma_N$ such that for all j, $0 \leqslant j \leqslant N$,

$$\int_0^1 \left\{ \sum_{i=1}^{N-1} u_i \psi_i(x) \right\} (a(x)\phi_j(x))' \,\mathrm{d}x + \int_0^1 \left\{ \sum_{i=0}^{N} \sigma_i \phi_i(x) \right\} \phi_j(x) \,\mathrm{d}x = 0 \tag{92}$$

and for all j, $1 \leqslant j \leqslant N-1$,

$$\int_0^1 \left\{ \sum_{i=0}^{N} \sigma_i \phi_i'(x) \right\} \psi_j(x) \,\mathrm{d}x - \int_0^1 b(x) \left\{ \sum_{i=1}^{N-1} u_i \psi_i(x) \right\} \psi_j(x) \,\mathrm{d}x = -\int_0^1 f(x)\psi_j(x) \,\mathrm{d}x. \tag{93}$$

Since each basis function is non-zero on at most two sub-intervals, these integrals extend over only one or two subintervals. This means that the foregoing set of equations can be rewritten in the form

$$u_{j-1} \int_{x_{j-1}}^{x_{j-1/2}} (a(x)\phi_j(x))' \,\mathrm{d}x + u_j \int_{x_{j-1/2}}^{x_{j+1/2}} (a(x)\phi_j(x))' \,\mathrm{d}x + u_{j+1} \int_{x_{j+1/2}}^{x_{j+1}} (a(x)\phi_j(x))' \,\mathrm{d}x$$
$$+\sigma_{j-1} \int_{x_{j-1}}^{x_j} \phi_{j-1}(x)\phi_j(x) \,\mathrm{d}x + \sigma_j \int_{x_{j-1}}^{x_{j+1}} (\phi_j(x))^2 \,\mathrm{d}x$$

$$+\sigma_{j+1}\int_{x_j}^{x_{j+1}} \phi_{j+1}(x)\phi_j(x)\,\mathrm{d}x = 0, \qquad 0 \leqslant j \leqslant N,$$

$$\sigma_{j-1}\int_{x_{j-1/2}}^{x_j} \phi'_{j-1}(x)\,\mathrm{d}x + \sigma_j \int_{x_{j-1/2}}^{x_{j+1/2}} \phi'_j(x)\,\mathrm{d}x + \sigma_{j+1}\int_{x_j}^{x_{j+1/2}} \phi'_{j+1}(x)\,\mathrm{d}x - u_j \int_{x_{j-1/2}}^{x_{j+1/2}} b(x)\,\mathrm{d}x$$
$$= -\int_{x_{j-1/2}}^{x_{j+1/2}} f(x)\,\mathrm{d}x, \qquad 1 \leqslant j \leqslant N-1.$$

The integrals in the above can be simplified considerably by using explicit expressions for the basis functions. Doing this gives

$$\frac{1}{2}a(x_{j-1/2})u_{j-1} + \frac{1}{2}\left(a(x_{j+1/2}) - a(x_{j-1/2})\right)u_j - \frac{1}{2}a(x_{j+1/2})u_{j+1} + \frac{x_j - x_{j-1}}{6}\sigma_{j-1}$$
$$+\left(\frac{x_j - x_{j-1}}{3} + \frac{x_{j+1} - x_j}{3}\right)\sigma_j + \frac{x_{j+1} - x_j}{6}\sigma_{j+1} = 0, \qquad 0 \leqslant j \leqslant N, \tag{94}$$

$$\tfrac{1}{2}\sigma_{j+1} - \tfrac{1}{2}\sigma_{j-1} - \left(\int_{x_{j-1/2}}^{x_{j+1/2}} b(x)\,\mathrm{d}x\right)u_j = -\int_{x_{j-1/2}}^{x_{j+1/2}} f(x)\,\mathrm{d}x, \qquad 1 \leqslant j \leqslant N-1. \tag{95}$$

This system of equations can be written in the compact form

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}\begin{pmatrix} \boldsymbol{\sigma} \\ \boldsymbol{u} \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \boldsymbol{F} \end{pmatrix}, \tag{96}$$

where $\boldsymbol{\sigma} = (\sigma_0, \ldots, \sigma_N)^T$, $\boldsymbol{u} = (u_1, \ldots, u_{N-1})^T$, A is a tridiagonal $(N+1)\times(N+1)$ matrix, and D is a diagonal $(N-1)\times(N-1)$ matrix. B and C are not square matrices, but they are still sparse in the sense that that they contain few non-zero elements.

A slightly different system of equations is obtained if, in the discrete form obtained from (86), the terms containing the unknown σ_j are not cancelled. The corresponding discrete equations are then of the form

$$\tfrac{1}{2}\sigma_{j+1} + \tfrac{1}{2}\sigma_j - \tfrac{1}{2}\sigma_j - \tfrac{1}{2}\sigma_{j-1} - \left(\int_{x_{j-1/2}}^{x_{j+1/2}} b(x)\,\mathrm{d}x\right)u_j = -\int_{x_{j-1/2}}^{x_{j+1/2}} f(x)\,\mathrm{d}x,$$
$$1 \leqslant j \leqslant N-1,$$

which can be rewritten as

$$\sigma_{j+1/2} - \sigma_{j-1/2} = -\int_{x_{j-1/2}}^{x_{j+1/2}} (f(x) - b(x)u_j)\,\mathrm{d}x, \qquad 1 \leqslant j \leqslant N-1$$

where the notation $\sigma_{i-1/2} = \frac{1}{2}(\sigma_{i-1} + \sigma_i)$ has been introduced. This is a natural conservative discretization of (84), (85). If $b = 0$, then this equation shows that the difference between the discrete fluxes at the midpoints is equal to the integral of the source term.

4.2. Some non-standard mixed finite element methods in one dimension

In the previous section, an illustration was given of the standard mixed finite element method, applied to the problem (84), (85). Before discussing the mixed finite element method in more detail, a number of non-standard variations on the method presented in section 4.1 are now considered.

The first non-standard mixed finite element method is obtained by using an alternative weak formulation. To this end, divide equation (84) by $-a(x)$ to get

$$-u'(x) + a^{-1}(x)\sigma(x) = 0 \tag{97}$$

which is possible since $a(x)$ is bounded away from zero. Then the system of equations consisting of (97) and (85) is put into variational form in the same manner as before. The weak problem is then

$$\begin{aligned}&\text{Find } (u,\sigma)\in U\times S \qquad \text{such that, for all} \quad \tau\in T,\\ &\int_0^1 u(x)\tau'(x)+\int_0^1 a^{-1}(x)\sigma(x)\tau(x)\,\mathrm{d}x=0\end{aligned} \tag{98}$$

$$\begin{aligned}&\text{and, for all} \quad v\in V,\\ &\int_0^1 \sigma'(x)v(x)\,\mathrm{d}x-\int_0^1 b(x)u(x)v(x)\,\mathrm{d}x=-\int_0^1 f(x)v(x)\,\mathrm{d}x.\end{aligned} \tag{99}$$

Just as in section 4.1, integration by parts has been applied to obtain the first equation and it has been assumed that $u(0)=u(1)=0$. Choosing suitable finite-dimensional spaces U_h, S_h, T_h and V_h we define the associated discrete problem

$$\begin{aligned}&\text{Find } (u_h,\sigma_h)\in U_h\times S_h \qquad \text{such that, for all} \quad \tau_h\in T_h,\\ &\int_0^1 u_h(x)\tau_h'(x)+\int_0^1 a^{-1}(x)\sigma_h(x)\tau_h(x)\,\mathrm{d}x=0\end{aligned} \tag{100}$$

$$\begin{aligned}&\text{and, for all} \quad v_h\in V_h,\\ &\int_0^1 \sigma_h'(x)v_h(x)\,\mathrm{d}x-\int_0^1 b(x)u_h(x)v_h(x)\,\mathrm{d}x=-\int_0^1 f(x)v_h(x)\,\mathrm{d}x.\end{aligned} \tag{101}$$

The rest of the procedure is now exactly the same as in the previous section, with the same choices for the spaces $U_h=V_h$ and $S_h=T_h$. Since the discretized version of (99) is the same as that obtained in the previous section, namely (95), only the discretized form of (98) will be given here. This is

$$\begin{aligned}u_{j-1}\int_{x_{j-1}}^{x_{j-1/2}}\phi_j'(x)\,\mathrm{d}x+u_j\int_{x_{j-1/2}}^{x_{j+1/2}}\phi_j'(x)\,\mathrm{d}x+u_{j+1}\int_{x_{j+1/2}}^{x_{j+1}}\phi_j'(x)\,\mathrm{d}x\\ +\sigma_{j-1}\int_{x_{j-1}}^{x_j}a^{-1}(x)\phi_{j-1}(x)\phi_j(x)\,\mathrm{d}x+\sigma_j\int_{x_{j-1}}^{x_{j+1}}a^{-1}(x)(\phi_j(x))^2\,\mathrm{d}x\\ +\sigma_{j+1}\int_{x_j}^{x_{j+1}}a^{-1}(x)\phi_{j+1}(x)\phi_j(x)\,\mathrm{d}x=0, \qquad 0\leqslant j\leqslant N.\end{aligned}$$

The coefficients of u_{j-1}, u_j, and u_{j+1} can now be calculated exactly, and the foregoing system of equations becomes

$$\begin{aligned}\tfrac{1}{2}u_{j-1}-\tfrac{1}{2}u_{j+1}+\sigma_{j-1}\int_{x_{j-1}}^{x_j}a^{-1}(x)\phi_{j-1}(x)\phi_j(x)\,\mathrm{d}x+\sigma_j\int_{x_{j-1}}^{x_{j+1}}a^{-1}(x)(\phi_j(x))^2\,\mathrm{d}x\\ +\sigma_{j+1}\int_{x_j}^{x_{j+1}}a^{-1}(x)\phi_{j+1}(x)\phi_j(x)\,\mathrm{d}x=0, \qquad 0\leqslant j\leqslant N.\end{aligned} \tag{102}$$

The discrete system of equations (102), (95) can then be written in a compact form similar to (96), the difference being that the matrices B and C are now related: $C=B^T$. The compact form is

$$\begin{pmatrix}A & B\\ B^T & D\end{pmatrix}\begin{pmatrix}\boldsymbol{\sigma}\\ \boldsymbol{u}\end{pmatrix}=\begin{pmatrix}\mathbf{0}\\ \boldsymbol{F}\end{pmatrix}. \tag{103}$$

Thus, the use of the variational formulation based on (97) rather than on (84) leads to a symmetric system matrix. This is certainly a desirable property, and it shows that this alternative formulation may be useful.

Just as for ordinary finite elements, alternative mixed finite element methods can be constructed. As an example of this the mixed finite element method involving inverse averages [56] is now described.

Dividing the first equation by the non-zero coefficient $-a$, and rearranging, the first-order equations (84), (85) become

$$-u'(x) + a^{-1}(x)\sigma(x) = 0,$$
$$-\sigma'(x) + b(x)u(x) = f(x).$$

Now multiply the first and second equations by test functions τ and v, respectively, and integrate to get

$$\int_0^1 (-u'\tau + a^{-1}\sigma\tau) = 0, \tag{104}$$
$$\int_0^1 (-\sigma' v + buv) = \int_0^1 fv. \tag{105}$$

Unlike what was done in the previous section, integration by parts is now applied to the second equation instead of the first. Furthermore, it is applied on each Dirichlet subinterval d_i separately, rather than on the whole interval $\overline{\Omega}$. Thus, the first term of the second equation can be written as

$$\int_0^1 \sigma' v = \sum_{i=0}^{N} \int_{d_i} \sigma' v = \sum_{i=0}^{N} \left([\sigma v]_{x_{i-1/2}}^{x_{i+1/2}} - \int_{x_{i-1/2}}^{x_{i+1/2}} \sigma v' \right).$$

If the test functions v are now chosen so that they are equal to a constant v_i on each d_i, this simplifies to

$$\int_0^1 \sigma' v = \sum_{i=0}^{N} (\sigma(x_{i+1/2}) - \sigma(x_{i-1/2}))v_i,$$

and the equations (104), (105) become

$$\int_0^1 (-u' + a^{-1}\sigma)\tau = 0,$$
$$-\sum_{i=0}^{N} (\sigma(x_{i+1/2}) - \sigma(x_{i-1/2}))v_i + \int_0^1 buv = \int_0^1 fv.$$

This suggests the following discrete variational problem

$$\begin{cases} \text{Find } u_h \in U_h, \quad \sigma_h \in S_h \qquad \text{such that for all} \quad v_h \in V_h, \\ \displaystyle\int_0^1 (-u_h' + a^{-1}\sigma_h)\tau_h = 0 \\ \text{and for all} \quad \tau_h \in T_h, \\ \displaystyle -\sum_{i=0}^{N} (\sigma_h(x_{i+1/2}) - \sigma_h(x_{i-1/2}))v_{h,i} + \int_0^1 bu_h v_h = \int_0^1 f v_h \end{cases}$$

where $v_{h,i}$ denotes the constant value of v_h on the Dirichlet subinterval d_i. It is not hard to verify that the discrete conditions (90), (91) are fulfilled by the discrete spaces $U_h = \text{span}\{\phi_i\}_1^{N-1}$, $V_h = \text{span}\{\psi_i\}_1^{N-1}$, and $S_h = T_h = \text{span}\{\xi_i\}_1^{N}$, where ϕ_i is the hat function for the node x_i, ψ_i is the characteristic function of the Dirichlet subinterval d_i and ξ_i is the characteristic function of the subinterval $\overline{\Omega}_i$. Note that the requirement that v_h is a constant on each d_i is also satisfied. Moreover, this is a Petrov–Galerkin problem for the discrete displacement u_h and a Bubnov–Galerkin problem for the discrete flux σ_h.

To obtain the corresponding linear system take $\tau_h = \xi_j$, for each $j = 1, \ldots, N$, and $v_h = \psi_j$ for each $j = 1, \ldots, N-1$. The equations become

$$\int_{x_{j-1}}^{x_j} (-u_h' + a^{-1}\sigma_h) = 0, \qquad j = 1, \ldots, N$$

$$-(\sigma_h(x_{i+1/2}) - \sigma_h(x_{i-1/2})) + \int_{x_{j-1/2}}^{x_{j+1/2}} bu_h = \int_{x_{j-1/2}}^{x_{j+1/2}} f, \qquad j = 1, \ldots, N-1.$$

Noting now that

$$u_h = \sum_{k=1}^{N-1} u_k \phi_k, \qquad \sigma_h = \sum_{k=1}^{N} \sigma_{k-1/2} \xi_k,$$

where $u_k = u_h(x_k)$ and $\sigma_{k-1/2} = \sigma_h(x_{k-1/2})$, the equations reduce to

$$-(u_j - u_{j-1}) + \left(\int_{x_{j-1}}^{x_j} a^{-1} \right) \sigma_{j-1/2} = 0, \qquad j = 1, \ldots, N$$

$$-(\sigma_{j+1/2} - \sigma_{j-1/2}) + \int_{x_{j-1/2}}^{x_{j+1/2}} bu_h = \int_{x_{j-1/2}}^{x_{j+1/2}} f, \qquad j = 1, \ldots, N-1.$$

Approximating the integrals in the second set of equations by the midpoint rule and recalling the relation

$$\overline{a}_{j-1/2} = \left(\frac{1}{x_j - x_{j-1}} \int_{x_{j-1}}^{x_j} a^{-1} \right)^{-1}$$

the equations become

$$-\frac{u_j - u_{j-1}}{x_j - x_{j-1}} + \overline{a}_{j-1/2}^{-1} \sigma_{j-1/2} = 0, \qquad j = 1, \ldots, N$$

$$-2\frac{\sigma_{j+1/2} - \sigma_{j-1/2}}{x_{j+1} - x_{j-1}} + b_j u_j = f_j, \qquad j = 1, \ldots, N-1.$$

The first set of equations can now be solved explicitly for the discrete fluxes $\sigma_{j-1/2}$ giving

$$\sigma_{j-1/2} = \overline{a}_{j-1/2} \frac{u_j - u_{j-1}}{x_j - x_{j-1}}, \qquad 1 \leqslant j \leqslant N. \tag{106}$$

Substituting this solution into the second set of equations gives

$$-\frac{2}{x_{j+1} - x_{j-1}} \left(\overline{a}_{j+1/2} \frac{u_{j+1} - u_j}{x_{j+1} - x_j} - \overline{a}_{j-1/2} \frac{u_j - u_{j-1}}{x_j - x_{j-1}} \right) + b_j u_j = f_j,$$
$$j = 1, \ldots, N-1.$$

This is a linear system of the form (44), where $\boldsymbol{u} = (u_1, \ldots, u_{N-1})^T$ and A is a symmetric, positive definite, M-matrix.

Note that this linear system of algebraic equations is identical to the system (71) obtained in section 3.2 by the non-standard ordinary finite element method using non-piecewise polynomial spaces. However, the present mixed finite element method has the advantage that it uses spaces involving only piecewise polynomial functions, and furthermore that it gives automatically the formula (106) above for the values of the discrete flux.

4.3. Extension to higher dimensions

As is the case for standard ordinary finite element methods, the theory for standard mixed methods is well developed. Important contributions have been made by Arnold and Brezzi [57], Brezzi [54] and Thomas [58]. For details of recent results we refer the reader to [59]. For practical applications, it is important to develop classes of spaces V_h and W_h which satisfy W_h-ellipticity (90) and the discrete Babuška–Brezzi condition (91). This has been done in [60] and [61], leading in the latter case to the frequently used Raviart–Thomas elements. Applications to semiconductor device problems started about 1987 in, for example, [15, 62, 63], with important and necessary improvements developed in [64–68]. Independently, a different type of mixed finite element method using inverse averages was developed in [56, 69]. Unfortunately, to the best of our knowledge, there are no easy-to-read introductions to the subject. This is mainly due to the complication introduced by the Babuška–Brezzi condition, and the frequent use of Sobolev spaces in the proofs of the error estimates. It is worth mentioning that [64] provides a rigorous introduction to the subject, with many one-dimensional examples. This is probably a good starting point for readers interested in mixed methods. In this section, only a brief summary of some of these methods is given and for simplicity the problems are restricted to two dimensions.

We consider here mixed finite element methods for the two-dimensional convection-diffusion problem (72). We introduce a new variable $\boldsymbol{\sigma} = a\nabla u$, which enables us to write the second-order equation in (72) as a system of first-order equations

$$-\nabla u + a^{-1}\boldsymbol{\sigma} = \mathbf{0},$$
$$-\nabla \cdot \boldsymbol{\sigma} + bu = f$$

with the boundary conditions

$$u = 0 \quad \text{on} \quad \Gamma_1 \qquad \text{and} \qquad \boldsymbol{\sigma} \cdot \boldsymbol{n} = 0 \quad \text{on} \quad \Gamma_2.$$

Taking the scalar product of the first with a vector-valued test function $\boldsymbol{\tau}$, integrating over Ω, and using a scalar test function v in the usual way for the second equation, gives

$$\int_\Omega (-\nabla u + a^{-1}\boldsymbol{\sigma}) \cdot \boldsymbol{\tau} = 0, \tag{107}$$

$$\int_\Omega (-\nabla \cdot \boldsymbol{\sigma} + bu)v = \int_\Omega fv. \tag{108}$$

Integrating the first term in (107) by parts and assuming that $\boldsymbol{\tau} \cdot \boldsymbol{n} = 0$ on Γ_2 we define the following variational problem

$$\begin{cases} \text{Find } u \in U,\ \boldsymbol{\sigma} \in S \qquad \text{such that for all} \quad \boldsymbol{\tau} \in T \\ \displaystyle\int_\Omega (u\nabla \cdot \boldsymbol{\tau} + a^{-1}\boldsymbol{\sigma} \cdot \boldsymbol{\tau}) = 0, \\ \text{and for all} \quad v \in V \\ \displaystyle\int_\Omega (-\nabla \cdot \boldsymbol{\sigma} + bu)v = \int_\Omega fv. \end{cases}$$

From this formulation we see that, unlike the variational formulation (74), the flux function $\boldsymbol{\sigma}$ needs to be smoother than the potential function u. This variational problem is called the dual mixed problem corresponding to (72). A discrete problem associated with this can then be defined by simply replacing the space S and V in the above by proper finite element subspaces S_h and V_h. One of the simplest choices for the spaces U_h and S_h are the Raviart–Thomas elements, the lowest order of which are defined as follows. Let E_h be a decomposition of $\overline{\Omega}$ into elements $\{E_i\}_1^N$, where each E_i is either a triangle or a rectangle. Then U_h is defined to

be the space of piecewise polynomials of degree zero on E_h, i.e. piecewise constant functions, a basis for which is $\{\psi_i\}_1^N$, where ψ_i is the characteristic function of E_i. To construct S_h we introduce the Raviart–Thomas functions on the standard unit triangle or standard unit square E defined by

$$RT(E) = \{\boldsymbol{v} = (v_1, v_2) : v_1 = a + bx_1, v_2 = c + dx_2, a, b, c, d \in \mathbb{R}\}$$

where it is assumed that $b = d$ if E is the unit triangle. The corresponding functions on any triangle or rectangle are obtained by an affine or bilinear mapping of the functions in $RT(E)$ to the new triangle or rectangle. Then the space S_h is defined as

$$S_h = \{\boldsymbol{\sigma}_h \in H_{\text{div}}(\Omega) : \boldsymbol{\sigma}_h|E_i \in RT(E_i), \text{ for all } E_i \in E_h \text{ and } \boldsymbol{\sigma}_h \cdot n = 0 \text{ on } \Gamma_2\}$$

where the notation $\boldsymbol{\sigma}_h|E_i$ means $\boldsymbol{\sigma}_h$ considered as a function on only the domain E_i and $H_{\text{div}}(\Omega) = \{v \in L^2(\Omega) : \text{div}\, v \in L^2(\Omega)\}$. It follows from this definition that the functions $\boldsymbol{\sigma}_h \in S_h$ must be continuous across edges common to two elements.

There are many possible choices for a basis of S_h. A convenient one is the following, because its basis functions are non-zero on at most two elements. Let e denote a common edge of any two adjacent elements. Then a basis function is defined for each such edge. The basis function $\boldsymbol{\tau}_e$ for the edge e is taken to be that function in S_h which has a normal component 1 on the edge e and zero normal component on all other edges. Then, any functions $u_h \in U_h$ and $\boldsymbol{\sigma}_h \in S_h$ can be written in the form

$$u_h(\boldsymbol{x}) = \sum_{i=1}^{N} u_i \psi_i(\boldsymbol{x}), \qquad \boldsymbol{\sigma}_h(\boldsymbol{x}) = \sum_e \sigma_e \boldsymbol{\tau}_e(\boldsymbol{x})$$

where u_i is the value of u_h on E_j and σ_e is the value of the normal components of $\boldsymbol{\sigma}_h$ on the edge e.

Just as in the one-dimensional case this choice of the basis functions leads to a system of equations with a system matrix of the form

$$\begin{pmatrix} A & B \\ B^T & 0 \end{pmatrix}, \tag{109}$$

where the matrix entries in A are

$$a_{ij} = \iint_\Omega a^{-1} \boldsymbol{\tau}_i \cdot \boldsymbol{\tau}_j.$$

The matrix B is rectangular with at most two non-zero entries per row. Its entries are

$$b_{ij} = \iint R_j \nabla \cdot \boldsymbol{\tau}_i.$$

With an appropriate scaling, it can be ensured that the non-zero elements of B are always ± 1. This is convenient when solving the linear system. Note that the foregoing construction of basis functions works both in the rectangular and triangular cases. The linear systems arising above are further analysed in section 4.4. It is not hard to verify that the lowest-order Raviart–Thomas finite element spaces defined in the above satisfy the discrete W_h-ellipticity and Babuška–Brezzi conditions (90) and (91). Hence, the spaces S_h and U_h can be used to approximate the solution of the problem (45). Note that the approximate flux $\boldsymbol{\sigma}_h$ is piecewise linear, while the function u_h is piecewise constant. Although this may seem contradictory, it is sometimes very convenient in practice since an accurate representation of the fluxes, rather than an accurate representation of the function u, is often desirable.

Higher-order Raviart–Thomas elements can also be defined, but this will not be done here, since the corresponding spaces are not used for semiconductor device modelling. The reason

for this will be explained in section 5. Readers interested in these higher-order finite element spaces are referred to the original paper by Raviart and Thomas [61].

The generalization of the second method discussed in section 4.2, namely the mixed finite element method involving inverse averages, is now described for the two-dimensional case. Its extension to three dimensions is completely analogous, see [69] for details.

Writing the first term in (108) as a sum of integrals over the Dirichlet tiles, applying Green's theorem to each integral and using the boundary conditions $v = 0$ on Γ_1 and $\boldsymbol{\sigma} \cdot \mathrm{n} = 0$ on Γ_2 gives

$$\int_\Omega (\nabla \cdot \boldsymbol{\sigma}) v = \sum_i \int_{d_i} (\nabla \cdot \boldsymbol{\sigma}) v = \sum_i \left(\int_{\partial d_i} \boldsymbol{\sigma} \cdot \boldsymbol{n} v - \int_{d_i} \boldsymbol{\sigma} \cdot \nabla v \right).$$

Assuming henceforth that the test functions v are chosen to be constant on each Dirichlet tile, we have $\nabla v = 0$ on each tile, so that the right-hand side reduces to a sum of boundary integrals. Thus the variational formulation of (107) and the above is

$$\begin{cases} \text{Find } u \in U,\ \boldsymbol{\sigma} \in S \qquad \text{such that for all} \quad \boldsymbol{\tau} \in T \\ \displaystyle\int_\Omega (-\nabla u + a^{-1}\boldsymbol{\sigma}) \cdot \boldsymbol{\tau} = 0, \\ \text{and for all} \quad v \in V \\ \displaystyle -\sum_i \int_{\partial d_i} \boldsymbol{\sigma} \cdot \boldsymbol{n} v + \int_\Omega b u v = \int_\Omega f v. \end{cases}$$

This is sometimes called the primal mixed formulation corresponding to (72), because the smoothness requirements for u and $\boldsymbol{\sigma}$ coincide with those of (74). It is not hard to verify that W-ellipticity (88) and the Babuška–Brezzi condition (89) are fulfilled for this problem. The corresponding discrete problem is

$$\begin{cases} \text{Find } u_h \in U_h,\ \boldsymbol{\sigma}_h \in S_h \qquad \text{such that for all} \quad \boldsymbol{\tau}_h \in T_h \\ \displaystyle\int_\Omega (-\nabla u_h + a^{-1}\boldsymbol{\sigma}_h) \cdot \boldsymbol{\tau}_h = 0, \\ \text{and for all} \quad v_h \in V_h \\ \displaystyle -\sum_i \int_{\partial d_i} \boldsymbol{\sigma}_h \cdot \boldsymbol{n} v_h + \int_\Omega b u_h v_h = \int_\Omega f v_h, \end{cases}$$

where it is assumed that the discrete test functions v_h are piecewise constant functions, constant on each d_i. From the above definitions it follows that each test function v_h is constant on all Dirichlet tiles d_i, and that the discrete W_h-ellipticity and Babuška–Brezzi conditions (90), (91) are satisfied, as required. The spaces U_h and V_h are defined by $U_h = \mathrm{span}\{\phi_i\}$, $V_h = \mathrm{span}\{\psi_i\}$, where ϕ_i is the standard piecewise linear or bilinear basis function for the node $\boldsymbol{x}_i$ of the triangulation (which may consist of triangles, rectangles or a mixture of both) and ψ_i is the characteristic function of the Dirichlet tile for the same node. The spaces of vector-valued functions are taken to be the same, i.e. $S_h = T_h$, and are defined as follows. Let $\boldsymbol{x}_i, \boldsymbol{x}_j$ be any two neighbouring nodes. Then their corresponding Dirichlet tiles d_i, d_j have a common line segment γ_{ij} on their boundaries. As in section 3, let b_{ij} denote the box formed by the union of the two triangles with the common base γ_{ij} and the vertices $\boldsymbol{x}_i, \boldsymbol{x}_j$ respectively. For each b_{ij} define a piecewise constant vector-valued function $\boldsymbol{\xi}_{ij}$, which is non-zero only on b_{ij} and at each point of b_{ij} is equal to the unit vector taken in one of the two possible senses parallel to the undirected line passing through $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. Note that $\boldsymbol{\xi}_{ij}$ is associated with the box b_{ij} and therefore $\boldsymbol{\xi}_{ji} = \boldsymbol{\xi}_{ij}$, because obviously $b_{ji} = b_{ij}$. Furthermore $\boldsymbol{\xi}_{ij} \cdot \boldsymbol{\xi}_{kl} = \delta_{ik}\delta_{jl}\delta_{il}\delta_{jk}$. Also the closure of the union of these boxes covers the whole of $\overline{\Omega}$. Now define $S_h = T_h = \mathrm{span}\{\boldsymbol{\xi}_{ij}\}$. Then, for any $\boldsymbol{\sigma}_h \in S_h$, $\boldsymbol{\sigma}_h(\boldsymbol{x}) = \sum_{b_{ij}} \sigma_{ij} \boldsymbol{\xi}_{ij}(\boldsymbol{x})$ where the summation is over all of the boxes

b_{ij}, and for any $u_h \in U_h$, $u_h(\boldsymbol{x}) = \sum_i u_i \phi_i(\boldsymbol{x})$. It is easy to check that $\boldsymbol{\sigma}_h(\boldsymbol{x}) \cdot n = 0$ on Γ_2 because when restricted to Γ_2, $\boldsymbol{\xi}_{ij}(\boldsymbol{x})$ is parallel to the tangent direction of Γ_2.

Note that the discrete problem is Petrov–Galerkin for u_h and Bubnov–Galerkin for $\boldsymbol{\sigma}_h$. Now take $v_h = \psi_j$ and $\boldsymbol{\tau}_h = \boldsymbol{\xi}_{ij}$ in the equations and get

$$\int_{b_{ij}} (-\nabla u_h + a^{-1}\boldsymbol{\sigma}_h) \cdot \boldsymbol{\xi}_{ij} = 0,$$
$$-\sum_i \int_{\partial d_i} \boldsymbol{\sigma}_h \cdot \boldsymbol{n}_i \psi_j + \int_{d_j} b u_h = \int_{d_j} f$$

where ∂d_i denotes the boundary of the Dirichlet tile d_i and $\boldsymbol{n}_i$ is the unit outward normal to ∂d_i. On d_j it is clear that

$$\nabla u_h \cdot \boldsymbol{\xi}_{ij} = \frac{u_j - u_i}{|\boldsymbol{x}_i - \boldsymbol{x}_j|}$$

and that

$$\boldsymbol{\sigma}_h = \sigma_{ij} \boldsymbol{\xi}_{ij}.$$

The first equation can then be written as

$$-\frac{u_j - u_i}{|\boldsymbol{x}_i - \boldsymbol{x}_j|} |b_{ij}| + \sigma_{ij} \int_{b_{ij}} a^{-1}(x)\, \mathrm{d}x = 0,$$

where $|b_{ij}|$ denotes the area of the box b_{ij}. Introducing the notation

$$\overline{a}_{ij} = \left(\frac{1}{|b_{ij}|} \int_{b_{ij}} a^{-1}(\boldsymbol{x})\, \mathrm{d}x \right)^{-1} \tag{110}$$

this becomes

$$\sigma_{ij} = \overline{a}_{ij} \frac{u_j - u_i}{|\boldsymbol{x}_i - \boldsymbol{x}_j|}. \tag{111}$$

Thus the first set of equations has been solved explicitly for $\boldsymbol{\sigma}_h$ in terms of u_h.

Now consider the second equation and note that

$$\int_{\partial d_i} \boldsymbol{\sigma}_h \cdot \boldsymbol{n}_i \psi_j = \delta_{ij} \int_{\partial d_j} \boldsymbol{\sigma}_h \cdot \boldsymbol{n}_j$$

because the value of ψ_j under the boundary integral sign is its limiting value on the boundary ∂d_i coming from the interior of the Dirichlet tile d_i, which is zero unless $i = j$. The equation can therefore be written in the form

$$-\int_{\partial d_j} \boldsymbol{\sigma}_h \cdot \boldsymbol{n} + \int_{d_j} bu = \int_{d_j} f.$$

Noting now that in the first term $\partial d_j = \cup_{i \in I_j} \partial d_{ji}$, where ∂d_{ji} denotes the line segment γ_{ji} directed so as to be part of the boundary ∂d_j, and using one-point quadrature rules for the last two terms, gives

$$-\sum_{i \in I_j} \int_{\partial d_{ji}} \boldsymbol{\sigma}_h \cdot \boldsymbol{n} + b_j u_j |d_j| = f_j |d_j|,$$

where $|d_j|$ denotes the area of d_j. But on d_{ji}, $\boldsymbol{\sigma}_h = \sigma_{ji} \boldsymbol{\xi}_{ji}$ and so

$$\boldsymbol{\sigma}_h \cdot \boldsymbol{n} = \sigma_{ji} \boldsymbol{\xi}_{ji} \cdot \boldsymbol{n} = \sigma_{ji}.$$

The equation therefore becomes

$$-\sum_{i\in I_j}\sigma_{ji}\frac{|\partial d_{ji}|}{|d_j|}+b_ju_j=f_j,$$

where $|\partial d_{ji}|$ denotes the length of ∂d_{ji}. Substituting (111) into this gives

$$\sum_{i\in I_j}\bar{a}_{ji}\frac{|\partial d_{ji}|}{|d_j|}\frac{u_j-u_i}{|\boldsymbol{x}_j-\boldsymbol{x}_i|}+b_ju_j=f_j \tag{112}$$

which is a system of linear algebraic equations of the form (44) identical to that obtained in section 3.3.

4.4. Practical considerations

In practical problems, it often happens that there are constraints on the discrete spaces that must be taken into account. Thus, not all combinations S_h and V_h developed in the theory can be used for practical problems such as the semiconductor device problem. In this section the one-dimensional case is used for illustrative purposes. Next, general constraints are formulated. As in section 3.4, the concept of an M-matrix plays an important role. In fact, it is often the first priority to make sure that discrete solutions satisfy a discrete maximum principle. Within the class of methods having this property, a specific method is then chosen which yields a reasonable degree of accuracy.

Consider the two-point boundary value problem

$$\begin{cases}(a(x)u'(x))'=f(x,u(x)), & x\in\Omega\\ u(0)=u_l, \qquad u(1)=u_r.\end{cases} \tag{113}$$

The corresponding linear system in (96) must now be solved. Unfortunately, the system matrix is neither an M-matrix nor positive definite. To demonstrate this, consider the simpler case in which $C=B^T$ and $D\equiv 0$. Then it is easy to verify that

$$\begin{pmatrix}A & B\\ B^T & 0\end{pmatrix}=\begin{pmatrix}A & 0\\ B^T & I\end{pmatrix}\begin{pmatrix}A^{-1} & 0\\ 0 & -B^TA^{-1}B\end{pmatrix}\begin{pmatrix}A & B\\ 0 & I\end{pmatrix},$$

which shows that the system matrix has positive as well as negative eigenvalues.

One possible remedy might be to eliminate the unknown flux values from the linear system, and then to solve the resulting system for the unknown values of u_h. Unfortunately, this does not work as can be seen from the following simple example in one space dimension. Taking the functions $a\equiv 1$, $f\equiv 0$, and using four mesh points on the interval $[0,1]$, the matrices A and B are found to be:

$$A=\frac{1}{24}\begin{pmatrix}2&1&0&0&0\\1&4&1&0&0\\0&1&4&1&0\\0&0&1&4&1\\0&0&0&1&2\end{pmatrix},$$

and

$$B=\begin{pmatrix}-1&0&0&0\\1&-1&0&0\\0&1&-1&0\\0&0&1&-1\\0&0&0&1\end{pmatrix}.$$

If σ is eliminated from the system, the system matrix for the vector of unknown values of u_h is $B^T A^{-1} B$. In this example

$$B^T A^{-1} B = \frac{1}{7} \begin{pmatrix} 201 & -99 & 27 & -9 \\ -99 & 129 & -81 & 27 \\ 27 & -81 & 129 & -99 \\ -9 & 27 & -99 & 201 \end{pmatrix},$$

which is positive definite, but is not an M-matrix as required.

The foregoing example shows that, even for a very simple one-dimensional problem, the mixed finite element method can lead to a system of equations which has undesirable properties, and so modifications are in order. In section 5 a remedy is discussed which has been developed within the framework of semiconductor device simulation, but which can also be used for problems in other areas. The approach consists of using quadrature rules to improve the properties of the system matrices. A more classical remedy for the aforementioned difficulties is that suggested by Fraeijs de Veubeke [70]. This approach consists of allowing σ_h to lie in a larger space than the space S_h. Fraeijs de Veubeke suggested taking σ_h in the space S_h^*, which consists of piecewise linear functions not necessarily continuous at the points of intersection of the subintervals. In this case, for each $v \in S_h$, v' is not square-integrable on $(0, 1)$ because of the discontinuity at the nodes. In fact v' behaves like a sum of δ-functions in $(0, 1)$, and so it is only in the dual space $H^{-1}(\Omega)$ of the space $H^1(\Omega)$. Thus, S_h^* is not a subspace of $H_{\text{div}}(\Omega)$. The same idea applies to the more general case. To this end, let S_h^* be a finite-dimensional space consisting of functions which are smooth in each element segment, but may not be continuous at the nodes. For these functions, the discrete mixed variational formulation (100) and (101) is not valid because S_h is not a subset of $H_{\text{div}}(\Omega)$. However, when v does not have continuous normal components but is still in $H_{\text{div}}(E)$ for every element in the decomposition E_h of Ω, integration by parts can be performed on each element separately. Hence

$$\int_\Omega v \cdot \nabla u = \sum_{E \in E_h} \int_E v \cdot \nabla u = - \sum_{E \in E_h} \int_E u \nabla \cdot v + \sum_{\partial E \in \text{Ed}_h} \int_{\partial E} u v \cdot n,$$

where Ed_h is the set of edges in the decomposition E_h of Ω. The function u, restricted to an edge ∂E, is called the Lagrange multiplier on ∂E. This Lagrange multiplier, denoted by λ, is a function defined on Ed_h. For simplicity we describe the method first in the one-dimensional case; later a two-dimensional case is considered.

In the one-dimensional case, the multiplier λ is defined on the internal nodes x_i of the decomposition. The space Λ_h of all nodal functions is an $(n-1)$-dimensional space. The function σ_h is chosen in the space of piecewise linear functions on the intervals, which is a $2n$-dimensional space. Let $\sigma_{i,r}$ be the right limit value of σ_h at a node x_i and $\sigma_{i,l}$ be the left limit value of σ_h at the same node. The continuity of σ_h at a point of intersection of two adjacent subintervals can be regained by adding a third equation

$$\sum_{i=1}^{n-1} \mu(\sigma_{i,r} - \sigma_{i,l}) = 0, \qquad \text{for all} \quad \mu \in \Lambda_h. \tag{114}$$

The mixed finite element method using Lagrange multipliers can then be summarized as follows. First define the spaces

$$S_h^* = \{\tau_h \in L^2(\Omega) | \tau_h|_I \in RT(I) \text{ for all } I \in I_h\},$$
$$\Lambda_h = \{\mu : \{x_1, \ldots, x_{n-1}\} \to R\}.$$

Then the Dirichlet boundary conditions are imposed by defining $\lambda_h(x_0) = u_1$ and $\lambda_h(x_n) = u_r$. The discrete problem is defined as follows

$$\begin{cases} \text{Find } (\sigma_h^*, u_h^*, \lambda_h) \in S_h^* \times V_h \times \Lambda_h \qquad \text{such that for all} \quad \tau_h \in S_h^* \\ \displaystyle\int_{x_0}^{x_n} a^{-1}\sigma_h^*\tau_h + \sum_{i=0}^{n-1}\int_{x_i}^{x_{i+1}} u_h^*\tau_h' = \sum_{i=0}^{n-1}(\lambda_h\tau_h|_{x_{i+1}^-} - \lambda_h\tau_h|_{x_i^+}), \\ \displaystyle\sum_{i=0}^{n-1}\int_{x_i}^{x_{i+1}} \phi_h(\sigma_h^*)' = \int_{x_0}^{x_n}\phi_h f, \qquad \text{for all} \quad \phi_h \in V_h, \\ \displaystyle\sum_{i=1}^{n-1}\mu(\sigma_{i,r}^* - \sigma_{i,l}^*) = 0 \qquad \text{for all} \quad \mu \in \Lambda_h. \end{cases} \tag{115}$$

Problem (115) has a unique solution $(\sigma_h^*, u_h^*, \lambda_h)$. Furthermore, if (σ_h, u_h) is the solution of (100) and (101), then

$$\sigma_h^* = \sigma_h, \qquad u_h^* = u_h.$$

Moreover,

$$\lambda_h|_{x_i} \sim u|_{x_i}, \qquad \text{for all} \quad 0 \leqslant i \leqslant n.$$

A proof of this can be found in [57, p 12]. The final statement above can be made more precise, but here the discussion is limited to indicating that the multipliers are approximations to the solution.

The system in (115) can be rewritten just as was done for the method without Lagrange multipliers. To this end, introduce the basis functions τ_{i1} and τ_{i2}, which are zero except on the interval T_i, where

$$\tau_{i1}(x) = \frac{x - x_{i+1}}{x_{i+1} - x_i}, \qquad \tau_{i2}(x) = \frac{x - x_i}{x_{i+1} - x_i}.$$

The $n-1$ basis functions μ_i in Λ_h are defined to be 1 at the interior node x_i and 0 at all other nodes $1 \leqslant i \leqslant n-1$. The basis functions of V_h are the same as before. Now write the unknown functions in terms of these basis functions:

$$\sigma_h = \sum_{i=0}^{n-1}(\sigma_{i1}\tau_{i1} + \sigma_{i2}\tau_{i2}), \qquad u_h = \sum_{i=0}^{n-1} u_i\phi_i, \qquad \lambda_h = \sum_{i=1}^{n-1}\lambda_i\mu_i.$$

Let the vectors σ, $\boldsymbol{u}$ and λ denote the column vectors of unknown coefficients. Then the system in (115) can be written in the compact form

$$\begin{bmatrix} A & B & C \\ B^T & 0 & 0 \\ C^T & 0 & 0 \end{bmatrix}\begin{bmatrix} \sigma \\ \boldsymbol{u} \\ \lambda \end{bmatrix} = \begin{bmatrix} \boldsymbol{b} \\ \boldsymbol{F} \\ \boldsymbol{0} \end{bmatrix}. \tag{116}$$

Here, A is a $2n \times 2n$ block diagonal matrix consisting of 2×2 matrices A_i, with

$$(A_i)_{jk} = \int_{x_i}^{x_{i+1}} a^{-1}\tau_{ij}\tau_{ik}.$$

B is a $2n \times n$ matrix which, on an element level, contains the 2×1 matrix $B_i = (1, 1)^T$, and C is a $2n \times (n-1)$ matrix containing element identity matrices, assuming that the basis functions have unit normal components. The right-hand side vector is

$$F_i = \int_{I_i} f,$$

and

$$\boldsymbol{b} = (u_l, 0, \ldots, 0, u_r)^T.$$

Since A is easy to invert, it is reasonable to write

$$\sigma = -A^{-1}(B\boldsymbol{u} + C\lambda) + A^{-1}\boldsymbol{b}.$$

Substituting this into the second set of equations in (116) gives

$$\boldsymbol{u} = -(B^T A^{-1} B)^{-1} B^T A^{-1}(C\lambda - \boldsymbol{b}) - (B^T A^{-1} B)^{-1}\boldsymbol{F}.$$

Finally, the results of these elimination steps can be substituted into the third set of equations in (116), which gives the following system for the Lagrange multipliers

$$C^T(-A^{-1} + A^{-1}B(B^T A^{-1} B)^{-1} B^T A^{-1})(C\lambda - \boldsymbol{b}) = -C^T A^{-1} B(B^T A^{-1} B)^{-1}\boldsymbol{F}. \tag{117}$$

To study the properties of this system, introduce the matrix

$$Q = -A^{-1} + A^{-1}B(B^T A^{-1} B)^{-1} B^T A^{-1}.$$

Then the system matrix of the system for λ in (117) is $C^T QC$. It is shown in [54, p 4] that this matrix is an M-matrix, and therefore a discrete maximum principle holds.

Unfortunately, this result holds only if the function f is independent of the unknown solution u. If f does depend on u, then the M-property may be destroyed. For example, if

$$f(u(x), x) = \nu u(x), \tag{118}$$

then the system in (116) is replaced by

$$\begin{bmatrix} A & B & C \\ B^T & -D & 0 \\ C^T & 0 & 0 \end{bmatrix} \begin{bmatrix} \sigma \\ \boldsymbol{u} \\ \lambda \end{bmatrix} = \begin{bmatrix} \boldsymbol{b} \\ \boldsymbol{F} \\ \boldsymbol{0} \end{bmatrix}, \tag{119}$$

where the diagonal matrix D contains the elements $d_{ii} = \nu(x_{i+1} - x_i)$, $0 \leqslant i \leqslant n-1$. The matrix Q is

$$Q = -A^{-1} + A^{-1}B(B^T A^{-1} B + D)^{-1} B^T A^{-1}.$$

When D is large, it is clear that $-C^T A^{-1} C$ itself must be an M-matrix. It is left to the reader to verify that this is not the case for the example given at the beginning of this section (see also [64]).

The difficulties sketched in the foregoing can be avoided by making changes to the mixed method. This has been described in detail in [64, pp 57–70]. Here we consider the simple case of problem (113) with $a(x) = \exp(-\psi(x))$ for some function ψ, f as in (118), and $u_l = u_r = 1$. Taking the same basis functions as in [62], the discretized system is of the form

$$\begin{pmatrix} A & B \\ B^T & -\nu D \end{pmatrix} \begin{pmatrix} \sigma \\ \boldsymbol{u} \end{pmatrix} = \begin{pmatrix} \boldsymbol{r}_1 \\ \boldsymbol{r}_2 \end{pmatrix},$$

where $\boldsymbol{u}$ and σ are the vectors of unknown values at the mesh points (note that we did not introduce Lagrange multipliers). D is a positive diagonal matrix, since u is approximated by piecewise constants. Eliminating σ we get

$$(-B^T A^{-1} B - \nu D)\boldsymbol{u} = \boldsymbol{r}_2 - B^T A^{-1}\boldsymbol{r}_1$$

or

$$(I + \nu(B^T A^{-1} B)^{-1} D)\boldsymbol{u} = (B^T A^{-1} B)^{-1}(\boldsymbol{r}_2 - B^T A^{-1}\boldsymbol{r}_1).$$

For $\nu = 0$, the solution is well behaved, even though the matrix $B^T A^{-1} B$ is not an M-matrix. But for large ν this is not the case. A remedy in that case is the lumping of the mass matrix A. Lumping basically means choosing a Lobatto quadrature rule to evaluate the entries of A, so that quadrature abcissae coincide with the nodes used in the finite element method. In this case we use

$$\int_{x_i}^{x_{i+1}} \exp(-\psi) f(x) \sim f(x_i) \int_{x_i}^{x_{i+1/2}} \exp(-\psi) + f(x_{i+1}) \int_{x_{i+1/2}}^{x_{i+1}} \exp(-\psi).$$

With this quadrature rule, only the diagonal elements of A are non-zero and they are given by

$$a_{ii} = \int_{x_{i-1/2}}^{x_{i+1/2}} \exp(-\psi).$$

The result is Il'in's method or, in the case of semiconductor device simulation, the Scharfetter–Gummel method provided that the integrals of $\exp(-\psi)$ are evaluated exactly, assuming that ψ is piecewise linear.

The remedy explained in the foregoing can also be applied to two-dimensional problems. For rectangular elements, this is straightforward (see [64]). For triangles, the situation is more complicated, and the proof that the resulting coefficient matrix has the M-property is rather involved. The reader is referred to [65, 67] for the details of the quadrature rules used in that case, and for proofs of the discrete maximum principle.

5. Application of finite element methods to semiconductor device models

5.1. History

Many papers have been published in which the finite element method is used as a means of obtaining discrete solutions of the full system of semiconductor device equations. The largest concentration of papers on this subject is found in the past ten years. This is not surprising, since most of the earlier efforts in the area of semiconductor device modelling were aimed at creating methods to simulate a particular device which could usually be modelled by a simple set of equations. In view of the difficulty of the problem, this was not a trivial exercise. Gradually, more general discretization methods and solution techniques matured, and now there are a number of accepted and/or preferred methods. Having established such methods, it is natural to explore others which may be beneficial in one respect or another. This is the main reason for the increased interest in the use of finite element methods. Another reason is the increased interest of mathematicians in the semiconductor device problem. In the early years of device modelling, only a few mathematicians expressed an interest in this industrial problem. Gradually it was recognized as a challenging problem, with many interesting mathematical features.

The first mention of the use of finite element methods in the context of semiconductor device modelling is probably in the papers of Buturla and Cottrell, in the 1970s [71, 72], although there are also other early contributors [73–76]. Most of these publications are concerned with the simulation of individual devices. Computer programs using the finite element method for a large class of semiconductor problems started to become available in the early 1980s [77–79]. In [79] the application is restricted to the off-state semiconductor problem, whereas in [77, 78] the full drift-diffusion model is used. On the other hand, it appears that the finite element method used in these papers is precisely the finite volume method (for a description, see [8]). Hence, strictly speaking, the method employed is not a finite element method, since a variational formulation is not used.

Although the finite element method is successful in many other disciplines, and particularly in the simulation of mechanical problems, its application to the simulation of semiconductor devices is not without problems. The main difficulty is due to the occurrence of boundary and interior layers in the devices which, as we have demonstrated in previous sections, usually necessitates the use of some form of upwinding or exponential fitting. Since this turns out to be hard to achieve for two- and three-dimensional problems, the finite volume method gradually became the preferred technique for discretizing the drift-diffusion equations. Although the latter is still the case, finite element methods are beginning to be used more often. The use

of adaptive meshing techniques makes more demands on the discretization which are more easily met by finite element methods. In this respect, the finite volume method is rather restricted, because there is not much theory and it is not clear how to obtain higher orders of approximation.

In the following sections a number of ordinary and mixed finite element methods will be presented that have been used for semiconductor device simulation. The overview is by no means complete, but an attempt has been made to extract the major developments. In addition to the references mentioned below, there is a multitude of papers in which the finite element method is used only as a tool, without any important modification of the method itself. These papers are not listed here, since this would make the bibliography unwieldy.

5.2. Ordinary finite element methods

From the preceding sections it is clear that the use of standard ordinary finite element methods for the discretization of the drift-diffusion equations is doomed to fail, unless measures are taken to capture the boundary and interior layers. One exception to this is the off-state problem. The use of finite element methods for this problem is considered first.

The off-state problem is described by a single partial differential equation, the nonlinear Poisson equation (41), supplemented by appropriate boundary conditions. It is assumed that the user knows the values of the Slotboom variables Φ_p and Φ_n or, in case the carrier concentrations are negligible, can specify appropriate approximations. In [80], it is shown how suitable piecewise constant values can be chosen depending on the boundary conditions for these variables. Since Poisson's equation is in self-adjoint form, there is no need for upwinding or exponential fitting (for an explanation of this, see [8, 50]). As a consequence, the use of standard finite element methods is feasible. Another attractive feature of the off-state problem is that the possible occurrence of oscillatory discrete solutions is not necessarily a problem. In fact, the carrier concentrations are calculated as a post-processing exercise, making use of the expressions

$$p = \exp\left(-\frac{q\psi}{kT}\right)\Phi_p,$$
$$n = \exp\left(\frac{q\psi}{kT}\right)\Phi_n.$$

Clearly, this method never leads to negative carrier concentrations. Hence, even though the discretized system of equations may produce a non-physical oscillatory approximation to the electric potential, the solution is physically acceptable. Needless to say, the quality of the solution improves if these non-physical oscillations are not present.

In [79], a standard finite element method is employed for the simulation of a large class of off-state semiconductor problems. The mesh used for the simulations is topologically equivalent to a rectangular grid, and consists of arbitrary quadrilateral elements. The spaces of trial and test functions are chosen to be piecewise bilinear functions. By using degenerate quadrilateral elements, the method also allows the use of triangular elements. In this case, the trial and test functions are linear on such elements.

Although the method has proved to be very useful for practical problems, a disadvantage is that the discrete electrostatic potential may exhibit non-physical oscillations. These oscillations may occur when the shape of the quadrilateral elements is distorted too far from a square. In other words, whenever the angles are much larger than $\frac{\pi}{2}$, or when the aspect ratio of the lengths of the edges is large. In [8] it is shown that, if rectangular elements are used, non-oscillatory

solutions can be guaranteed only when

$$\frac{1}{\sqrt{2}} < \frac{h_x}{h_y} < \sqrt{2}.$$

Here, h_x and h_y are the lengths of the sides of the rectangle in the x and y directions, respectively. If this condition is not fulfilled, the coefficient matrix may violate the conditions necessary for it to be an M-matrix. Hence, oscillatory solutions may appear, depending on the boundary conditions and on the location of the rectangles violating the above condition. This may happen also for triangular meshes, when obtuse angles occur [8].

For the on-state problem the full set of drift-diffusion equations must be considered. As has been demonstrated in [80], the use of standard finite element methods with piecewise linear basis and test functions leads to oscillatory solutions and negative carrier concentrations. Hence, for a finite element method to yield a physically acceptable solution, it is necessary to change the discrete function spaces. In several papers this problem is circumvented by using Scharfetter–Gummel expressions for the current densities, effectively reducing the finite element method to a finite volume method, see for example [77, 78]. Alternatively, however, the Scharfetter–Gummel expressions can be used to obtain a suitable bilinear form, which can be used in the variational formulation. One of the first papers describing this technique is [81], and a brief description of this method follows. The starting point is the set of time-dependent drift-diffusion equations, with constant electron and hole mobilities and a simple Shockley–Read–Hall expression for the recombination/generation term. To simplify the notation and the analysis, both the dependent and independent variables are scaled, and the resulting system of equations is

$$-\Delta\psi = \alpha(p - n + D(x)), \tag{120}$$

$$\frac{\partial n}{\partial t} = \gamma_n \nabla \cdot \boldsymbol{J}_n - R(n, p), \tag{121}$$

$$\frac{\partial p}{\partial t} = \gamma_p \nabla \cdot \boldsymbol{J}_p - R(p, n). \tag{122}$$

The scaled current densities are

$$\boldsymbol{J}_n = \mu_n(\nabla n - n\nabla\psi), \tag{123}$$

$$\boldsymbol{J}_p = -\mu_p(\nabla p + p\nabla\psi). \tag{124}$$

The above system of equations have to be solved on a bounded domain $\Omega \times (0, T)$. Without loss of generality, we assume that the system satisfies the following initial and homogeneous Dirichlet and Neumann boundary conditions

$$\psi(x, 0) = \psi_0(x), \qquad n(x, 0) = n_0(x), \qquad p(x, 0) = p_0(x), \qquad x \in \Omega,$$
$$\psi(x, t) = n(x, t) = p(x, t) = 0 \qquad x \in \Gamma_1, \quad t \in (0, T),$$
$$\nabla\psi(0, t) \cdot \boldsymbol{n} = \boldsymbol{J}_n(0, t) \cdot \boldsymbol{n} = \boldsymbol{J}_p(0, t) \cdot \boldsymbol{n} = 0 \qquad x \in \Gamma_2, \quad t \in (0, T),$$

where Γ_1 and Γ_2 denote respectively the parts of the boundary on which Dirichlet and Neumann conditions are specified. A problem with non-homogeneous boundary conditions can be transformed into this form by subtracting an appropriate known function in each of the equations (120)–(122), as demonstrated in section 3.2 for the one-dimensional problem. It is assumed here that Ω is a two-dimensional domain, unless otherwise stated.

Before the above system is transformed to variational form, the following scaled Slotboom variables are introduced

$$w = \exp(-\psi)n, \qquad z = \exp(\psi)p, \tag{125}$$

so that in these variables

$$\boldsymbol{J}_n = \mu_n \exp(\psi)\nabla w, \tag{126}$$

$$\boldsymbol{J}_p = -\mu_p \exp(-\psi)\nabla z. \tag{127}$$

Multiplying each of the equations in (120)–(122) by a function v from the space

$$V = \{v \in H^1(\Omega) | v_{|\Gamma_1} = 0\},$$

and using Green's theorem repeatedly, the following variational form is obtained

$$\int_\Omega \nabla\psi \cdot \nabla v - \alpha \int_\Omega (p - n + D)v = 0, \tag{128}$$

$$\int_\Omega \frac{\partial n}{\partial t} v + \gamma_n \int_\Omega \mu_n \exp(\psi)\nabla w \cdot \nabla v + \int_\Omega R(n, p)v = 0, \tag{129}$$

$$\int_\Omega \frac{\partial p}{\partial t} v + \gamma_p \int_\Omega \mu_p \exp(-\psi)\nabla z \cdot \nabla v + \int_\Omega R(n, p)v = 0, \tag{130}$$

which must hold for all $v \in V$.

To obtain a discrete variational formulation, $\overline{\Omega}$ is covered by triangular elements K (in [81], the case of quadrilaterals is also discussed), the union of which is denoted by $\overline{\Omega_h}$. If $\hat{K}$ denotes the unit reference element, the element K can be described geometrically by

$$\boldsymbol{x}(\xi_1, \xi_2) = \sum_{j=1}^{3} \boldsymbol{x}^j M_j(\xi_1, \xi_2),$$

where the $\boldsymbol{x}^j$ are the vertices of the triangular element K, and the M_j are the shape functions

$$M_1(\xi_1, \xi_2) = 1 - \xi_1 - \xi_2, \qquad M_2(\xi_1, \xi_2) = \xi_1, \qquad M_3(\xi_1, \xi_2) = \xi_2.$$

We now introduce the space of piecewise linear elements

$$W_h = \left\{ v \in C^0(\overline{\Omega_h}) \mid v|_K = \sum_{j=1}^{3} v_j M_j \right\},$$

and the associated space

$$V_h = \{v \in W_h | v = 0 \text{ on } \Gamma_1\}.$$

The unknown functions ψ, n, p, w, and z are approximated by functions ψ_h, n_h, p_h, w_h, and z_h, respectively, all of which are assumed to be in W_h. It is assumed that the discrete approximations satisfy the relationships

$$w_h(\boldsymbol{x}^j) = \exp(-\psi_h(\boldsymbol{x}^j))n_h(\boldsymbol{x}^j),$$
$$z_h(\boldsymbol{x}^j) = \exp(\psi_h(\boldsymbol{x}^j))p_h(\boldsymbol{x}^j).$$

The discrete variational equations can now be formulated. The difficulty lies in the approximation of the second integrals in equations (129) and (130), since it is known that a straightforward treatment using standard quadrature formulae leads to unsatisfactory results. Ideally, in deriving such approximations, the techniques used to obtain the Scharfetter–Gummel expressions should be employed. This can be done as follows. Assume that the electron current density $\boldsymbol{J}_n$ is approximated by a vector function $\boldsymbol{\sigma}_n$, which is constant on each element. Then

$$\boldsymbol{\sigma}_n \approx \mu_n \exp(\psi)\nabla w,$$

and so

$$\mu_n \nabla \hat{w} \approx \exp(-\hat{\psi}) J^T \boldsymbol{\sigma}_n. \tag{131}$$

Here, $\hat{w}$ and $\hat{\psi}$ are the functions corresponding to w and ψ on the reference element $\hat{K}$, and J is the Jacobian matrix of the mapping from this element to the element K.

Integrating the first component of (131) along the ξ_1-axis, and the second component along the ξ_2-axis, gives

$$\mu_n \begin{pmatrix} w(\boldsymbol{x}^2) - w(\boldsymbol{x}^1) \\ w(\boldsymbol{x}^3) - w(\boldsymbol{x}^1) \end{pmatrix} \approx D_n J^T \boldsymbol{\sigma}_n.$$

Here, D_n is a 2×2 diagonal matrix with diagonal entries

$$(D_n)_{1,1} = \int_0^1 \exp(-\hat{\psi}(\xi_1, 0))\, \mathrm{d}\xi_1$$

$$(D_n)_{2,2} = \int_0^1 \exp(-\hat{\psi}(0, \xi_2))\, \mathrm{d}\xi_2.$$

Replacing ψ and w by their discrete analogues, the latter two integrals can be evaluated exactly, which gives

$$D_n = \mathrm{diag}(B(\psi_h(\boldsymbol{x}^1) - \psi_h(\boldsymbol{x}^2)), B(\psi_h(\boldsymbol{x}^1) - \psi_h(\boldsymbol{x}^3)))$$

where B is the familiar Bernoulli function

$$B(t) = \begin{cases} \dfrac{t}{\exp(t) - 1} & t \neq 0 \\ 1 & t = 0. \end{cases} \tag{132}$$

Rearranging, this leads to

$$\boldsymbol{\sigma}_n = \mu_n \exp(\psi_h(\boldsymbol{x}^1))(J^T)^{-1} D_n J^T \nabla w_h \tag{133}$$

which is an expression for the discrete analogue of the electron current density.

Finally, to arrive at the discrete formulation, (133) is used to define an approximation of the second integral in (129). In the discrete problem, the integral

$$\int_\Omega \mu_n \exp(\psi) \nabla w \cdot \nabla v$$

is replaced by

$$\sum_K \mu_n \exp(\psi(\boldsymbol{x}^1)) \int_K (J^T)^{-1} D_n J^T \nabla w_h \cdot \nabla v_h.$$

In a similar fashion, a discrete analogue of (130) can be derived.

The discrete variational formulation thus obtained has several attractive features. First of all, it can be shown [81] that the formulation is symmetric with respect to w_h and v_h (or z_h and v_h in the case of holes). It can also be shown that a discrete maximum principle holds if the triangulation is of acute type, i.e. if all triangles in the triangulation are non-obtuse. As a result of the latter property, it is proved in [81] that a discrete solution exists, and that the discrete carrier concentrations are non-negative.

In [81], Zlámal gives a detailed analysis of the above method. In order to simplify the computations, he introduces a slightly different method which he refers to as a partly linear scheme. For the latter method, uniqueness of the discrete solution can also be demonstrated. The paper contains descriptions of the method with quadrilateral elements in the two-dimensional case, and with simplicial and tetrahedral elements in the three-dimensional case.

Although Zlámal's finite element method has a number of attractive features, some comment is required. First of all, the method is different from conventional finite element methods in the sense that one-dimensional arguments are used to obtain suitable

approximations to the integrals or, in mathematical terms, the bilinear forms. Secondly, the carrier concentrations are approximated by piecewise linear functions. In view of their inherent exponential character, this may not be ideal. Thirdly, and probably most importantly, the success of the method depends heavily on the fact that the current densities are assumed to be piecewise constant.

It is not clear how to extend Zlámal's approach to yield higher-order finite element methods with similar properties. This is unfortunate, because a suitable family of finite element methods, if available, would have nice properties such as the ability to handle complex geometries and a wide choice of convergence rates. In [82], a class of inverse-average finite element methods is introduced. This work is extended in [83] to the three-dimensional case. Unfortunately, these two papers contain only straightforward generalizations of the method described in [81] above, and no proposals for higher-order methods. Finally, it should be noted that the inverse-average methods fit into the framework of the generalized finite element methods developed in [84]. A closer look at the examples in the latter paper reveals that again no suggestions for higher-order methods are made.

5.3. The ordinary finite element method with inverse averages

We now consider the application of the method with inverse averages described in sections 3.2 and 3.3 to the semiconductor device equations (120)–(122) in the scaled Slotboom variables (125) with the current densities defined in (126) and (127). We take the two-dimensional case for simplicity and follow the notation in section 3.3.

Let E_h be a decomposition of the solution domain Ω into either rectangles or Delaunay triangles or of a combination of both. For each vertex x_i of E_h, the associated Dirichlet tile is denoted by d_i.

Comparing (120)–(122) in the Slotboom variables with (72) we see that each of these has the same form as (72) plus a time derivative. The coefficient functions corresponding to the three equations are respectively 1, $\gamma_n\mu_n$ and $\gamma_p\mu_p$. Thus, we can simply apply the result (82) to (120)–(122) with the corresponding coefficient functions.

The Poisson equation is of the form (72) with $a(\boldsymbol{x}) = 1$. Thus, applying (72) to this equation we get

$$\sum_{i\in I_j} \frac{|\partial d_{ji}|}{|d_j|}\frac{(\psi_j - \psi_i)}{|\boldsymbol{x}_j - \boldsymbol{x}_i|} - \alpha(p_j - n_j) = \alpha D_j, \qquad j = 1, \ldots, N, \tag{134}$$

where ψ_j, n_j and p_i are nodal approximations to ψ, n and p respectively and $D_i = D(\boldsymbol{x}_j)$.

Each of the two continuity equations (121) and (122) differs from (72) by a time-derivative term. Thus, the variational formulation (77) in section 3.3 for the stationary problem needs some slight modifications. Equivalently, we may multiply (121) and (122) by the basis function ψ_j defined in (78), integrate the resulting equations over the tile d_i and apply Green's formula to the second-order terms. This yields

$$\int_{d_j} \frac{\partial(\exp(\psi)w)}{\partial t} - \gamma_n \int_{\partial d_j} \mu_n \exp(\psi)\nabla w \cdot \boldsymbol{n}_j + \int_{d_i} R(\exp(\psi)w, \exp(-\psi)z) = 0,$$

$$\int_{d_j} \frac{\partial(\exp(-\psi)z)}{\partial t} - \gamma_p \int_{\partial d_j} \mu_p \exp(-\psi)\nabla z \cdot \boldsymbol{n}_j + \int_{d_i} R(\exp(\psi)w, \exp(-\psi)z) = 0,$$

for $j = 1, \ldots, N$. Replacing ψ, w and z respectively by ψ_h, w_h and z_h and using the one-point quadrature rule we obtain

$$\frac{\partial(\exp(\psi_j)w_j)}{\partial t}|d_j| - \gamma_n \int_{\partial d_j} \mu_n \exp(\psi_h)\nabla w_h \cdot \boldsymbol{n}_j + R_j(\exp(\psi_j)w_j, \exp(-\psi_j)z_j)|d_j| = 0,$$

$$\frac{\partial(\exp(-\psi_j)z_j)}{\partial t}|d_j| - \gamma_p \int_{\partial d_j} \mu_p \exp(-\psi)\nabla z_h \cdot \boldsymbol{n}_j + R_j(\exp(\psi_j)w_j, \exp(-\psi_j)z_j)|d_j| = 0,$$

for $j = 1, \ldots, N$. Each of these is of the form (79) with an extra time derivative. Thus, using the same basis function ϕ_i defined in section 3.3 and following the same argument, we have the following linear system corresponding to (82)

$$\frac{\partial(\exp(\psi_j)w_j)}{\partial t} + \gamma_n \sum_{i\in I_j} \frac{\mu_n \overline{\exp(\psi_h)}_{ji}|\partial d_{ji}|}{|d_j|}\frac{(w_j - w_i)}{|\boldsymbol{x}_j - \boldsymbol{x}_i|} + R_j(\exp(\psi_j)w_j, \exp(-\psi_j)z_j) = 0,$$

$$\frac{\partial(\exp(-\psi_j)z_j)}{\partial t} + \gamma_n \sum_{i\in I_j} \frac{\mu_p \overline{\exp(-\psi_h)}_{ji}|\partial d_{ji}|}{|d_j|}\frac{(z_j - z_i)}{|\boldsymbol{x}_j - \boldsymbol{x}_i|} + R_j(\exp(\psi_j)w_j, \exp(-\psi_j)z_j) = 0,$$

for $j = 1, \ldots, N$, where $\overline{\exp(\psi_h)}_{ji}$ and $\overline{\exp(-\psi_h)}_{ji}$ respectively are the inverse average approximations of $\exp(\psi_h)$ and $\exp(-\psi_h)$ on the segment connecting $\boldsymbol{x}_j$ and $\boldsymbol{x}_i$ defined in (80). We now discuss the evaluation of these approximations. From the discretized Poisson equation (134), or the construction of the basis function ϕ_j in section 3.3, we see that ψ_h is linear along the edge from $\boldsymbol{x}_j$ to $\boldsymbol{x}_i$. Thus, if we parametrize ψ_h along the edge so that

$$\psi_h(s) = \psi_j + \frac{\psi_i - \psi_j}{|\boldsymbol{x}_j - \boldsymbol{x}_i|}s, \qquad 0 \leqslant s \leqslant |\boldsymbol{x}_j - \boldsymbol{x}_i|$$

then, from this and (80) we have

$$\begin{aligned}\overline{\exp(\psi_h)}_{ji} &= \left(\frac{1}{|\boldsymbol{x}_j - \boldsymbol{x}_i|}\int_0^{|\boldsymbol{x}_j - \boldsymbol{x}_i|} \exp\left(-\psi_j + \frac{\psi_i - \psi_j}{|\boldsymbol{x}_j - \boldsymbol{x}_i|}s\right) \mathrm{d}s\right)^{-1} \\ &= \exp(\psi_j)\cdot\frac{\psi_j - \psi_i}{\exp(\psi_j - \psi_i) - 1} \\ &= \exp(\psi_j)B(\psi_j - \psi_i),\end{aligned}$$

where $B(\cdot)$ is the Bernoulli function defined in (132). Similarly we have

$$\overline{\exp(-\psi_h)}_{ji} = \exp(-\psi_j)B(\psi_i - \psi_j).$$

Substituting these into the above discretized equations for w and z we obtain

$$\frac{\partial(\exp(\psi_j)w_j)}{\partial t} + \gamma_n \sum_{i\in I_j} \frac{\mu_n \exp(\psi_j)B(\psi_j - \psi_i)|\partial d_{ji}|}{|d_j|}\frac{(w_j - w_i)}{|\boldsymbol{x}_j - \boldsymbol{x}_i|} + R_j(\exp(\psi_j)w_j, \exp(-\psi_j)z_j) = 0, \tag{135}$$

$$\frac{\partial(\exp(-\psi_j)z_j)}{\partial t} + \gamma_n \sum_{i\in I_j} \frac{\mu_p \exp(-\psi_j)B(\psi_i - \psi_j)|\partial d_{ji}|}{|d_j|}\frac{(z_j - z_i)}{|\boldsymbol{x}_j - \boldsymbol{x}_i|} + R_j(\exp(\psi_j)w_j, \exp(-\psi_j)z_j) = 0, \tag{136}$$

for $j = 1, \ldots, N$. These, along with (134), form a nonlinear algebraic system for the nodal approximations to w, z and ψ, which can be solved either by Newton's method or by Gummel's method [24].

From the transformation in (125) we see that the Slotboom variables w and z explicitly depend exponentially on ψ. The same is true for the discrete variables w_h and z_h and for the coefficients of the nonlinear systems (135) and (136). This may cause some numerical problems when solving these systems. This drawback can be relieved by applying the inverse transformation to (125) at the discrete level, i.e., for $i = 1, \ldots, N$, we put

$$w_j = \exp(-\psi_j)n_j, \qquad \text{and} \qquad z_j = \exp(\psi_j)p_j. \tag{137}$$

Substituting these into (135) and (136) respectively we get

$$\frac{\partial n_j}{\partial t} + \gamma_n \sum_{i\in I_j} \frac{\mu_n|\partial d_{ji}|}{|d_j|} \frac{B(\psi_j-\psi_i)n_j - \exp(\psi_j-\psi_i)B(\psi_j-\psi_i)n_i}{|\boldsymbol{x}_j-\boldsymbol{x}_i|} + R_j(n_j, p_j) = 0,$$

$$\frac{\partial p_j}{\partial t} + \gamma_n \sum_{i\in I_j} \frac{\mu_p|\partial d_{ji}|}{|d_j|} \frac{B(\psi_i-\psi_j)p_j - \exp(\psi_i-\psi_j)B(\psi_i-\psi_j)p_i}{|\boldsymbol{x}_j-\boldsymbol{x}_i|} + R_j(n_j, p_j) = 0,$$

for $j = 1, \ldots, N$. From the definition (132) of the Bernoulli function it is easy to verify that $B(-t) = \exp(t)B(t)$ for any real number t. Using this the above systems become

$$\frac{\partial n_j}{\partial t} + \gamma_n \sum_{i\in I_j} \frac{\mu_n|\partial d_{ji}|}{|d_j|} \frac{B(\psi_j-\psi_i)n_j - B(\psi_i-\psi_j)n_i}{|\boldsymbol{x}_j-\boldsymbol{x}_i|} + R_j(n_j, p_j) = 0, \tag{138}$$

$$\frac{\partial p_j}{\partial t} + \gamma_n \sum_{i\in I_j} \frac{\mu_p|\partial d_{ji}|}{|d_j|} \frac{B(\psi_i-\psi_j)p_j - B(\psi_j-\psi_i)p_i}{|\boldsymbol{x}_j-\boldsymbol{x}_i|} + R_j(n_j, p_j) = 0, \tag{139}$$

for $j = 1, \ldots, N$. These systems are nonsymmetric, but have a symmetric structure, in the sense that if the entry $c_{i,j} \neq 0$ then $c_{j,i} \neq 0$.

When the three coupled nonlinear systems (134), (138) and (139) are decoupled by Gummel's algorithm, the discretized Poisson equation is linear and the two discretized continuity equations are nonlinear because of the recombination term R. If we assume that $\partial R/\partial n, \partial R/\partial p \geqslant 0$, then the linearized form of both (138) and (139) is $(A + D)\boldsymbol{x} = \boldsymbol{b}$ where D is a diagonal matrix with non-negative entries and A corresponds to the first term in (138) or (139). It is easy to show that A is a nonsymmetric M-matrix for both cases. These nonsymmetric linear systems can be solved effectively by some conjugate gradient-like methods such as the CGS or the Bi-CGSTAB methods (cf [85, 86]).

The final goal of device modelling is probably to find the current–voltage characteristic of a device. Thus, when the equations (120)–(122) are solved numerically, we need to evaluate the terminal current flowing into or out of an ohmic contact using the numerical solutions ψ_h, n_h and p_h from (134), (138) and (139). For simplicity we discuss this only for the stationary case, that is for $\partial n/\partial t = \partial p/\partial t = 0$. We assume that the Dirichlet boundary Γ_1 of the device consists of a finite number of disjoint ohmic contacts. For each $c \in \Gamma_1$, let $\{\boldsymbol{x}_i^c\}_1^{N_c}$ denote the mesh nodes on c. We also assume that the boundary Γ is meshed in such a way that for any $c \in \Gamma_1$, $c = \Gamma_1 \cap (\cup_{j=1}^{N_c} \partial d_j^c)$ where d_j^c denotes the Dirichlet tile associated with x_j^c.

Let ψ^c be a piecewise constant function satisfying

$$\psi^c(x) = \begin{cases} 1 & \text{if } x \in \cup_{j=1}^{N_c} d_j^c \\ 0 & \text{otherwise.} \end{cases} \tag{140}$$

Multiplying both (121) and (122) by ψ^c and integrating by parts we obtain

$$\int_c \boldsymbol{J}_n \cdot \boldsymbol{n} + \sum_{j=1}^{N_c} \int_{\partial d_j^c \setminus c} \boldsymbol{J}_n \cdot \boldsymbol{n} - \gamma_n^{-1} \int_\Omega R(n, p)\psi^c = 0,$$

$$\int_c \boldsymbol{J}_p \cdot \boldsymbol{n} + \sum_{j=1}^{N_c} \int_{\partial d_j^c \setminus c} \boldsymbol{J}_p \cdot \boldsymbol{n} + \gamma_p^{-1} \int_\Omega R(n, p)\psi^c = 0,$$

where $\boldsymbol{J}_n$ and $\boldsymbol{J}_p$ are current densities for n and p respectively defined in (123) and (124) or (126) and (127). Thus, the outflow currents J_n^c and J_p^c through c due to $\boldsymbol{J}_n$ and $\boldsymbol{J}_p$ are defined respectively by

$$J_n^c = \int_c \boldsymbol{J}_n \cdot \boldsymbol{n} = -\sum_{j=1}^{N_c} \int_{\partial d_j^c \setminus c} \boldsymbol{J}_n \cdot \boldsymbol{n} + \gamma_n^{-1} \int_\Omega R(n, p)\psi^c, \tag{141}$$

and

$$J_p^c = \int_c \boldsymbol{J}_p \cdot \boldsymbol{n} = -\sum_{j=1}^{N_c} \int_{\partial d_j^c \setminus c} \boldsymbol{J}_p \cdot \boldsymbol{n} - \gamma_p^{-1} \int_\Omega R(n, p)\psi^c. \tag{142}$$

Replacing $\boldsymbol{J}_n$ and $\boldsymbol{J}_p$ by the approximate fluxes $\boldsymbol{J}_{n,h}$ and $\boldsymbol{J}_{p,h}$ defined by

$$\boldsymbol{J}_{n,h} = \mu_n \exp(\psi_h)\nabla w_h, \qquad \boldsymbol{J}_{p,h} = -\mu_p \exp(-\psi_h)\nabla z_h,$$

and using the one-point quadrature rule, we obtain the following approximate outflow currents

$$\begin{aligned} J_{n,h}^c &= -\sum_{j=1}^{N_c} \int_{\partial d_j^c \setminus c} \boldsymbol{J}_{n,h} \cdot \boldsymbol{n} + \sum_{j=1}^{N_c} \frac{R_j(n_j, p_j)}{\gamma_n} |d_j^c| \\ &= \sum_{j=1}^{N_c} \left[\sum_{i \in I_j, \boldsymbol{x}_i \notin c} \frac{\mu_n \exp(\psi_j) B(\psi_j - \psi_i) |\partial d_{ji}|}{|d_j^c|} \frac{w_j - w_i}{|\boldsymbol{x}_j - \boldsymbol{x}_i|} + \frac{R_j(n_j, p_j)}{\gamma_n} |d_j^c| \right] \end{aligned} \tag{143}$$

$$\begin{aligned} J_{p,h}^c &= -\sum_{i=1}^{N_c} \int_{\partial d_i^c \setminus c} \boldsymbol{J}_{p,h} \cdot \boldsymbol{n} - \sum_{j=1}^{N_c} \frac{R_j(n_j, p_j)}{\gamma_p} |d_j^c| \\ &= \sum_{j=1}^{N_c} \left[\sum_{i \in I_j, \boldsymbol{x}_i \notin c} \frac{\mu_p \exp(-\psi_j) B(\psi_i - \psi_j) |\partial d_{ji}|}{|d_j^c|} \frac{z_j - z_i}{|\boldsymbol{x}_j - \boldsymbol{x}_i|} - \frac{R_j(n_j, p_j)}{\gamma_n} |d_j^c| \right] \end{aligned} \tag{144}$$

where n_j, p_j, w_j and z_j are related by the equations in (137). Similar expressions of $J_{n,h}^c$ and $J_{p,h}^c$ in terms of n_h and p_h can also be obtained using the transformation (137). Finally, the total outflow current through the ohmic contact c is

$$J_h^c = J_{n,h}^c + J_{p,h}^c.$$

It can be shown that this approximation satisfies

$$\sum_{c \in \Gamma_1} J_h^c = 0$$

and so the terminal currents are conservative. We omit this discussion for brevity and refer the reader to [43].

We comment that the resulting linear systems (138) and (139) from this method coincide with those from the well known box method [77] for semiconductor device simulation, which combines the finite element method on triangular meshes proposed in [87] with the one-dimensional current representation in [88]. The latter idea was also proposed independently by Allen–Southwell [52] and Il'in [53]. Using the finite element framework with inverse averages discussed in section 3.2, the above method is reformulated in [43] as a Petrov–Galerkin finite element method. Stability and error bounds for this method are then established using some standard techniques of finite element analysis, which are similar to those described in section 3.1.

Although the above method is presented here only in two dimensions, it can be easily extended to three dimensions. In fact, the systems (134), (138) and (139) and the terminal currents $J_{n,h}^c$ and $J_{p,h}^c$ in (143) and (144) are applicable in three dimensions, if we replace the mesh dependent weights $|d_j|$ and $|\partial d_{ji}|$ respectively by the corresponding weights in three dimensions, i.e. the volume of the Dirichlet tile associated with the node $\boldsymbol{x}_j$ and the area of intersection of the Voronoi polyhedra associated with the nodes $\boldsymbol{x}_j$ and $\boldsymbol{x}_i$.

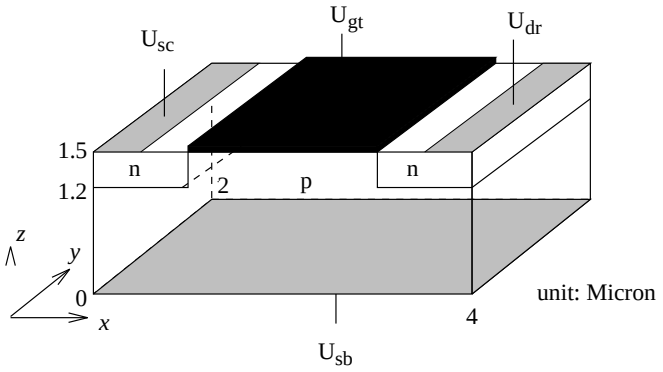

Figure 10. An MOS transistor; the ohmic contacts are shaded.

To demonstrate the effectiveness of the method, the stationary problem for the model MOS transistor depicted in figure 10 is considered. In this model problem we use the doping function

$$D(x) = \begin{cases} 10^{18}, & 0 \leqslant x \leqslant 1, \quad 0 \leqslant y \leqslant 2, \quad 1.2 \leqslant z \leqslant 1.5, \\ 2 \times 10^{18}, & 3 \leqslant x \leqslant 4, \quad 0 \leqslant y \leqslant 2, \quad 1.2 \leqslant z \leqslant 1.5, \\ -10^{16}, & \text{otherwise.} \end{cases}$$

The height of the gate is 5×10^{-3} micron and the applied biases are chosen to be

$$U_{\mathrm{sb}} = U_{\mathrm{sc}} = 0, \qquad U_{\mathrm{dr}} = 0.2\ \mathrm{V}, \qquad U_{\mathrm{gt}} = 1, 2, \ldots, 25\ \mathrm{V}.$$

The device is two-dimensional, but it is solved as a three-dimensional problem.

To solve the problem numerically for these data, we use a tetrahedral mesh obtained by dividing each brick element of a $45 \times 5 \times 22$ non-uniform hexahedral mesh into five tetrahedra. However, $|\partial d_{ji}| = 0$ in (138) and (139) for all edges not parallel to one of the coordinate

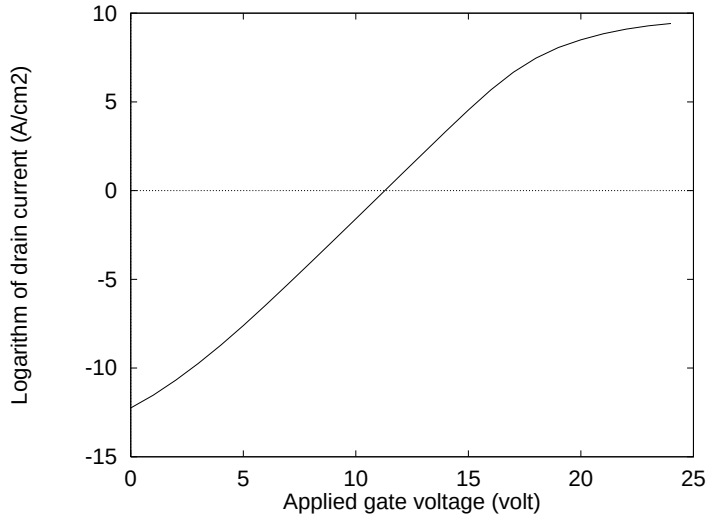

Figure 11. $I - V$ characteristic of the MOS transistor.

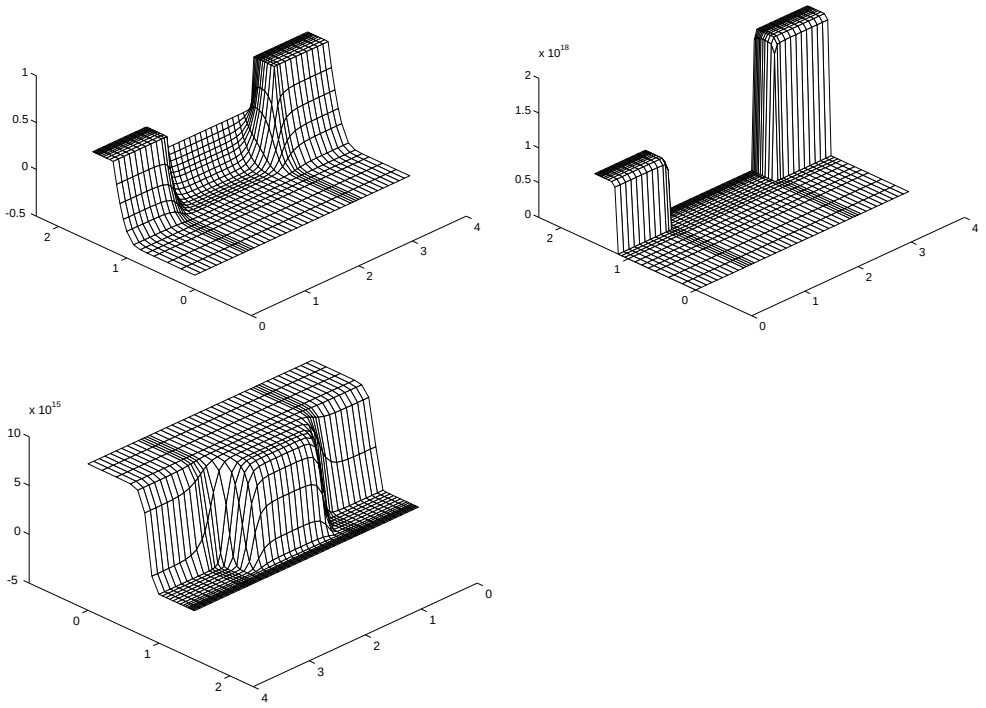

Figure 12. Plots of ψ, n and p with $U_{\mathrm{sc}} = U_{\mathrm{sb}} = 0$, $U_{\mathrm{dr}} = 0.2$ V and $U_{\mathrm{gt}} = 10$ V for the MOS transistor.

axes, and so the tetrahedral mesh is equivalent to the hexahedral mesh. The current–voltage characteristic for the applied biases is shown in figure 11, and the electrostatic potential ψ and the electron and hole concentrations n and p at the cross section $y = 1$ are plotted in figure 12. Note that the plot for p has a different orientation from that for ψ and n.

5.4. Mixed finite element methods

First the nonlinear Poisson equation is considered, which can be used on its own to simulate the behaviour of semiconductor devices in situations with low current (see [8]). Consider the simulation of a simple one-dimensional diode, with the doping function

$$D(x) = \begin{cases} -10^{18}, & x \in [0, 5 \times 10^{-4}), \\ 0, & x = 5 \times 10^{-4}, \\ 10^{18}, & x \in (5 \times 10^{-4}, 10^{-3}]. \end{cases}$$

Note that there is no difference between the potentials on the left and right contacts, so that both of the Slotboom variables can be taken to be equal to one. Hence, the equation to be solved is

$$\nabla(\varepsilon \nabla \psi) = -q(n_{\mathrm{int}} \exp(-\psi/u_T) - n_{\mathrm{int}} \exp(\psi/U_T) + D), \tag{145}$$

with the boundary conditions

$$\psi(0) = -U_T \log\left(\frac{|D(0)|}{n_{\mathrm{int}}}\right),$$

$$\psi(10^{-3}) = U_T \log\left(\frac{|D(10^{-3})|}{n_{\text{int}}}\right).$$

A discretization with the uniform mesh spacing h using a mixed finite element method leads to the following discrete equations for the Lagrange multipliers at the nodes

$$\varepsilon\left(\frac{\lambda_{i-1} - \lambda_i}{h} + \frac{\lambda_{i+1} - \lambda_i}{h}\right) = \frac{f_{i+1} + f_i}{2},$$

where

$$f_i = -qn_{\text{int}} \int_{x_i}^{x_{i+1}} \exp(-\psi/U_T) - \exp(\psi/U_T) + D/n_{\text{int}}.$$

The integrals are now approximated using the midpoint rule. The nonlinear system of equations is solved using Newton's method. In table 1 the results of the simulations on a number of different meshes are presented. The percentage of negative elements in the table indicates how many of the non-zero off-diagonal matrix elements are negative. Since the diagonal elements are positive, an M-matrix is obtained only when all off-diagonal elements are negative, that is when the percentage of negative off-diagonal elements equals 100%. The amplitude of the wiggles is a measure of the size of the oscillations that occur in the solution.

Table 1. Results of a mixed FEM applied to one-dimensional diode.

Number of subintervals	Percentage of negative off-diagonal elements	Amplitude of wiggles
10	0.0	0.318
40	0.0	0.409
100	0.0	0.270
200	0.2	0.096
400	0.5	0.0013
1000	0.6	0.0
1300	100.0	0.0

From table 1 it is seen that with a uniform mesh it is necessary to take about 1300 or more subintervals to obtain an M-matrix. If the instabilities are remedied, the numerical approximations obtained for the electrostatic potential are monotone and non-oscillatory. Furthermore, all off-diagonal elements in the system matrices are non-positive. This is not surprising, since these properties were already established theoretically.

In [62], the mixed finite element method using the lowest-order Raviart–Thomas elements is applied to the current continuity equations using a prescribed electrostatic potential. Since zero recombination is used (right-hand side identically equal to zero), a discrete maximum principle holds. This is confirmed by the numerical results. However, as has been demonstrated in section 4.4, problems may be expected if the recombination term is present. This problem can be circumvented by using a non-standard mixed finite element method, which is described in the next section.

5.5. Non-standard mixed finite element methods

To remedy the problems associated with a non-zero recombination term, the approach described at the end of section 4.4 can be used. Special quadrature rules ensure that the mass matrix has exactly the right properties, and proofs can be given that the coefficient matrix of the system for the unknown potentials and/or carrier concentrations has the M-property. This has been described in detail in [65,67]. A remaining problem is that, in the case of a triangular mesh, the

triangles must all be non-obtuse. This is in contrast to the finite volume method, where obtuse triangles are allowed as long as they are compensated for by sufficiently acute neighbouring triangles [8, 9].

Another remedy, which is similar to the one in the previous paragraph, is suggested in [89]. Here, also, quadrature rules are used to obtain essentially diagonal mass matrices. Since the methods developed in this way are similar to classical Galerkin methods, error estimates are obtained immediately. The technique is also shown to be applicable to nonlinear problems, as is confirmed by the numerical results obtained when it is applied to a number of simple semiconductor devices.

A different approach, which was developed in the context of semiconductor device simulation, is to design new basis functions. Hence, the Raviart–Thomas elements are replaced by new triangular elements. In [90, 91], one such approach is described. Instead of the normal set of piecewise linear vector basis functions $\{\tau_1, \tau_2, \tau_3\}$ satisfying

$$\int_{e_i} \boldsymbol{\tau}_j \cdot \boldsymbol{n} = \delta_{ij},$$

Brezzi *et al* [90] suggest the use of τ_1 and τ_2 which are constant on an element (and orthogonal to each other), and a piecewise linear vector basis function τ_3 which satisfies

$$\int_{e_i} \boldsymbol{\tau}_3 \cdot \boldsymbol{n} = \delta_{i3},$$

and the special requirement

$$\int_T \boldsymbol{\tau}_1 \cdot \boldsymbol{\tau}_3 = \int_T \boldsymbol{\tau}_2 \cdot \boldsymbol{\tau}_3 = 0.$$

Clearly, this ensures that the mass matrix is a diagonal matrix. The disadvantage of this approach is that the role of the edges is no longer symmetric. In each triangle, one of the edges must be chosen to play a special role. The common edge of two neighbouring triangles may play this role in one triangle, but not in the other triangle. Nevertheless, the approach leads to coefficient matrices with the desired properties, and the numerical results in [90] give rise to some optimism that this method may work for more complicated problems. In [93], a more detailed analysis of these new basis functions is given, and numerical results for the full system of drift-diffusion equations are presented. One of the conclusions of this work is that the special role of some of the edges may lead to numerical problems, which can sometimes be avoided by a simple renumbering of the element edges (such that other edges play the special role). This was observed, for example, for a bipolar transistor under high forward bias conditions. To date, these problems have not been investigated any further.

5.6. The mixed finite element method with inverse averages

In this section we discuss the application of the second non-standard mixed finite element method from section 4.3 to the device equations (120)–(122). For brevity we consider only the case of the steady-state equations in two dimensions, i.e. $\partial n/\partial t = \partial p/\partial t = 0$ in (121) and (122). The time-dependent problem can be treated in the same way as that in section 5.2 and the three-dimensional extension can be found in [69].

The Poisson equation (120) has exactly the same form as that of (72) with $a = 1$, $b = 0$ and $f = \alpha(p - n + D)$. Thus, applying (112) to this equation we get

$$\sum_{i \in I_j} \frac{|\partial d_{ji}|}{|d_j|} \frac{\psi_j - \psi_i}{|\boldsymbol{x}_j - \boldsymbol{x}_i|} = \alpha(p_j - n_j + D_j), \tag{146}$$

for $j = 1, \ldots, N$. This linear algebraic system is identical to (134).

The two continuity equations (121) and (122) in the Slotboom variables (125) have the same form as (72) with a replaced by $\exp(\psi)$ and $\exp(-\psi)$ respectively. Thus, the application of (112) to these two equations yields

$$\sum_{i \in I_j} \overline{\exp(\psi_h)}_{ji} \frac{|\partial d_{ji}|}{|d_j|} \frac{w_j - w_i}{|\boldsymbol{x}_j - \boldsymbol{x}_i|} + R_j(\exp(\psi_j) w_j, \exp(\psi_j) z_j) = 0, \tag{147}$$

and

$$\sum_{i \in I_j} \overline{\exp(-\psi_h)}_{ji} \frac{|\partial d_{ji}|}{|d_j|} \frac{z_j - z_i}{|\boldsymbol{x}_j - \boldsymbol{x}_i|} + R_j(\exp(\psi_j) w_j, \exp(\psi_j) z_j) = 0, \tag{148}$$

for $j = 1, \ldots, N$, where ψ_h denotes the numerical solution of (146) and $\overline{\exp(\psi_h)}_{ji}$ and $\overline{\exp(-\psi_h)}_{ji}$ are the inverse averages of $\exp(\psi_h)$ and $\exp(-\psi_h)$ on b_{ji} defined in (110). Obviously (146)–(148) form a coupled nonlinear algebraic system for ψ_h, w_h and z_h. As in section 5.2, this system can be decoupled by Gummel's method and the decoupled nonlinear sub-systems (147) and (148) can then be linearized by a Newton-like method. All these decoupled and linearized sub-systems are symmetric and their coefficient matrices are M-matrices if the recombination-generation term $R(n, p)$ satisfies $\partial R/\partial n, \partial R/\partial p \geqslant 0$.

It now remains to evaluate the inverse averages $\overline{\exp(\psi_h)}_{ji}$ and $\overline{\exp(-\psi_h)}_{ji}$ in (147) and (148). From the construction of the region b_{ji} in section 3.3 (cf figure 8), we see that it consists of two triangles sharing the edge joining $\boldsymbol{x}_j$ and $\boldsymbol{x}_i$. We denote the two triangles by t_{ji}^1 and t_{ji}^2 and assume that they have vertices $\boldsymbol{x}_j$, $\boldsymbol{x}_i$, $\boldsymbol{x}_{ji}^1$ and $\boldsymbol{x}_j$, $\boldsymbol{x}_i$, $\boldsymbol{x}_{ji}^2$ respectively. From the construction of the meshes we know that $\boldsymbol{x}_{ji}^1$ and $\boldsymbol{x}_{ji}^2$ are the circumcentres of the two triangular elements sharing the edge connecting $\boldsymbol{x}_j$ and $\boldsymbol{x}_i$. Since the numerical solution ψ_h from (146) is piecewise linear on Ω, it is linear on each of the two triangles. Thus, we can rewrite the averages of $\exp(\psi_h)$ and $\exp(-\psi_h)$ on b_{ji} in the form

$$\overline{\exp(\psi_h)}_{ji} = \left[\frac{1}{|b_{ji}|} \sum_{k=1}^{2} |t_{ji}^k| \left(\frac{1}{|t_{ji}^k|} \int_{t_{ji}^k} \exp(-\psi_h) \right) \right]^{-1} \tag{149}$$

and

$$\overline{\exp(-\psi_h)}_{ji} = \left[\frac{1}{|b_{ji}|} \sum_{k=1}^{2} |t_{ji}^k| A_{t_{ji}^k}(\exp(\psi_h)) \right]^{-1}. \tag{150}$$

This shows that $\overline{\exp(\psi_h)}_{ji}$ and $\overline{\exp(-\psi_h)}_{ji}$ can be expressed respectively as linear combinations of the averages of $\exp(-\psi_h)$ and $\exp(\psi_h)$ on the two triangles t_{ji}^k $(k = 1, 2)$.

Let us now consider the evaluation of the average $A_t(\exp(\phi))$ of $\exp(\phi)$ on an arbitrary triangle t with vertices $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ and $\boldsymbol{x}_3$ where ϕ is a linear function on t. That is

$$A_t(\exp(\phi)) = \frac{1}{|t|} \int_t \exp(\phi(\boldsymbol{x}))\, \mathrm{d}x$$

where $|t|$ denotes the area of t, $\boldsymbol{x} = (x_1, x_2)$ and $\mathrm{d}x = \mathrm{d}x_1\, \mathrm{d}x_2$. Consider a linear coordinate transformation from $\boldsymbol{s}_1 = (0, 0)$, $\boldsymbol{s}_2 = (1, 0)$ and $\boldsymbol{s}_3 = (0, 1)$ in the $\boldsymbol{s}$-plane to the vertices $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ and $\boldsymbol{x}_3$ of t of the form

$$\boldsymbol{x} = \boldsymbol{c} + D\boldsymbol{s}$$

where $\boldsymbol{c}$ is a 2×1 vector and D is a 2×2 matrix. Note that the elements of $\boldsymbol{c}$ and D are uniquely determined by the vertices $\boldsymbol{x}_i$ and $\boldsymbol{s}_i$, $i = 1, 2, 3$. This transformation establishes a

1-1 correspondence between t and the reference triangle $\hat{t}$ with vertices $\boldsymbol{s}_1$, $\boldsymbol{s}_2$ and $\boldsymbol{s}_3$. Using this transformation the above average can be rewritten as

$$\begin{aligned} A_t(\exp(\phi)) &= \frac{1}{|\hat{t}|\det(D)}\int_{\hat{t}} \det(D)\exp(\phi(\boldsymbol{s}))\,\mathrm{d}s_1\,\mathrm{d}s_2 \\ &= \frac{1}{|\hat{t}|}\int_{\hat{t}} \exp(\phi(\boldsymbol{x}))\,\mathrm{d}s_1\,\mathrm{d}s_2, \end{aligned}$$

where det D is the Jacobian of the transformation. Since $\phi(\boldsymbol{x})$ is linear on t, $\phi(\boldsymbol{s})$ is linear on $\hat{t}$, and so we assume that

$$\phi(\boldsymbol{s}) = \alpha_1 + \alpha_2 s_1 + \alpha_3 s_2$$

where α_i $(i = 1, 2, 3)$ are constants. Substituting this into the above integral, and noting that $|\hat{t}| = 1/2$, we have

$$\begin{aligned} A_t(\exp(\phi)) &= 2\int_0^1 \int_0^{1-s_1} \exp(\alpha_1 + \alpha_2 s_1 + \alpha_3 s_2)\,\mathrm{d}s_1\,\mathrm{d}s_2 \\ &= \frac{2\exp(\alpha_1)}{\alpha_3}\int_0^1 \exp(\alpha_2 s_1)(\exp(\alpha_3(1-s_1)) - 1)\,\mathrm{d}s_1 \\ &= \frac{2\exp(\alpha_1)}{\alpha_3}\left(\exp(\alpha_3)\frac{\exp(\alpha_2-\alpha_3)-1}{\alpha_2-\alpha_3} - \frac{\exp(\alpha_2)-1}{\alpha_2}\right) \\ &= \frac{2\exp(\alpha_1)}{\alpha_3}(\exp(\alpha_3)B^{-1}(\alpha_2-\alpha_3) - B^{-1}(\alpha_2)) \end{aligned}$$

where $B(\cdot)$ is the Bernoulli function defined in (132). Writing $\phi_i = \phi(\boldsymbol{x}_i)$ for $i = 1, 2, 3$, it is easy to verify that

$$\alpha_1 = \phi_1, \qquad \alpha_2 = \phi_2 - \phi_1 \qquad \text{and} \qquad \alpha_2 = \phi_3 - \phi_1,$$

and

$$A_t(\exp(\phi)) = \frac{2}{\phi_3 - \phi_1}(\exp(\phi_3)B^{-1}(\phi_2 - \phi_3) - \exp(\phi_1)B^{-1}(\phi_2 - \phi_1)). \tag{151}$$

From this we see that $\phi_1 \neq \phi_3$. If $\phi_1 = \phi_3$, we can permute the vertices $\boldsymbol{x}_i$ $(i = 1, 2, 3)$ so that $\phi_1 \neq \phi_3$. If $\phi_1 = \phi_2 = \phi_3$, the average becomes $\exp(\phi_1)$. Also, when the absolute values of the differences between ϕ_i $(i = 1, 2, 3)$ are small, it is necessary to use Taylor expansions about zero to obtain accurate evaluations of the right-hand side of (151). A discussion of this for tetrahedral elements can be found in [69].

Replacing t by t_{ji}^k and ϕ in (151) by $-\psi_h$ and ψ_h respectively we obtain

$$A_{t_{ji}^k}(\exp(-\psi_h)) = \frac{2}{\psi_j - \psi_{ji}^k}(\exp(-\psi_{ji}^k)B^{-1}(\psi_{ji}^k - \psi_i) - \exp(-\psi_j)B^{-1}(\psi_j - \psi_i)),$$

$$A_{t_{ji}^k}(\exp(\psi_h)) = \frac{2}{\psi_{ji}^k - \psi_j}(\exp(\psi_{ji}^k)B^{-1}(\psi_i - \psi_{ji}^k) - \exp(\psi_j)B^{-1}(\psi_i - \psi_j)),$$

where $\psi_k = \psi_h(\boldsymbol{x}_k)$ and $\psi_{ji}^k = \psi_h(\boldsymbol{x}_{ji}^k)$ for $k = 1, 2$. Substituting these into (149) and (150) respectively we obtain the inverse averages in (147) and (148).

Comparing (147) and (148) with the systems (138) and (139) we see that they have the same form, but the inverse averages of $\exp(\psi_h)$ and $\exp(-\psi_h)$ are different. The inverse averages in (138) and (139) are evaluated along the edge joining $\boldsymbol{x}_j$ and $\boldsymbol{x}_i$ but the inverse averages in (147) and (148) are evaluated over the region b_{ji} containing that edge. Obviously the former can be regarded as approximations or lumped versions of the latter. Therefore, we may expect that the latter give better approximations to the coefficient functions.

The evaluation of the terminal currents for this scheme is similar to that in section 5.4. In fact, using the same notation and argument as in section 5.4, we see that the terminal currents flowing out of an ohmic contact c are given by (141) and (142). From the construction of the space S_h for the second method in section 4.3 and (111) we have that the approximate current densities obtained from this method are

$$\boldsymbol{J}_{n,h} = \sum_{b_{ji}} \overline{\exp(\psi_h)}_{ji} \frac{w_j - w_i}{|\boldsymbol{x}_i - \boldsymbol{x}_j|},$$

$$\boldsymbol{J}_{p,h} = \sum_{b_{ji}} \overline{\exp(-\psi_h)}_{ji} \frac{z_j - z_i}{|\boldsymbol{x}_i - \boldsymbol{x}_j|}.$$

Thus, replacing $\boldsymbol{J}_n$ and $\boldsymbol{J}_p$ in (141) and (142) by these approximations respectively and using the one-point quadrature rule, we have the approximate terminal currents

$$\begin{aligned} J^c_{n,h} &= -\sum_{j=1}^{N_c} \int_{\partial d^c_j \setminus c} \boldsymbol{J}_{n,h} \cdot \boldsymbol{n} + \sum_{j=1}^{N_c} \frac{R_j(n_j, p_j)}{\gamma_n} |d^c_j| \\ &= \sum_{j=1}^{N_c} \left[\sum_{i \in I_j, \boldsymbol{x}_i \notin c} \frac{\mu_n \overline{\exp(\psi_h)}_{ji} \partial d_{ji}|}{|d_j|} \frac{(w_j - w_i)}{|\boldsymbol{x}_j - \boldsymbol{x}_i|} + \frac{R_j(n_j, p_j)}{\gamma_n} |d^c_j| \right], \\ J^c_{p,h} &= -\sum_{i=1}^{N_c} \int_{\partial d^c_i \setminus c} \boldsymbol{J}_{p,h} \cdot \boldsymbol{n} - \sum_{j=1}^{N_c} \frac{R_j(n_j, p_j)}{\gamma_p} |d^c_j| \\ &= \sum_{j=1}^{N_c} \left[\sum_{i \in I_j, \boldsymbol{x}_i \notin c} \frac{\mu_p \overline{\exp(-\psi_h)}_{ji} |\partial d_{ji}|}{|d_j|} \frac{(z_j - z_i)}{|\boldsymbol{x}_j - \boldsymbol{x}_i|} - \frac{R_j(n_j, p_j)}{\gamma_n} |d^c_j| \right], \end{aligned}$$

where the inverse averages are defined in (149) and (150). Finally, the total computed current J^c_h flowing out of the contact c is given by

$$J^c_h = J^c_{n,h} + J^c_{p,h}.$$

Similarly to the case in section 5.4, it can be shown that

$$\sum_{c \in \Gamma_1} J^c_h = 0,$$

and so the computed terminal currents are conservative.

To demonstrate the usefulness of the scheme in practice, we present some numerical results from [69] for the three-dimensional diode depicted in figure 13. The device is doped uniformly with a concentration of 10^{16} cm^{-3} in both the n and p regions, so that the doping function D equals 10^{16} in the n region and -10^{16} in the p-region. A zero bias is maintained on the cathode.

To solve this problem we use a tetrahedral mesh obtained by dividing each element of a non-uniform brick mesh with 6859 ($19 \times 19 \times 19$) nodes into five tetrahedra. Element edges, that are not perpendicular to one of the axes, contribute nothing to the system matrix because $|\partial d_{ji}| = 0$ in (146)–(148) for each of these edges. The computed values of n and p at two cross sections for 0.5 V forward bias are plotted in figure 14 and the computed values of the anode current density for various applied biases are shown in figure 15. The graph of the latter is linear and its slope $\approx 38.6 \approx (q/kT)$, where q is the electronic charge, k is Boltzmann's constant and T is the absolute temperature, which is taken to be 300 K (room temperature). This is in agreement with the one-dimensional theoretical results (see, for example, [92]).

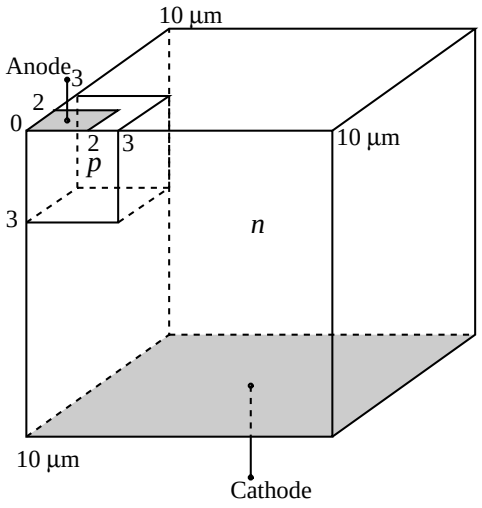

Figure 13. A three-dimensional diode; the ohmic contacts are shaded.

5.7. Miscellaneous

The methods described in the foregoing sections can be regarded as the main developments in the application of finite element methods and mixed finite element methods to the simulation of the behaviour of semiconductor devices. As mentioned before, we have not attempted to give a complete account of the literature on the subject. Instead, we believe that it is enlightening to give a brief survey of activities in this area, especially because there is now an increased interest in the use of mixed finite element methods for the semiconductor device problem. One reason for this is the recognition that mixed finite element methods are similar to finite volume methods, the preferred discretization technique for semiconductor device simulation. In the following, some additional remarks are made about the use of finite element methods for semiconductor device simulation.

Early papers on finite element methods for semiconductor device simulation are [73, 94], but it appears that some of the conclusions given in these papers are premature. The authors use linear triangular elements and Hermite bicubic elements, and claim that current conservation holds exactly. In our opinion, this is only the case in the limit as the mesh spacing approaches zero. Nevertheless, it is interesting to see that as early as 1974 the possibility of applying finite element methods was considered.

Finite element methods can also be used for the off-state semiconductor problem formulated as a free boundary problem. In this case, the problem domain is divided into a number of subdomains, some of which are assumed to be entirely depleted of carriers, and others to contain no space charge. The mathematical model then consists of the Poisson equation, which has either a zero right-hand side or a right-hand side which is equal to the doping function. The boundaries between the subdomains are unknown, and the determination of these is part of the problem. In [66], a mixed finite element method is applied to the solution of such problems. It is shown how a variational formulation can be obtained, and how a mixed method using the lowest-order Raviart–Thomas elements can be used. The algorithm makes use of a formulation in terms of Lagrange multipliers.

Three-dimensional problems have not been discussed in this paper, mainly because so far the number of papers on this specific subject is rather limited. One paper addressing this topic

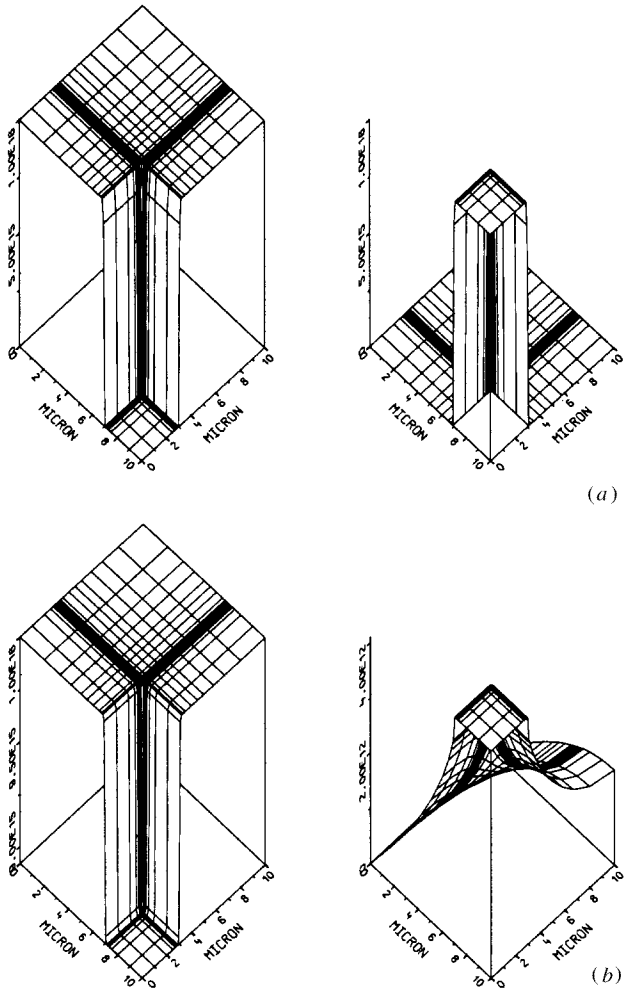

Figure 14. Electron and hole concentrations for the diode at 0.5 V forward bias: (*a*) the section at $y = 0$ μm (*b*) the section at $y = 3.1$ μm.

is [95], where a mixed finite element method is proposed based on hexahedral elements and Van Welij edge elements [96]. In [69], a tetrahedral mixed finite element method is proposed which belongs to the class of inverse average type schemes discussed in section 4.3. The method is shown to lead to a symmetric and positive definite coefficient matrix which satisfies the M-property. Furthermore, as has been shown in the two-dimensional case, the currents are convergent and conservative.

Several authors have suggested patch tests as a validation for finite element methods applied to the discretization of semiconductor device problems. Patch tests were originally developed in the area of mechanics [97]. One of the first patch tests for the semiconductor device problem was suggested in [98]. Later, similar tests have been described in [99–101].

An entirely different application of mixed finite element methods is suggested in [102,103] and, more recently, in [104]. In these references, a finite volume method is used for the discretization of the semiconductor device problem. However, to find a suitable representation

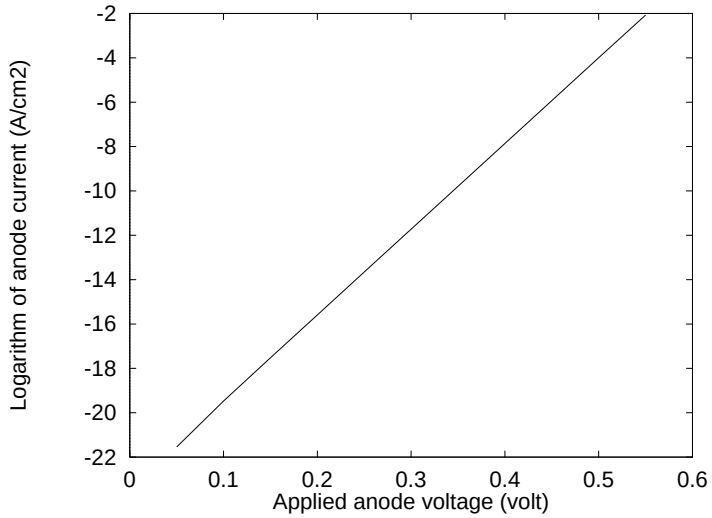

Figure 15. I–V characteristic of the diode at forward biases, slope $\approx q/KT$.

of the current densities and the electric field, a mixed finite element representation is suggested. This technique solves the frequently encountered problem of the representation of these quantities, and can either be used as an interpolation tool for postprocessing purposes, or as a means of obtaining a global representation for the electric field needed in a mixed finite element discretization of the current continuity equations. In the latter case, the mixed finite element method can be used to discretize Poisson's equation, whereas the continuity equations can be discretized using the finite volume method. Such combined approaches are becoming increasingly popular, the advantage being that the mixed finite element approximation of the electrostatic potential yields a globally valid representation. More information on this technique can be found in [104].

Finite element methods have also been suggested for the hydrodynamic problem [105, 106]. This work is motivated by the fact that the time-dependent hydrodynamic equations are similar to the Euler equations of gas dynamics. The only difference is in the second order terms occurring in the equations for the carrier temperatures, terms which are not present in the Euler equations. Whereas the Euler equations constitute a hyperbolic system of equations, the hydrodynamic system of equations is said to be incompletely parabolic, indicating that the second-order coefficient matrix is rank deficient (see [9] and the references therein). Because of the resemblance to the Euler equations, it is expected that methods from the area of computational fluid dynamics may also be applicable to hydrodynamic device simulations. For this reason, the application of streamline upwind Petrov–Galerkin methods [107, 108] was investigated. Details of this approach are given in [105, 106]. In the former paper, only unipolar examples are given, whereas the second paper also contains numerical results for bipolar devices. The latter are important in view of the fact that many methods may work for unipolar devices, but often fail for bipolar devices. For this reason, simulations of a simple one-dimensional diode are often enlightening. Both references provide much detail about the method employed, and are strongly recommended for further reading.

Other papers devoted to the solution of the hydrodynamic model using finite element methods are [109–112]. The finite element approach adopted by Chen [112], which was originally described for the classical hydrodynamic model in [111], consists of discretizing

Poisson's equation using the mixed finite element method with the lowest order Raviart–Thomas elements. This implies that the electric field is approximated with piecewise linear functions on the elements. This representation of the electric field is then used in the discrete equations corresponding to the quantum hydrodynamic equations. The latter are not discretized by a mixed finite element method, instead a Runge–Kutta discontinuous Galerkin method is employed.

Similar combinations of methods, using the finite element method only for the discretization of Poisson's equation to obtain a global representation of the electric field, have also been described for other types of semiconductor device simulations. In [113], a combination of a mixed finite element method with an alternating direction implicit (ADI) method is presented for the modelling of semiconductor heterojunction structures. In [114], a standard quadratic finite element method is combined with a particle simulation method, such as the Monte Carlo method.

6. Conclusion

Mathematical models of semiconductor device behaviour lead to singular perturbation problems for complicated nonlinear systems of parabolic and elliptic partial differential equations. For industrial design purposes there is a need for parameter-robust numerical solutions of these problems with guaranteed pointwise accuracy. The standard finite element methods used, for example, in structural engineering do not provide adequate numerical solutions. Today, most successful numerical methods for these semiconductor device problems are based on the Scharfetter–Gummel finite difference method and its variants and extensions. Inverse averaging techniques are useful for generating finite element variants of the Scharfetter-Gummel method for one-and two-dimensional problems. However all such variants are fitted operator methods and do not provide parameter-robust numerical solutions with guaranteed pointwise accuracy unless they are used in conjunction with an appropriate non-uniform mesh. To obtain such numerical solutions monotone numerical methods, finite difference or finite element, on special non-uniform Shishkin meshes are required. The development of such methods is in its infancy and is undoubtedly one of the most promising directions for future numerical research on these problems.

References

[1] Markowich P A 1984 A singular perturbation analysis of the fundamental semiconductor device equations *SIAM J. Appl. Math.* **5** 896–928
[2] Miller J J H, O'Riordan E and Shishkin G I 1996 *Fitted Numerical Methods for Singular Perturbation Problems* (Singapore: World Scientific)
[3] Frederiksen T M 1982 *Intuitive IC Electronics* (New York: McGraw-Hill)
[4] Sze S M 1981 *Physics of Semiconductor Devices* (New York: Wiley)
[5] Sze S M 1983 *VLSI Technology* (Singapore: McGraw-Hill)
[6] Cercignani C 1975 *Theory and Application of the Boltzmann Equation* (Edinburgh: Scottish Academic)
[7] Markowich P A, Ringhofer C A, and Schmeiser C 1990 *Semiconductor Equations* (Vienna: Springer)
[8] Schilders W H A 1998 *Numerical Methods for Semiconductor Device Simulation* vol 1 (Vienna: Springer) in press
[9] Schilders W H A 1998 *Numerical Methods for Semiconductor Device Simulation* vol 2 (Vienna: Springer) in press
[10] Mock M S 1972 On equations describing steady state carrier distributions in a semiconductor device *Comment. Pure Appl. Math.* **25** 781–92
[11] Mock M S 1974 An initial value problem from semiconductor device theory *SIAM J. Math. Anal.* **5** 597–612

[12] Mock M S 1982 An example of nonuniqueness of stationary solutions in semiconductor device models *COMPEL* **1** 165–74
[13] Mock M S 1983 *Analysis of Mathematical Models of Semiconductor Devices* (Dublin: Boole)
[14] Markowich P A 1986 *The Stationary Semiconductor Device Equations* (Vienna: Springer)
[15] Brezzi F 1986 Theoretical and numerical problems in semi-conductor devices *Numerical Analysis* ed D F Griffiths and G A Watson pp 24–31
[16] Gajewski H and Gröger K 1984 On the basic equations for carrier transport in semiconductors *Report* P-MATH-39/84, Akad. Wissenschaften DDR, Berlin
[17] Gajewski H 1985 On uniqueness and stability of steady-state carrier distributions in semiconductors *Proceedings Equadiff Conf. 1985* (Berlin: Springer)
[18] Jerome J W 1985 Consistency of semiconductor modelling: an existence/stability analysis for the stationary Van Roosbroeck system, *SIAM J. Appl. Math.* **45** 565–90
[19] Kerkhoven T 1988 A spectral analysis of the decoupling algorithm for semiconductor simulation *SIAM J. Numer. Anal.* **25** 1299–312
[20] Kerkhoven T 1990 Efficiency and acceleration of steady-state decoupling algorithms *Lectures in Appl. Math.* **25** 151–8
[21] Seidman T I 1980 Steady state solutions of diffusion reaction systems with electrostatic convection *Nonlinear Anal.* **4** 623–37
[22] Seidman T I 1986 Time-dependent solutions of a nonlinear system arising in semiconductor theory—II. Boundedness and periodicity *Nonlinear Anal.* **10** 491–502
[23] Gilbarg D and Trudinger N S 1984 *Elliptic Partial Differential Equations of Second Order* (Berlin: Springer)
[24] Gummel H K 1964 A self-consistent iterative scheme for one-dimensional steady state transistor calculations,*IEEE Trans. Electron Devices* **11** 455–65
[25] Hestenes M R and Stiefel E 1952 Methods of conjugate gradients for solving linear systems *J. Res. Natl Bur. Stand.* **49** 409–36
[26] Reid J K 1971 On the method of conjugate gradients for the solution of large sparse systems of linear equations *Large Sparse Sets of Linear Equations*, ed J K Reid (New York: Academic) pp 231–54
[27] Ciarlet P G 1978 *The Finite Element Method for Elliptic Problems* (Amsterdam: North-Holland)
[28] Ciarlet P G and Lions J L 1991 *Handbook of Numerical Analysis* vols II and III (Amsterdam: North Holland)
[29] Temam R 1977 *Navier–Stokes Equations* (Amsterdam: North Holland)
[30] Strang G and Fix G J 1973 *An Analysis of the Finite Element Method*, (Englewood Cliffs, NJ: Prentice-Hall)
[31] Oden J T and Reddy J N 1976 *An Introduction to the Mathematical Theory of Finite Elements* (New York: Wiley–Interscience)
[32] Axelsson O and Barker V A 1984 *Finite Element Solution of Boundary Value Problems* (New York: Academic)
[33] Silvester P P and Ferrari R L 1990 *Finite Elements for Electrical Engineers* 2nd edn (Cambridge: Cambridge University Press)
[34] Johnson C 1987 *Numerical Solutions of Partial Differential Equations by the Finite Element Method* (Cambridge: Cambridge University Press)
[35] Babuška I and Aziz A K 1972 Survey lectures on the mathematical foundations of the finite element method, *The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations* (New York: Academic)
[36] Brenner S C and Scott L R 1994 *The Mathematical Theory of Finite Element Methods* (New York: Springer)
[37] Fletcher C A J 1984 *Computational Galerkin Methods* (New York: Springer)
[38] Christie I, Griffiths D F, Mitchell A R and Zienkiewicz O C 1976 Finite element methods for second order differential equations with significant first derivatives *Int. J. Num. Meth. Eng.* **10** 1389–96
[39] Heinrich J C, Huyakorn P S, Mitchell A R, and Zienkiewicz O C 1977 An upwind finite element scheme for two-dimensional convective transport equations, *Int. J. Num. Meth. Eng.* **11** 131–43
[40] Miller J J H 1976 Construction of a FEM for a singularly perturbed problem in 2 dimensions *Numerische Behandlung von Differential-gleichungen, insbesondere mit der Methode der finiten Elemente* Conference Oberwolfach, ISNM, vol 31 (Birkhäuser) pp 165–9
[41] Barrett J W and Morton K W 1980 Optimal finite element solutions to diffusion-convection problems in one dimension *Int. J. Num. Meth. Eng.* **15** 1457–74
[42] Hughes T J R 1979 *Finite Element Methods for Convection Dominated Flows, AMD* vol 34 (New York: American Society of Mechanical Engineers)
[43] Miller J J H and Wang S 1994 An analysis of the Scharfetter–Gummel box method for the stationary semiconductor device equations, *RAIRO Modél. Math. Anal. Numér.* **28** 123–40
[44] Babuška I and Osborn J E 1978 Numerical Treatment of eigenvalue problems for differential equations with discontinuous coefficients *Math. Comput.* **32** 991–1023

[45] Hemker P W 1977 A numerical study of stiff two-point boundary value problems *PhD Thesis* Mathematical Centre, Amsterdam
[46] Barrett J W and Morton K W 1984 Approximate symmetrization and Petrov–Galerkin methods for diffusion-convection problems *Comput. Meth. Appl. Mech. Eng.* **45** 97–122
[47] Morton K W 1996 Numerical solution of convection-diffusion problems (London: Chapman and Hall)
[48] Hughes T J R and Brooks A N 1979 A multidimensional upwind scheme with no crosswind diffusion *Finite Element Methods for Convection Dominated Flows (AMD)* vol 34, ed T J R Hughes (New York: American Society of Mechanical Engineers)
[49] Varga R S 1962 *Matrix Iterative Analysis* (Englewood Cliffs, NJ: Prentice-Hall)
[50] Doolan E P, Miller J J H, and Schilders W H A 1980 *Uniform Numerical Methods for Problems with Initial and Boundary Layers* (Dublin: Boole)
[51] Protter M H and Weinberger H F 1984 *Maximum Principles in Differential Equations* (New York: Springer)
[52] Allen D N de G and Southwell R V 1955 Relaxation methods applied to determine the motion, in 2-D, of a viscous fluid past a fixed cylinder *Q. J. Mech. Appl. Math.* **VIII** 129–45
[53] Il'in A M 1969 Differencing scheme for a differential equation with a small parameter affecting the highest derivative **Math. Notes 6** 596–602
[54] Brezzi F 1974 On the existence, uniqueness and approximation of saddle-point problems arising from Lagrangian multipliers, *RAIRO Modél. Math. Anal. Numér.* **8** 129–51
[55] Babuška I 1971 Error bound for the finite element method *Num. Math.* **16** 322–33
[56] Miller J J H and Wang S 1991 A triangular mixed finite element method for the stationary semiconductor device equations *RAIRO Modél. Math. Anal. Numér.* **25** 441–63
[57] Arnold D N and Brezzi F 1985 Mixed and non-conforming finite element methods: implementation, postprocessing and error estimates *RAIRO Modél. Math. Anal. Numér.* **19** 7–32
[58] Thomas J M 1977 Sur l'analyse numerique des methodes d'elements finis hybrides et mixtes *PhD Thesis* Univ. Pierre et Marie Curie, Paris
[59] Brezzi F and Fortin M 1991 *Mixed and Hybrid Finite Element Methods* (Berlin: Springer)
[60] Nedelec J C 1980 Mixed finite elements in R^3 *Numer. Math.* **35** 315–41
[61] Raviart P A and Thomas J M 1977 A mixed finite element method for second order elliptic problems *Mathematical Aspects of the Finite Element Method (Lecture Notes in Mathematics 606)* (Berlin: Springer) pp 292–315
[62] Brezzi F, Marini L D and Pietra P 1987 Méthodes d'eléments finis mixtes et schéma de Scharfetter–Gummel, *C. R. Acad. Sci. I, Math.* **305** 599–604
[63] Brezzi F, Marini L D and Pietra P 1989 Numerical simulation of semiconductor devices *Comput. Methods Appl. Mech. Eng.* **75** 493–514
[64] Couperus H D 1988 Mixed finite element methods applied to the one- and two-dimensional stationary semiconductor equations *Master's Thesis* Utrecht University, The Netherlands
[65] Gulikers P J A 1989 Mixed finite element methods applied to the 2-D stationary semiconductor equations *Master's Thesis* Technical University of Eindhoven, Eindhoven
[66] Marini L D and Savini A 1984 Accurate computation of electric field in reverse-biased semiconductor devices: a mixed finite-element approach *COMPEL* **3** 123–35
[67] Polak S J 1988 Mixed FEM for $\Delta u = \alpha u$ *Philips Internal Report* UDR/MSW/088/SP053
[68] Polak S J, Schilders W H A and Couperus H D 1988 A finite element method with current conservation *Proc. SISDEP 3 Conf.* ed G Baccarani and M Rudan (Bologna: Tecnoprint)
[69] Miller J J H and Wang S 1994 A tetrahedral mixed finite element method for the stationary semiconductor device equations *SIAM J. Numer. Anal.* **31** 196–216
[70] Fraeijs de Veubeke B X 1965 Displacement and equilibrium models in the finite element method *Stress Analysis* ed O C Zienkiewicz and G Hollister (New York: Wiley)
[71] Cottrell P E and Buturla E M 1975 Steady state analysis of field effect transistors via the finite element method *Technical Digest IEDM-75* (New York: Institute of Electrical and Electronic Engineers) pp 51–4
[72] Buturla E M, Cottrell P E, Grossman B M, Salsburg K A, Lawlor M B and McMullen C T 1980 Three-dimensional finite element simulation of semiconductor devices used in ICs *1980 IEEE International Solid-State Circuits Conference Digest of Technical Papers* (New York: Institute of Electrical and Electronic Engineers) pp 76–7
[73] Barnes J J and Lomax R J 1974 Two-dimensional finite element simulation of semiconductor devices *Electron. Lett.* **10** 341–3
[74] Hachtel G D, Mack M H, O'Brien R R and Quinn H F 1976 Two-dimensional finite element modeling of NPN devices *IEDM-76 Technical Digest* (New York: Institute of Electrical and Electronic Engineers) pp 166–9
[75] Hohl J H 1978 Variational principles for semiconductor device modelling with finite elements *IBM J. Res. Dev.*

22 159–67

[76] Lomax R J and Oh S Y 1978 Finite elements for semiconductor device simulation *11th Annual Asimolar Conf. on Circuits, Systems and Computers* (New York: Institute of Electrical and Electronic Engineers) pp 35–40

[77] Buturla E M, Cottrell P E, Grossman B M and Salsburg K A 1981 Finite-element analysis of semiconductor devices: the FIELDAY program *IBM J. Res. Dev.* **25** 218–31

[78] Buturla E M and Cottrell P E 1979 Two-dimensional static and transient simulation of mobile carrier transport in a semiconductor *Numerical Analysis of Semiconductor Devices, Procs. NASECODE I Conf.* ed J J H Miller (Dublin: Boole)

[79] Polak S J, Wachters A, Vaes H M J, de Beer A and den Hagen C 1979 A continuation method for the calculation of electrostatic potential in semiconductors *Numerical Analysis of Semiconductor Devices, Procs. NASECODE I Conf.* ed J J H Miller (Dublin: Boole)

[80] Polak S J, den Heijer C, Schilders W H A and Markowich P A 1987 Semiconductor device modeling from the numerical point of view *Int. J. Numer. Meth. Eng.* **24** 763–838

[81] Zlámal M 1986 Finite element solution of the fundamental equations of semiconductor devices. I *Math. Comput.* **46** 27–43

[82] Markowich P A and Zlámal M 1988 Inverse-average-type finite element discretizations of self-adjoint second-order elliptic problems *Math. Comput.* **51** 431–49

[83] Liu F and Miller J J H 1992 Inverse average type tetrahedral finite-element schemes for the stationary semiconductor device equations, *J. Comput. Appl. Math.* **44** 77–94

[84] Babuška I and Osborn J E 1983 Generalized finite element methods: their performance and their relation to mixed methods *SIAM J. Numer. Anal.* **20** 510–36

[85] Sonneveld P 1989 CGS, A fast Lanczos-type solver for nonsymmetric linear systems *SIAM J. Sci. Statist. Comput* **10** 36-52

[86] Van der Vorst H A 1992 Bi-CGSTAB: a fast and smoothly converging variant of Bi-CG for the solution of nonsymmetric linear systems *SIAM J. Sci. Statist. Comput.* **13** 631–44

[87] MacNeal R H 1953 An asymmetrical finite difference network *Q. Appl. Math.* **11** 295–310

[88] Scharfetter D and Gummel H K 1969 Large-signal analysis of a silicon read diode oscillator *IEEE Trans. Elec. Dec. Dec.* **ED-16** 64–77

[89] Chen Z 1990 Hybrid variable finite elements for semiconductor devices *Comput. Math. Applic.* **19** 65–73

[90] Brezzi F, Marini L D and Pietra P 1989 Mixed exponential fitting schemes for current continuity equations, *Proc. NASECODE-VI Conf.* ed J J H Miller (Dublin: Boole)

[91] Marini L D and Pietra P 1990 New mixed finite element schemes for current continuity equations *COMPEL* **9** 257–68

[92] Neudeck G W 1983 *Modular Series on Solid State Devices, Vol II: The PN Junction Diode* (Reading, MA: Addison-Wesley)

[93] Lambert A 1990 Aspects of isothermal and nonisothermal semiconductor device modelling *Master's Thesis* University of Strathclyde

[94] Barnes J J and Lomax R J 1977 Finite-element methods in semiconductor device simulation *IEEE Trans. Electron. Dev.* **24** 1082–9

[95] Fitzsimons C J, Miller J J H, Wang S and Wu C 1990 Hexahedral finite elements for the stationary semiconductor device equations *Comput. Methods Appl. Mech. Eng.* **84** 43–57

[96] van Welij J S 1986 Basis functions matching tangential components on element edges *Proc. SISDEP-2 Conf.* ed K Board and D R J Owen (Swansea: Pineridge) pp 371–83

[97] Zienkiewicz O C and Taylor R L 1991 *The Finite Element Method. Vol 1, Basic Formulation and Linear Problems* (London: McGraw-Hill)

[98] Huang M D 1985 The constant-flow patch test—a unique guideline for the evaluation of discretization schemes for the current continuity equations *IEEE Trans. Electron. Dev.* **32** 2139–64

[99] Montrone F 1995 A robust finite volume method for 3D device simulation *Proc. 8th ECMI Conf.* ed H Neunzert (Kaiserslautern)

[100] Sacco R, Gatti E and Gotusso L 1995 The patch test as a validation of a new finite element for the solution of convection-diffusion equations *Comput. Methods Appl. Mech. Eng.* **124** 113–24

[101] Sacco R and Saleri F 1997 Stabilized mixed finite volume methods for convection-diffusion problems *Report* 291/P, Politecnico di Milano, Dipartimento di matematica

[102] Schilders W H A and Polak S J 1990 Generalisations of the box method using the mixed finite element method *1990 VLSI Process/Device Modeling Workshop* pp 174–9

[103] Schilders W H A 1991 Generalisations of the box method with applications to the semiconductor problem *Proc. IMACS Conf. (Dublin)* pp 1688–9

[104] Sacco R and Saleri F 1997 Mixed finite volume methods for semiconductor device simulation *Numer. Meth.*

Part. Diff. Eq. **13** 215–36

[105] Aluru N R, Raefsky A, Pinsky P M, Law K H, Goossens R J G and Dutton R W 1993 A finite element formulation for the hydrodynamic semiconductor device equations *Comput. Methods Appl. Mech. Eng.* **107** 269–98

[106] Aluru N R, Law K H, Raefsky A, Pinsky P M and Dutton R W 1995 Numerical solution of two-carrier hydrodynamic semiconductor equations employing a stabilized finite element method *Comput. Methods Appl. Mech. Eng.* **125** 187–220

[107] Hughes T J R, Mallet M and Mizukami A 1986 A new finite element formulation for computational fluid dynamics: II. Beyond SUPG *Comput. Methods Appl. Mech. Eng.* **54** 341–55

[108] Johnson C 1986 Streamline diffusion methods for problems in fluids *Finite elements in fluids* vol VI, ed R H Gallagher *et al* (London: Wiley) pp 251–61

[109] Chen Z and Cockburn B 1994 Error estimates for a finite element method for the drift-diffusion semiconductor device equations *SIAM J. Numer. Anal.* **31** 1062–89

[110] Chen Z and Cockburn B 1995 Analysis of a finite element method for the drift-diffusion semiconductor device equations *Numer. Math.* **71** 1–28

[111] Chen Z, Cockburn B, Jerome J W and Shu C W 1995 Mixed-RDKG finite element methods for the 2-D hydrodynamic model for semiconductor device simulation, *VLSI Designs* **3** 1–14

[112] Chen Z 1996 A finite element method for the quantum hydrodynamic model for semiconductor devices *Comput. Math. Applic.* **31** 17–26

[113] Hecht F, Marrocco A, Caquot E and Filoche M 1991 Semiconductor device modelling for heterojunctions structures with mixed finite elements *COMPEL* **10** 425–38

[114] Martin R C and Ghoniem N M 1991 A hybrid finite-element/particle simulation method for the analysis of semiconductor transients and bipolar transport

[115] Farrell P A, Miller J J H, O'Riordan E and Shishkin G I 1998 Singularly perturbed differential equations with discontinuous source terms *Proc. Int. Workshop on Analytical and Computation Methods for Convection-Dominated and Singularly Perturbed Problems, Lozenetz, Bulgaria, August 27–31, 1998* in press