Massively parallel methods for semiconductor device modelling

R. K. Coomer * and I. G. Graham †
School of Mathematical Sciences
University of Bath, Bath BA2 7AY, U.K.

Abstract

In this paper we describe, analyse and implement a parallel iterative method for the solution of the steady-state drift diffusion equations governing the behaviour of a semiconductor device in two space dimensions. The unknowns in our model are the electrostatic potential and the electron and hole quasi-Fermi potentials. Our discretisation consists of a finite element method with mass lumping for the electrostatic potential equation and a hybrid finite element with local current conservation properties for the continuity equations. A version of Gummel's decoupling algorithm which only requires the solution of positive definite symmetric linear systems is used to solve the resulting nonlinear equations. We show that this method has an overall rate of convergence which only degrades logarithmically as the mesh is refined. Indeed the (inner) nonlinear solves of the electrostatic potential equation converge quadratically, with a mesh independent asymptotic constant. We also describe an implementation on a MasPar MP-1 data parallel machine, where the required linear systems are solved by the preconditioned congugate gradient method. Domain decomposition methods are used to parallelise the required matrix-vector multiplications and to build preconditioners for these very poorly conditioned systems. Our preconditioned linear solves also have a rate of convergence which degrades logarithmically as the grid is refined relative to subdomain size, and their performance is resilient to the severe layers which arise in the coefficients of the underlying elliptic operators. Extensive parallel experiments are given.

Mathematics Subject Classification (1991): 65F10, 65H10, 65N22, 65N30, 65N55

Keywords: Semiconductor Device, Drift-Diffusion Equations, Finite Element Methods, Gummel's Method, Preconditioned Conjugate Gradient Methods, Schur Complement, Domain Decomposition, Massively Parallel.

Running Head: Semiconductor Device Modelling

*Supported by an SERC Postgraduate Studentship.

†Please address correspondence to I.G. Graham. Email address: igg@maths.bath.ac.uk.

1 Introduction

In this paper we describe, analyse and implement a parallel iterative method for the solution of the steady-state drift-diffusion equations governing the electrical behaviour of semiconductor devices in two space dimensions. The drift-diffusion equations (three nonlinear PDEs in gradient form) may be posed using several different choices of variables. Here we model the density of electrons and holes using the "quasi-Fermi potentials". The corresponding continuity equations then contain nonlinear elliptic operators with coefficients depending exponentially on the unknowns. These suffer a very large numerical range which causes severe ill-conditioning in the matrices arising from numerical discretisations. The device model consists of two such continuity equations, coupled with a quasilinear elliptic equation defining the electrostatic potential. The equations are discretised by a variant of the finite element method with mass-lumping for the zeroth-order terms. For the continuity equations a special piecewise-constant average of the coefficients in the differential operator is used. This ensures that the scheme has certain weak current-conservation properties.

The resulting finite-dimensional nonlinear system is solved by a variant of the classical "Gummel " fixed-point iteration method. This is nothing more than a nonlinear block Gauss-Seidel iteration applied to the discretisations of the three coupled PDEs. Alternatively it can be viewed as a decoupled approximate block Newton method, where only (approximations to) the diagonal blocks of the Jacobian are inverted at each step. Each iterate of our Gummel algorithm requires first the solution of the (semilinear) electrostatic potential equation, with the quasi-Fermi potentials frozen. Using this updated electrostatic potential and the existing guesses for the quasi-Fermi potentials, a new electron quasi-Fermi potential is found by solving an appropriate linearised continuity equation. A new hole quasi-Fermi potential is found analogously and this process is repeated to convergence.

In this paper we provide answers to four questions.

1. How is the convergence of the (outer) Gummel iteration affected by refinement of the finite element mesh? It has been known since the work of Kerkhoven and Jerome ([18], [15]) that the fixed point map defined by (a related version of) Gummel's method is a contraction in some appropriately defined ball, provided the applied bias on the device is sufficiently small. In this paper we show that the Lipschitz constant of the fixed point map for our version of Gummel's method grows only logarithmically in $1/h$ (as the mesh diameter $h \to 0$), provided the meshes are refined in a regular manner.

2. How are the semilinear equations arising in the calculation of the electrostatic potential inside each Gummel iterate to be solved? Because these equations may become singularly perturbed, standard Newton convergence theory predicts a convergence ball whose radius may be so small as to have no practical meaning. In this paper we propose a certain quasi-Newton method which computes sequences of upper and lower solutions, and converges quadratically from any starting upper and lower solution pair. The required starting solutions are trivial to find. The rate of quadratic convergence is shown to be mesh-independent.

3. How can the linear systems which arise throughout the implemenation of the algorithm be effectively solved in parallel? The linear systems arising from the continuity equations present the greatest difficulty, since the exponential coefficients (which in a typical application vary at least between $10^{\pm 8}$) cause severe ill-conditioning. We describe the "massively parallel" solution of these systems by a domain decomposition preconditioned conjugate gradient method (CGM). We use a (variant of) the method proposed by Smith [30], [31] in which the CGM is applied to the (Schur complement) system arising when interior nodes of subdomains are locally eliminated. A preconditioner is constructed from local solves on (the interiors) of the edges of the subdomains together with the solution of an appropriate global coarse grid problem. This is an "additive Schwarz" method (see, e.g. [3], [11], [12]. [32], [30], [31] [23]). The preconditioned problem has conditioning independent of the jumps of the coefficients of the PDE across subdomain boundaries, and only growing logarithmically as the fine grid is refined relative to the coarse grid. Thus the degradation of performance of the inner (CGM) iterates is no worse than that of the outer (Gummel) iterates as the mesh is refined.

4. How should this algorithm be implemented on a massively parallel machine? Such machines have a large number of processors (typically some multiple of 1024), each with its own fast memory. These operate in SIMD lockstep, i.e. at any instant all processors are implementing the same instruction. Interchange of data between neighbouring processors is usually quicker than exchanges between random processors in the array. The slowest communication is between the array of processors and the outside world. Such machines naturally lead the user to work with a large number of subdomains each of which has a small number of nodes. We briefly describe implementation issues arising from such an architecture, and how we have dealt with them. Finally we give some numerical results for the semiconductor system on the MasPar MP-1.

Gummel's method in its continuous or discrete form has been well-studied and proved to converge for small enough applied bias ([15], [16], [18], [19], [20], [21]). In practice Gummel's method is still used far away from equilibrium, but usually in conjunction with continuation in some suitable parameter (e.g. the bias) – see, for example, [14]. Such methods compute solutions for a sequence of parameter values and restart with a smaller stepsize if divergence occurs. Clearly this can require a great many linear iteration steps leading to computation times which are much longer than those which may be reasonably required by a succesful interactive design system. One aim of this paper is to develop parallel methods which have the potential for speeding up this process. For another approach to accelerating Gummel's method see [29].

A body of work related to the present paper is [22], [8] and the references therein. There the semiconductor equations are discretised using a finite difference method on a tensor product grid and the resulting nonlinear equations are solved using a Gummel iteration with a Jacobi method for each of the (nonlinear) fractional steps. A convergence theory for the (inner) Jacobi iterations is given and an implementation on a massively parallel machine is described.

Mesh-independence results for *Newton's* method for general nonlinear systems have appeared at various points of the recent literature - see, e.g. [10] and the referenes therein. However, as is usual for Newton's method, these results depend on good estimates for the inverse of the Jacobian near the root. Such estimates are not needed for the analysis of Gummel's method which we give here. Indeed it is not clear that such estimates can be obtained for the complicated PDE problem containing several small parameters which we consider here.

The theoretical results concerning our version of Gummel's method are given in §3. The domain decomposition method is described in §4 and the computational results in §5. In §2 we set up the discretisation and describe the Gummel algorithm.

2 Discretization and outer (Gummel) iteration

The (scaled) drift-diffusion equations for device modelling may be written in the form

$$-\lambda^2\Delta\psi + \delta\{\exp(\psi - v) - \exp(w - \psi)\} - d = 0, \tag{2.1}$$

$$-\nabla.(\exp(\psi - v)\nabla v) - \sigma\rho_v r(\psi, v, w) = 0, \tag{2.2}$$

$$-\nabla.(\exp(w - \psi)\nabla w) + \sigma\rho_w r(\psi, v, w) = 0. \tag{2.3}$$

Here ψ is the electrostatic potential and v and w are the electron and hole quasi-Fermi potentials respectively. Using the scaling in [24, §(3.1)], the parameters λ and δ are given by the formulae

$$\lambda = l^{-1}\sqrt{(\epsilon U_T)/(q\tilde{D})}, \qquad \delta = n_i/\tilde{D},$$

where ϵ is the permittivity, U_T is the thermal voltage, q is the elementary charge, n_i is the intrinsic concentration, l is a characteristic length and $\tilde{D}$ is the maximum absolute value of the doping profile D. Then $d = D/\tilde{D}$, is the scaled doping profile. In the continuity equations (2.2), (2.3), Einstein's relations ([24, p.89]) have been assumed to hold. Then the parameters ρ_v and ρ_w are the reciprocals of the electron and hole mobilities (which are assumed constant) and $\sigma = l^2/(\tilde{D}U_T)$ is a scaling factor. The function r models the recombination and generation of holes and electrons in the device. (see e.g. [24]). The most commonly used model is that proposed by Shockley, Read and Hall:

$$r(\psi, v, w) = \frac{\exp(w - v) - 1}{\exp(w - \psi) + \exp(\psi - v) + 2} \quad . \tag{2.4}$$

We will consider this system on a polygonal domain $\Omega \subset \mathbf{R}^2$ with boundary $\partial\Omega$, subject to mixed boundary conditions given as follows. At the "Ohmic contacts" $\partial\Omega_D = \cup_i \partial\Omega_{D_i}$ (where the $\partial\Omega_{D_i}$ are a finite number of closed nonempty subsets of $\partial\Omega$), we impose piecewise constant Dirichlet conditions on v and w:

$$v|_{\partial\Omega_{Di}} = w|_{\partial\Omega_{Di}} = \alpha_i = \text{ constant}, \quad \text{ for each } i. \tag{2.5}$$

This corresponds to the imposition of an applied voltage (or *bias*) $\alpha_i U_T$ at each of the contacts $\partial\Omega_{D_i}$. Dirichlet conditions for ψ at the contacts are then obtained by requiring that the space charge there should be zero., i.e.

$$\delta\{\exp(\psi - v) - \exp(w - \psi)\} - d = 0, \quad \text{on} \ \ \partial\Omega_D, \tag{2.6}$$

or, equivalently,

$$\psi|_{\partial\Omega_{Di}} = \alpha_i + \sinh^{-1}\left(\frac{d|_{\partial\Omega_{D_i}}}{2\delta}\right) \quad \text{for each } i, \tag{2.7}$$

On $\partial\Omega_N := \partial\Omega\backslash\partial\Omega_D$ the device is assumed to be insulated, i.e.

$$\frac{\partial\psi}{\partial n} = \frac{\partial u}{\partial n} = \frac{\partial v}{\partial n} = 0 \quad \text{on} \ \ \partial\Omega_N. \tag{2.8}$$

We can now see heuristically why the solutions of this system may exhibit layer behaviour. Consider for example the zero-current case ($\alpha_i = 0$) for a simple $p - n$ diode with equal doping levels. This can be modelled by choosing d to have the values ± 1 in the n, p regions respectively. Then, from (2.7),

$$\psi = \pm\beta, \qquad \text{where} \ \ \beta = \sinh^{-1}(1/(2\delta)) \tag{2.9}$$

at contacts in the n, p regions. When the parameter λ in (2.1) is small, (2.1) is singularly perturbed. Then ψ changes rapidly in a small region (layer) around the interface in d and remains essentially constant in most of the domain. When the α_i become different from zero, layers also arise in v, w. Singular perturbation theory has been used to explain these phenomena in detail ([24]).

We triangulate Ω and discretise (2.1) – (2.3) using linear finite elements. To facilitate the domain decomposition methods in §4, our triangulation is obtained by first dividing Ω into open convex quadrilateral subdomains or "substructures" $\Omega^{(i)}$ such that $\overline{\Omega} = \cup_i \overline{\Omega^{(i)}}$ and $\Omega^{(i)} \cap \Omega^{(j)} = \emptyset$ when $i \neq j$. Each substructure vertex is assumed to belong to no more than four substructures, and the intersection of the closure of any two substructures is either empty or consists entirely of a common edge and the two associated vertices. We divide each substructure into two triangles to obtain a "coarse grid" with maximum mesh diameter H. We then further refine the coarse grid to form a "fine grid" with maximum mesh diameter h. For theoretical purposes, we assume throughout that both coarse grid and fine grid are conforming, and that as the mesh is refined both grids are *regular* and satisfy an *inverse assumption* [6, p.124, p.140]. Such meshes are usually referred to as *quasi-uniform*. We also assume that the fine grid is of *weakly acute type*, (i.e. each angle of its triangles is no greater then $\pi/2$). Finally we assume that the points in $\partial\Omega_D \cap \overline{\partial\Omega_N}$ (i.e. the *collision points* between Dirichlet and Neumann boundary conditions) are vertices of substructures and hence are nodes of both the coarse and fine grids.

We introduce the usual spaces $L_2(\Omega)$ with inner product $(\cdot,\cdot)$ and norm $\|\cdot\|_{L_2(\Omega)}$, and $H^1(\Omega)$ with semi-norm $|\cdot|_{H^1(\Omega)}$ and norm $\|\cdot\|_{H^1(\Omega)}$. If $\boldsymbol{f}, \boldsymbol{g} \in (L^2(\Omega))^2$ then $(\boldsymbol{f}, \boldsymbol{g}) := \int_\Omega \boldsymbol{f}\cdot\boldsymbol{g}$, with $\cdot$ denoting the usual dot product on $\mathbf{R}^2$.

Let S_h denote the space of all *piecewise linear* functions on Ω subordinate to the fine grid. The usual basis functions for S_h are $\{\phi_p\}$ where ϕ_p is 1 at node p and zero elsewhere.

Let Σ_h denote the space of all *piecewise constant* functions subordinate to the fine grid. The usual finite element method for (2.1) – (2.3) would consist of seeking $\Psi, V, W \in S_h$ satisfying (2.5), (2.7) such that the equations

$$\lambda^2(\nabla\Psi,\ \nabla\phi_p) + (\delta\{\exp(\Psi - V) - \exp(W - \Psi)\} - d,\ \phi_p) = 0, \tag{2.10}$$

$$(\exp(\Psi - V)\nabla V,\ \nabla\phi_p) - (\sigma\rho_v r(\Psi, V, W), \phi_p) = 0, \tag{2.11}$$

$$(\exp(W - \Psi)\nabla W, \nabla\phi_p) + (\sigma\rho_w r(\Psi, V, W), \phi_p) = 0, \tag{2.12}$$

are satisfied for all nodes $p \notin \partial\Omega_D$.

We shall in fact consider a slightly modified scheme. Firstly we shall use "mass lumping" for the zeroth-order term in (2.10). This approximation is useful in practice, since it yields a simple diagonal approximation to a complicated nonlinear term. Moreover, iterative schemes for solving the approximation of (2.10) (with v and w fixed) then have very useful montonicity properties which are not present if the same methods are applied to (2.10) itself (see §3). Mass lumping may be achieved by introducing the nodal quadrature rule:

$$\int_\Omega f \ \approx\ \sum_T \frac{1}{3}\mathcal{A}(T)\sum_{p_T} f(p_T), \tag{2.13}$$

where the outer sum is over all triangles T of the fine grid, the inner sum is over the three nodes p_T of T and $\mathcal{A}(T)$ denotes the area of T. This rule is exact when $f \in S_h$ and may be used to approximate the standard linear finite element method for sufficiently smooth linear elliptic problems without loss of order of accuracy (in the energy norm) [6, Theorem 4.1.6]. For continuous f it is useful to write (2.13) as

$$\int_\Omega f \approx \sum_p w_p f(p), \tag{2.14}$$

where the sum is over all nodes p of the fine grid and each weight w_p is simply one third the sum of the areas of all the triangles which meet at node p. This induces the discrete bilinear form

$$\langle f, g\rangle = \sum_T \frac{1}{3}\mathcal{A}(T)\sum_{p_T}(fg)(p_T) = \sum_p w_p f(p)g(p), \tag{2.15}$$

which we use to approximate the second term in (2.10)

The other approximations we shall make are in (2.11), (2.12). If $X \in S_h$ we define $\overline{X} \in \Sigma_h$ by

$$\exp(\overline{X}|_T) \ = \left\{\frac{1}{\mathcal{A}(T)}\int_T \exp(-X)\right\}^{-1}, \text{ for each triangle } T.$$

Then $\exp(\overline{X})$ is the *harmonic average* of $\exp(X)$ (see Appendix) and can be calculated analytically. We replace the exponential coefficients in the second-order terms of (2.11), (2.12) by their harmonic averages. Moreover we replace $r(\Psi, V, W)$ by its standard average $r_h(\Psi, V, W)$ computed as in (A.13). The zeroth order terms may then be computed exactly by the quadrature rule (2.13). Following the discussion in the Appendix, these approximations (which do not degrade the rate of convergence in the energy norm) ensure that the resulting (piecewise constant) approximations to the electron and hole currents ($\exp(\psi - v)\nabla v$ and $\exp(w - \psi)\nabla w$) have weak conservation properties. Applying these approximations in (2.12)-(2.15) yields a nonlinear system to be solved for Ψ, V, W.

One variant of *Gummel's method* ([13]) for the solution of this system consists of iterating the map

$$\mathcal{G} : (V, W) \longmapsto (\tilde{V}, \tilde{W}), \tag{2.16}$$

defined as follows.

Step 1. (Fractional Step) Find $\tilde{\Psi} \in S_h$ satisfying (2.7) and such that

$$\lambda^2(\nabla\tilde{\Psi}, \nabla\phi_p) + \langle\delta\{\exp(\tilde{\Psi} - V) - \exp(W - \tilde{\Psi})\} - d,\ \phi_p\rangle = 0, \quad p \notin \partial\Omega_D. \tag{2.17}$$

Step 2. Find $\tilde{V}$ satisfying (2.5) and such that

$$(\exp(\overline{\tilde{\Psi} - V})\nabla\tilde{V}, \nabla\phi_p) - \langle\sigma\rho_v r_h(\tilde{\Psi}, V, W), \phi_p\rangle = 0, \quad p \notin \partial\Omega_D. \tag{2.18}$$

Step 3. Find $\tilde{W}$ satisfying (2.5) and such that

$$(\exp(\overline{W - \tilde{\Psi}})\nabla\tilde{W}, \nabla\phi_p) + \langle\sigma\rho_w r_h(\tilde{\Psi}, \tilde{V}, W), \phi_p\rangle = 0, \quad p \notin \partial\Omega_D. \tag{2.19}$$

In the next section we examine the convergence of this iteration and also the convergence of an (inner) iterative method for solving the semilinear problem (2.17).

3 Convergence of Gummel Iteration

In this section we will discuss implementation and convergence of the map $\mathcal{G}$ defined in (2.16) – (2.19) Following [15], [18] we do this using the contraction mapping theorem in the set

$$B = \{(V, W) \in S_h \times S_h\ :\ \underline{\alpha} \le V, W \le \overline{\alpha}\},$$

where

$$\overline{\alpha} = \max_i\ \alpha_i, \qquad \underline{\alpha} = \min_i\ \alpha_i, \tag{3.1}$$

and α_i are the applied voltages appearing in (2.5). We equip B with the norm

$$\|(V, W)\|_B = \{\|V\|_{H^1(\Omega)} + \|W\|_{H^1(\Omega)}\},$$

with respect to which B is a complete metric space.

In Theorem 3.6 below we will demonstrate the Lipschitz continuity of $\mathcal{G}$. We shall show that its Lipschitz constant is less than 1 provided $\underline{\alpha}$ and $\overline{\alpha}$ are sufficiently close to zero when h is fixed. (A result analogous to this is also obtained in [18], but with a different choice of variables.) Here we shall also show that the Lipschitz constant of $\mathcal{G}$ grows only logarithmically in h as the mesh is refined.

We also obtain an iterative method for solving (2.17) which converges monotonically, quadratically and mesh independently from specific starting guesses which we shall define. The monotonicity of this method also gives us *a priori* mesh-independent bounds on $\tilde{\Psi}$. These results about (2.17) are the subject of the next subsection, while the properties of $\mathcal{G}$ are proved in the subsection following.

3.1 Calculation and properties of the electrostatic potential

In this subsection we assume that

$$(V, W) \in B \tag{3.2}$$

is fixed and we consider the solution of (2.17) for $\tilde{\Psi}$. Instead of using a standard Newton method, we devise a quasi-Newton scheme which is guaranteed to converge monotonically from initial upper and lower solutions (which are easy to construct). (More details and background on monotone Newton schemes will be given in [7].) To describe this method, we introduce the following notation.

If $\boldsymbol{z}$ is any vector, then $\mathrm{diag}\{\boldsymbol{z}\}$ denotes the diagonal matrix with $\boldsymbol{z}$ on its main diagonal. If f is any univariate function, then $f(\boldsymbol{z})$ denotes the vector with $f(\boldsymbol{z})_p = f(z_p)$. We write $\boldsymbol{z} \geq \boldsymbol{0}$ to mean $z_p \geq 0$ for all p. If $\boldsymbol{z}$ and $\boldsymbol{y}$ have the same number of components, we write $\boldsymbol{z} \geq \boldsymbol{y}$ if $\boldsymbol{z} - \boldsymbol{y} \geq \boldsymbol{0}$, in which case we denote the set $\{\boldsymbol{\Phi} : \boldsymbol{y} \leq \boldsymbol{\Phi} \leq \boldsymbol{z}\}$ by $[\boldsymbol{y}, \boldsymbol{z}]$.

Now write $\boldsymbol{\Psi}, \boldsymbol{V}$ and $\boldsymbol{W}$ for the vectors of values of Ψ, V and W at the nodes on $\overline{\Omega}\backslash\partial\Omega_D$ and write $\boldsymbol{\Psi}_D, \boldsymbol{V}_D$ and $\boldsymbol{W}_D$ for the vectors of values of Ψ, V and W at the nodes on $\partial\Omega_D$ given by (2.5),(2.7). Let $\boldsymbol{d}$ be the vector with $d_p = d(p)$ for each node p, and (considering $\boldsymbol{V}$ and $\boldsymbol{W}$ to be fixed), define

$$\boldsymbol{g}(\boldsymbol{\Psi}) = \mathrm{diag}\{\boldsymbol{w}\}\,[\delta\,\{\exp(\boldsymbol{\Psi} - \boldsymbol{V}) \,-\, \exp(\boldsymbol{W} - \boldsymbol{\Psi})\} - \boldsymbol{d}],$$

where $\boldsymbol{w}$ is the vector of weights of the quadrature rule (2.14).

Then, using (2.15), (2.17) may be written as the problem of finding a solution $\tilde{\boldsymbol{\Psi}}$ to $\boldsymbol{F}(\boldsymbol{\Psi}) = \boldsymbol{0}$, where

$$\boldsymbol{F}(\boldsymbol{\Psi}) = \lambda^2[K\boldsymbol{\Psi} + K_D\boldsymbol{\Psi}_D] + \boldsymbol{g}(\boldsymbol{\Psi}). \tag{3.3}$$

Here K represents the coupling between nodes of $\overline{\Omega}\backslash\partial\Omega_D$ induced by the finite element approximation to the Laplacian, whereas K_D represents the analogous coupling between nodes on $\overline{\Omega}\backslash\partial\Omega_D$ and nodes on $\partial\Omega_D$. Obviously K is just the stiffness matrix arising from the finite element approximation of the Laplacian with homogeneous Dirichlet condition on $\partial\Omega_D$ and (homogeneous) natural boundary condition on $\partial\Omega_N$. Also, observe that if $\boldsymbol{1}$ denotes the unit vector at nodes on $\overline{\Omega}\backslash\partial\Omega_D$ and $\boldsymbol{1}_D$ denotes the unit vector at nodes on $\partial\Omega_D$, then we have the useful formula

$$K\boldsymbol{1} + K_D\boldsymbol{1}_D = \boldsymbol{0}. \tag{3.4}$$

The system (3.3) has Jacobian

$$\begin{aligned} J(\boldsymbol{\Psi}) &= \lambda^2 K \;+\; \delta\mathrm{diag}\{\boldsymbol{w}\}\mathrm{diag}\{\exp(\boldsymbol{\Psi} - \boldsymbol{V}) + \exp(\boldsymbol{W} - \boldsymbol{\Psi})\} \\ &=: \lambda^2 K \;+\; \boldsymbol{g}_\Psi(\boldsymbol{\Psi}), \end{aligned} \tag{3.5}$$

where $\boldsymbol{g}_\Psi$ is the (diagonal) Jacobian matrix of $\boldsymbol{g}$.

Our quasi-Newton method for solving (3.3) may then be written

- Assume we have *lower* and *upper solutions* $\boldsymbol{\Lambda}^0, \boldsymbol{\Omega}^0$ which satisfy

$$\boldsymbol{\Lambda}^0 \leq \boldsymbol{\Omega}^0 \quad \text{and} \quad \boldsymbol{F}(\boldsymbol{\Lambda}^0) \leq \boldsymbol{0} \leq \boldsymbol{F}(\boldsymbol{\Omega}^0). \tag{3.6}$$

- Then, for $k \geq 0$, set

$$\begin{aligned} \boldsymbol{\Lambda}^{k+1} &= \boldsymbol{\Lambda}^k - (J^k)^{-1}\boldsymbol{F}(\boldsymbol{\Lambda}^k), && (3.7)\\ \boldsymbol{\Omega}^{k+1} &= \boldsymbol{\Omega}^k - (J^k)^{-1}\boldsymbol{F}(\boldsymbol{\Omega}^k), && (3.8) \end{aligned}$$

  where

$$J^k := \max\left\{J(\boldsymbol{\Lambda}^k), J(\boldsymbol{\Omega}^k)\right\}. \qquad (3.9)$$

(The maximum in (3.9) is taken *elementwise.*) Before we analyse this method we make two remarks.

Remark 3.1 Since the fine grid is assumed to be of weakly acute type, the matrix K is irreducible, positive definite and symmetric, with non-positive off-diagonal elements. For *any* vector $\boldsymbol{\Psi}$, $J(\boldsymbol{\Psi})$ also has this property. Hence by [33, p.85, Corollary 3], both K and $J(\boldsymbol{\Psi})$ are non-singular and

$$K^{-1} > 0, \qquad J(\boldsymbol{\Psi})^{-1} > 0 \qquad (3.10)$$

(i.e. all elements of both these inverses are strictly positive). Hence our quasi-Newton method is always well-defined. $\square$

Remark 3.2 Lower and upper solutions are easy to construct. For example, set

$$\Lambda^0 = \underline{\alpha} + \min_{\boldsymbol{x}\in\overline{\Omega}} \sinh^{-1}\left(\frac{d(\boldsymbol{x})}{2\ \delta}\right)$$

and define

$$\boldsymbol{\Lambda}^0 = \Lambda^0\mathbf{1}, \qquad \boldsymbol{\Lambda}_D^0 = \Lambda^0\mathbf{1}_D.$$

Then (3.3), (3.4) yield

$$\boldsymbol{F}(\boldsymbol{\Lambda}^0) = \lambda^2 K_D(\boldsymbol{\Psi}_D - \boldsymbol{\Lambda}_D^0) + \boldsymbol{g}(\boldsymbol{\Lambda}^0).$$

The first term is non-positive since $\boldsymbol{\Psi}_D \geq \boldsymbol{\Lambda}_D^0$ and since K_D contains only non-positive elements. Also since $(V, W) \in B$, we have

$$\boldsymbol{g}(\boldsymbol{\Lambda}^0) \leq \mathrm{diag}(\boldsymbol{w})[2\delta\sinh((\Lambda^0 - \underline{\alpha})\mathbf{1}) - \boldsymbol{d}] \leq \mathbf{0}.$$

So Λ^0 is a lower solution. Similarly $\boldsymbol{\Omega}^0 := \Omega^0\mathbf{1}$ is an upper solution when

$$\Omega^0 = \overline{\alpha} + \max_{\boldsymbol{x}\in\overline{\Omega}} \sinh^{-1}\left(\frac{d(\boldsymbol{x})}{2\ \delta}\right) \qquad \square.$$

The analysis of our quasi-Newton method is based on a generalisation of the ideas in [26], where a monotone Newton convergence theorem is given in §13.3.4. However that theorem requires that the function $\boldsymbol{F}$ should be *order convex* in the set in which the solution $\tilde{\Psi}$ is sought. Since in general $\tilde{\boldsymbol{\Psi}}$ may contain both negative and positive entries, it is clear that $\boldsymbol{F}$ given by (3.3) is not order convex. (Consider for example the special case $\underline{\alpha} = \overline{\alpha} = 0$. Then $\boldsymbol{g}(\boldsymbol{\Psi}) = \mathrm{diag}(\boldsymbol{w})[2\delta\sinh(\boldsymbol{\Psi}) - \boldsymbol{d}]$.) Here we prove a more general convergence analysis which is applicable to (3.6)-(3.9). It has recently been pointed out to us (F. A. Potra, private communication) that the basic convergence properties of Λ^k, Ω^k proved below are a consequence of the much more general theory in [28]. However neither the *a priori* bounds of Theorem 3.3 or the mesh-independence of the convergence proved in Corollary 3.4 are in [28] and so we take the opprotunity here to give a full elementary proof of the convergence as well.

LEMMA 3.1 *Let $\left\{\boldsymbol{\Lambda}^k\right\}_{k=1}^{\infty}, \left\{\boldsymbol{\Omega}^k\right\}_{k=1}^{\infty}$ be defined by (3.6)-(3.9). Suppose that $\boldsymbol{\Lambda}^0 \le \boldsymbol{\Lambda}^k \le \boldsymbol{\Omega}^k \le \boldsymbol{\Omega}^0$ for some k. Then, for any $\boldsymbol{\Phi} \in [\boldsymbol{\Lambda}^k, \boldsymbol{\Omega}^k]$ we have*

$$\boldsymbol{F}(\boldsymbol{\Omega}^k) - \boldsymbol{F}(\boldsymbol{\Phi}) - J^k(\boldsymbol{\Omega}^k - \boldsymbol{\Phi}) \le \mathbf{0} \le \boldsymbol{F}(\boldsymbol{\Lambda}^k) - \boldsymbol{F}(\boldsymbol{\Phi}) - J^k(\boldsymbol{\Lambda}^k - \boldsymbol{\Phi}) \tag{3.11}$$

and

$$\boldsymbol{\Lambda}^{k+1} \le \boldsymbol{\Phi} - (J^k)^{-1}\boldsymbol{F}(\boldsymbol{\Phi}) \le \boldsymbol{\Omega}^{k+1}. \tag{3.12}$$

Proof By (3.5) and (3.9), we can write

$$J^k = \lambda^2 K + \boldsymbol{g}_\Psi^k, \quad \text{where} \quad \boldsymbol{g}_\Psi^k = \max\left\{\boldsymbol{g}_\Psi(\boldsymbol{\Lambda}^k), \boldsymbol{g}_\Psi(\boldsymbol{\Omega}^k)\right\}. \tag{3.13}$$

Now let $\boldsymbol{\Phi} \in [\boldsymbol{\Lambda}^k, \boldsymbol{\Omega}^k]$. Then, using the mean value theorem,

$$\begin{aligned} \boldsymbol{F}(\boldsymbol{\Omega}^k) - \boldsymbol{F}(\boldsymbol{\Phi}) - J^k(\boldsymbol{\Omega}^k - \boldsymbol{\Phi}) &= \boldsymbol{g}(\boldsymbol{\Omega}^k) - \boldsymbol{g}(\boldsymbol{\Phi}) - \boldsymbol{g}_\Psi^k(\boldsymbol{\Omega}^k - \boldsymbol{\Phi}) \\ &= (\boldsymbol{g}_\Psi(\boldsymbol{\eta}) - \boldsymbol{g}_\Psi^k)(\boldsymbol{\Omega}^k - \boldsymbol{\Phi}), \end{aligned}$$

for some $\boldsymbol{\eta} \in [\boldsymbol{\Lambda}^k, \boldsymbol{\Omega}^k]$. Now, since $\boldsymbol{g}_\Psi(\boldsymbol{\eta})$ is diagonal and since $\boldsymbol{g}_\Psi(\boldsymbol{\eta})_{ii}$ is a convex function of η_i, we have

$$\boldsymbol{g}_\Psi(\boldsymbol{\eta})_{ii} \le \max\left\{\boldsymbol{g}_\Psi^k(\boldsymbol{\Lambda}^k)_{ii}, \boldsymbol{g}_\Psi(\boldsymbol{\Omega}^k)_{ii}\right\} \equiv (\boldsymbol{g}_\Psi^k)_{ii}.$$

Therefore the left-hand inequality of (3.11) follows. The right-hand inequality is proved analogously.

To obtain (3.12), use (3.7), (3.10) and (3.11) to write (for any $\boldsymbol{\Phi} \in [\boldsymbol{\Lambda}^k, \boldsymbol{\Phi}^k]$),

$$\begin{aligned} \boldsymbol{\Phi} - (J^k)^{-1}\boldsymbol{F}(\boldsymbol{\Phi}) &= \boldsymbol{\Lambda}^{k+1} - (\boldsymbol{\Lambda}^k - \boldsymbol{\Phi}) + (J^k)^{-1}(\boldsymbol{F}(\boldsymbol{\Lambda}^k) - \boldsymbol{F}(\boldsymbol{\Phi})) \\ &\ge \boldsymbol{\Lambda}^{k+1} - (\boldsymbol{\Lambda}^k - \boldsymbol{\Phi}) + (J^k)^{-1}J^k(\boldsymbol{\Lambda}^k - \boldsymbol{\Phi}) \\ &= \boldsymbol{\Lambda}^{k+1}, \end{aligned}$$

which is the left-hand inequality of (3.12). Again, the right-hand inequality is analogous. □

LEMMA 3.2 *Let $\left\{\boldsymbol{\Lambda}^k\right\}_{k=1}^{\infty}, \left\{\boldsymbol{\Omega}^k\right\}_{k=1}^{\infty}$ be defined by (3.6)-(3.9). Then for all $k \ge 0$*

$$\boldsymbol{\Lambda}^0 \le \boldsymbol{\Lambda}^k \le \boldsymbol{\Omega}^k \le \boldsymbol{\Omega}^0 \tag{3.14}$$

$$\boldsymbol{F}(\boldsymbol{\Lambda}^k) \le \mathbf{0} \le \boldsymbol{F}(\boldsymbol{\Omega}^k) \tag{3.15}$$

$$\boldsymbol{\Omega}^{k+1} \le \boldsymbol{\Omega}^k \tag{3.16}$$

$$\boldsymbol{\Lambda}^k \le \boldsymbol{\Lambda}^{k+1} \tag{3.17}$$

Proof First note that if (3.14) and (3.15) hold then (by (3.7)-(3.10)), so do (3.16) and (3.17). Also note that, by Remark 3.2, $\mathbf{\Lambda}^0$ and $\mathbf{\Omega}^0$ satisfy (3.14) and (3.15). Hence the result is true for $k=0$. Now suppose that the result holds for some $k \geq 0$. We need to show that (3.14) and (3.15) hold for $k+1$. Putting $\mathbf{\Phi} = \mathbf{\Lambda}^k$ in (3.12) yields

$$\mathbf{\Lambda}^{k+1} = \mathbf{\Lambda}^k - (J^k)^{-1} \boldsymbol{F}(\mathbf{\Lambda}^k) \leq \mathbf{\Omega}^{k+1}.$$

Hence we have $\mathbf{\Lambda}^0 \leq \mathbf{\Lambda}^k \leq \mathbf{\Lambda}^{k+1} \leq \mathbf{\Omega}^{k+1} \leq \mathbf{\Omega}^k \leq \mathbf{\Omega}^0$. On the other hand, putting $\mathbf{\Phi} = \mathbf{\Lambda}^{k+1}$ in (3.11) and using (3.7) we obtain

$$\mathbf{0} \leq \boldsymbol{F}(\mathbf{\Lambda}^k) - \boldsymbol{F}(\mathbf{\Lambda}^{k+1}) - J^k(\mathbf{\Lambda}^k - \mathbf{\Lambda}^{k+1}) = -\boldsymbol{F}(\mathbf{\Lambda}^{k+1})$$

Thus $\boldsymbol{F}(\mathbf{\Lambda}^{k+1}) \leq \mathbf{0}$. Similarly we can prove that $\boldsymbol{F}(\mathbf{\Omega}^{k+1}) \geq \mathbf{0}$, and the result follows by induction. $\square$

We can now prove the convergence of (3.6)-(3.9).

THEOREM 3.3 *The sequences* $\left\{\mathbf{\Lambda}^k\right\}_{k=1}^{\infty}, \left\{\mathbf{\Omega}^k\right\}_{k=1}^{\infty}$ *defined by (3.6)-(3.9) converge to the same limit* $\tilde{\mathbf{\Psi}}$*, which is the unique solution of* $\boldsymbol{F}(\mathbf{\Psi}) = \mathbf{0}$*. Furthermore,*

$$\underline{\alpha} + \min_{\boldsymbol{x} \in \overline{\Omega}} \sinh^{-1}\left(\frac{d(\boldsymbol{x})}{2\delta}\right) \leq \tilde{\mathbf{\Psi}} \leq \overline{\alpha} + \max_{\boldsymbol{x} \in \overline{\Omega}} \sinh^{-1}\left(\frac{d(\boldsymbol{x})}{2\delta}\right).$$

Proof By Lemma 3.2, $\left\{\mathbf{\Omega}^k\right\}_{k=1}^{\infty} \subset [\mathbf{\Lambda}^0, \mathbf{\Omega}^0]$ is a bounded monotonically decreasing sequence. Hence

$$\mathbf{\Omega}^k \searrow \tilde{\mathbf{\Omega}} \in [\mathbf{\Lambda}^0, \mathbf{\Omega}^0].$$

Similarly it follows that

$$\mathbf{\Lambda}^k \nearrow \tilde{\mathbf{\Lambda}} \in [\mathbf{\Lambda}^0, \mathbf{\Omega}^0].$$

Lemma 3.2 implies also that $\tilde{\mathbf{\Lambda}} \leq \tilde{\mathbf{\Omega}}$.

We now show that $\boldsymbol{F}(\tilde{\mathbf{\Lambda}}) = \mathbf{0} = \boldsymbol{F}(\tilde{\mathbf{\Omega}})$. To do this use (3.13) and the properties of $\mathbf{\Lambda}^k$ and $\mathbf{\Omega}^k$ proved in Lemma 3.2 to obtain the trivial bound:

$$||J^k||_\infty \leq \lambda^2 ||K||_\infty + ||\boldsymbol{g}_\Psi^k||_\infty \leq \lambda^2 ||K||_\infty + ||\boldsymbol{g}_\Psi^0||_\infty =: M,$$

with M is independent of k. Using this and (3.8), we then have

$$||\boldsymbol{F}(\mathbf{\Omega}^k)||_\infty = ||J^k(\mathbf{\Omega}^k - \mathbf{\Omega}^{k+1})||_\infty \leq M||\mathbf{\Omega}^k - \mathbf{\Omega}^{k+1}||_\infty \to 0, \text{ as } k \to \infty.$$

Hence, by continuity, $\boldsymbol{F}(\tilde{\mathbf{\Omega}}) = \mathbf{0}$. Similarly $\boldsymbol{F}(\tilde{\mathbf{\Lambda}}) = \mathbf{0}$.

Now observe that if $\tilde{\mathbf{\Psi}}$ is *any* solution of $\boldsymbol{F}(\mathbf{\Psi}) = \mathbf{0}$ then, again using the mean value theorem,

$$\mathbf{0} = \boldsymbol{F}(\tilde{\mathbf{\Omega}}) - \boldsymbol{F}(\tilde{\mathbf{\Psi}}) = \lambda^2 K(\tilde{\mathbf{\Omega}} - \tilde{\mathbf{\Psi}}) + \boldsymbol{g}(\tilde{\mathbf{\Omega}}) - \boldsymbol{g}(\tilde{\mathbf{\Psi}}) = J(\boldsymbol{\eta})(\tilde{\mathbf{\Omega}} - \tilde{\mathbf{\Psi}}),$$

for some vector $\boldsymbol{\eta}$. By (3.10), $J(\boldsymbol{\eta})$ is invertible and therefore $\tilde{\mathbf{\Omega}} = \tilde{\mathbf{\Psi}}$. Hence $\tilde{\mathbf{\Omega}} = \tilde{\mathbf{\Lambda}}$ is the unique solution of $\boldsymbol{F}(\mathbf{\Psi}) = \mathbf{0}$, which we now denote $\tilde{\mathbf{\Psi}}$.

To obtain the required bounds on $\tilde{\mathbf{\Psi}}$ simply use the fact that $\tilde{\mathbf{\Psi}}$ lies between any pair of lower and upper solutions and recall Remark 3.2. $\square$

COROLLARY 3.4 *Suppose method (3.6)-(3.9) is applied to starting vectors* $\boldsymbol{\Lambda}^0, \boldsymbol{\Omega}^0$, *which are bounded independently of* h *(such as those in Remark 3.2). Then*

$$||\boldsymbol{\Omega}^{k+1} - \boldsymbol{\Lambda}^{k+1}||_2 \leq C||\boldsymbol{\Omega}^k - \boldsymbol{\Lambda}^k||_2^2, \quad k \geq 0,$$

with a constant C *depending on* λ *and* δ *but independent of* h *and* k.

Proof By (3.9), (3.10), (3.13) and the fact that $\boldsymbol{g}_\Psi$ is a non-negative matrix, we have

$$(\lambda^2 K)^{-1} J^k = I + (\lambda^2 K)^{-1} \boldsymbol{g}_\Psi^k \geq I.$$

Then, using (3.9) and (3.10) again it follows that

$$(\lambda^2 K)^{-1} \geq (J^k)^{-1} > 0. \tag{3.18}$$

Now, for any symmetric matrix A, let $\mu\mathrm{max}(A)$ denote its largest eigenvalue. Then since $(J^k)^{-1}$ is positive definite and symmetric we have, by the Rayleigh Quotient Theorem,

$$\mu\mathrm{max}((J^k)^{-1}) = \frac{\boldsymbol{x}^T((J^k)^{-1})\boldsymbol{x}}{\boldsymbol{x}^T\boldsymbol{x}}$$

for some $\boldsymbol{x}$. Since $(J^k)^{-1}$ is also positive, the Perron-Frobenius Theorem implies that we can choose $\boldsymbol{x}$ to be non-negative. Hence, using (3.18) it follows that

$$\mu\mathrm{max}((J^k)^{-1}) = \frac{\boldsymbol{x}^T((J^k)^{-1})\boldsymbol{x}}{\boldsymbol{x}^T\boldsymbol{x}} \leq \frac{\boldsymbol{x}^T(\lambda^2 K)^{-1}\boldsymbol{x}}{\boldsymbol{x}^T\boldsymbol{x}} \leq \mu\mathrm{max}((\lambda^2 K)^{-1}).$$

Hence

$$||(J^k)^{-1}||_2 \leq ||(\lambda^2 K)^{-1}||_2 = \lambda^{-2}||K^{-1}||_2 \leq C\lambda^{-2}h^{-2},$$

with C independent of h, where the last inequality is found, for example, in [17, §7.7]. Hence, using (3.7), (3.8) and the mean value theorem again,

$$\begin{aligned}
||\boldsymbol{\Omega}^{k+1} - \boldsymbol{\Lambda}^{k+1}||_2 &= ||\boldsymbol{\Omega}^k - \boldsymbol{\Lambda}^k - (J^k)^{-1}(\boldsymbol{F}(\boldsymbol{\Omega}^k) - \boldsymbol{F}(\boldsymbol{\Lambda}^k))||_2 \\
&= ||(J^k)^{-1}[J^k(\boldsymbol{\Omega}^k - \boldsymbol{\Lambda}^k) - (\boldsymbol{F}(\boldsymbol{\Omega}^k) - \boldsymbol{F}(\boldsymbol{\Lambda}^k)]||_2 \\
&\leq ||(J^k)^{-1}||_2||J^k(\boldsymbol{\Omega}^k - \boldsymbol{\Lambda}^k) - (\boldsymbol{F}(\boldsymbol{\Omega}^k) - \boldsymbol{F}(\boldsymbol{\Lambda}^k))||_2 \\
&\leq ||(J^k)^{-1}||_2||(J^k - J(\boldsymbol{\eta}))(\boldsymbol{\Omega}^k - \boldsymbol{\Lambda}^k)||_2 \\
&\leq Ch^{-2}||J^k - J(\boldsymbol{\eta})||_2||\boldsymbol{\Omega}^k - \boldsymbol{\Lambda}^k||_2,
\end{aligned} \tag{3.19}$$

for some $\boldsymbol{\eta} \in [\boldsymbol{\Lambda}^k, \boldsymbol{\Omega}^k]$. Now $J^k - J(\boldsymbol{\eta})$ is a diagonal matrix, so

$$|(J^k - J(\boldsymbol{\eta}))_{p,p}| \leq \delta|w_p|\ |\{(\exp(\gamma_p - V_p) - \exp(W_p - \gamma_p)\}\{\Omega_p^k - \Lambda_p^k\}|, \tag{3.20}$$

where $\boldsymbol{\gamma} \in [\boldsymbol{\Lambda}^k, \boldsymbol{\Omega}^k] \subset [\boldsymbol{\Lambda}^0, \boldsymbol{\Omega}^0]$. Since $\boldsymbol{\Lambda}^0$, $\boldsymbol{\Omega}^0$ are bounded independently of h and since by (3.2), $\boldsymbol{V}$ and $\boldsymbol{W}$ are similarly bounded, it follows that

$$\begin{aligned}
\|J^k - J(\boldsymbol{\eta})\|_2 &\leq C\max_p |w_p|\|\boldsymbol{\Omega}^k - \boldsymbol{\Lambda}^k\|_\infty \\
&\leq C\max_p |w_p|\|\boldsymbol{\Omega}^k - \boldsymbol{\Lambda}^k\|_2
\end{aligned} \tag{3.21}$$

Also, by recalling (2.14), w_p is bounded by one third of the area of the circle with radius h centered on node p. Then (3.19)-(3.21) yield the required result. □

3.2 Calculation of the electron and hole quasi-Fermi potentials

In the previous subsection we have discussed the solution of (2.17) for $\tilde{\Psi}$, given that $(V, W) \in B$. We have shown that $\tilde{\Psi}$ is pointwise bounded independently of h, and may be computed by the quasi-Newton method (3.6)-(3.9). In this subsection we obtain the Lipschitz continuity of the map $\mathcal{G}$ defined following (2.16). The first step, given in Theorem 3.5, is to examine the continuity of the fractional step (2.17). Then in Theorem 3.6 we examine the continuity of steps (2.18), (2.19). To do this we make the simplifying assumption (in (2.2),(2.3))

$$r = 0. \tag{3.22}$$

This assumption is for theoretical purposes. It allows us to employ a discrete maximum principle to bound the solutions of (2.18),(2.19). It is made by all the other convergence analyses of Gummel's method of which we are aware ([16], [18]- [20]). It can be physically justified to some extent in the case where the device is a $p - n$ diode in reverse bias, but in general the recombination/generation rate is an essential part of the physical model and is significant in size at least in part(s) of Ω. It remains an open question to repeat the present analysis without the assumption (3.22).

In the next two results, (V^i, W^i), $i = 1, 2$ will denote two arbitrary elements of B. For each i, $\tilde{\Psi}^i$ will be the corresponding solutions of (2.17), and $(\tilde{V}^i, \tilde{W}^i)$ will be the corresponding solutions of (2.18), (2.19), with all solutions satisfying the appropriate Dirichlet boundary conditions. Our methods of proof are adapted from [19], with changes necessary to handle the different set of variables used here, as well as to deal with the approximations due to mass-lumping and harmonic averaging. In addition, through the use of a discrete Sobolev inequality, we are able to obtain the behaviour of the Lipschitz constant of $\mathcal{G}$ with respect to h.

Throughout the following proofs C will denote a generic constant which is always independent of h, but may depend on other parameters as stated.

THEOREM 3.5 *For each $M > 0$, there exists a constant C which is independent of h such that*

$$|\tilde{\Psi}^1 - \tilde{\Psi}^2|_{H^1(\Omega)} \leq C \, \|(V^1, W^1) - (V^2, W^2)\|_B,$$

for all $(V^1, W^1), (V^2, W^2) \in B$, provided $\max\{|\underline{\alpha}|, |\overline{\alpha}|\} \leq M$.

Proof. By Theorem 3.3, given $(V^i, W^i) \in B$, for $i = 1, 2$, there is a unique $\tilde{\Psi}^i$ satisfying (2.7) and

$$\lambda^2 (\nabla \tilde{\Psi}^i, \nabla \phi) + \langle \delta \{\exp(\tilde{\Psi}^i - V^i) - \exp(W^i - \tilde{\Psi}^i)\} - d, \phi \rangle = 0$$

for all test functions $\phi \in S_h$ which vanish on $\partial \Omega_D$. Subtracting the case $i = 1$ from the case $i = 2$ and putting $\phi = \tilde{\Psi}^2 - \tilde{\Psi}^1$ yields

$$\begin{aligned} &\lambda^2 (\nabla (\tilde{\Psi}^2 - \tilde{\Psi}^1), \nabla (\tilde{\Psi}^2 - \tilde{\Psi}^1)) \\ &\quad + \delta \langle \exp(\tilde{\Psi}^2 - V^2) - \exp(W^2 - \tilde{\Psi}^2) - \exp(\tilde{\Psi}^1 - V^1) + \exp(W^1 - \tilde{\Psi}^1), \tilde{\Psi}^2 - \tilde{\Psi}^1 \rangle \\ &\quad = 0. \end{aligned}$$

We rewrite this as

$$\lambda^2(\nabla(\tilde\Psi^2 - \tilde\Psi^1),\ \nabla(\tilde\Psi^2 - \tilde\Psi^1)) + \delta\langle t_1,\ \tilde\Psi^2 - \tilde\Psi^1\rangle = -\delta\langle t_2,\ \tilde\Psi^2 - \tilde\Psi^1\rangle, \tag{3.23}$$

where

$$\begin{aligned} t_1 &= \{\exp(\tilde\Psi^2) - \exp(\tilde\Psi^1)\}\exp(-V^2) + \{\exp(-\tilde\Psi^1) - \exp(-\tilde\Psi^2)\}\exp(W^2), \\ t_2 &= \exp(\tilde\Psi^1)\{\exp(-V^2) - \exp(-V^1)\} + \exp(-\tilde\Psi^1)\{\exp(W^1) - \exp(W^2)\}. \end{aligned}$$

It is easily shown by the mean value theorem that

$$\langle t_1,\ \tilde\Psi^1 - \tilde\Psi^2\rangle \geq 0. \tag{3.24}$$

Now by (2.15) and the Cauchy-Schwarz inequality,

$$\begin{aligned} |\langle t_2,\ \tilde\Psi^2 - \tilde\Psi^1\rangle| &\leq \sum_T \mathcal{A}(T)\|t_2\|_{L_\infty(T)}\|\tilde\Psi^2 - \tilde\Psi^1\|_{L_\infty(T)} \\ &\leq \left\{\sum_T \mathcal{A}(T)\|t_2\|^2_{L_\infty(T)}\right\}^{1/2} \left\{\sum_T \mathcal{A}(T)\|\tilde\Psi^2 - \tilde\Psi^1\|^2_{L_\infty(T)}\right\}^{1/2}. \end{aligned} \tag{3.25}$$

Now using the fact that $(V^i, W^i) \in B$ for each i and the bounds on $\tilde\Psi^i$ which follow from Theorem 3.3, we have

$$\|t_2\|_{L_\infty(T)} \leq C\{\|V^2 - V^1\|_{L_\infty(T)} + \|W^2 - W^1\|_{L_\infty(T)}\}.$$

But since $V^2 - V^1 \in S_h$ we can use a standard inverse inequality (see, for example [6, Theorem 3.2.6]):

$$\|V^2 - V^1\|_{L_\infty(T)} \leq C\mathcal{A}(T)^{-1/2}\|V^2 - V^1\|_{L_2(T)}.$$

Using an identical bound for $\|W^2 - W^1\|_{L_\infty(T)}$, we obtain

$$\|t_2\|_{L_\infty(T)} \leq C\mathcal{A}(T)^{-1/2}\left\{\|V^2 - V^1\|_{L_2(T)} + \|W^2 - W^1\|_{L_2(T)}\right\}.$$

Hence

$$\mathcal{A}(T)\|t_2\|^2_{L_\infty(T)} \leq 2C^2\{\|V^2 - V^1\|^2_{L_2(T)} + \|W^2 - W^1\|^2_{L_2(T)}\}.$$

Thus, recalling that C is a generic constant,

$$\begin{aligned} \sum_T \mathcal{A}(T)\|t_2\|^2_{L_\infty(T)} &\leq C\left\{\|V^2 - V^1\|^2_{L_2(\Omega)} + \|W^2 - W^1\|^2_{L_2(\Omega)}\right\} \\ &\leq C\|(V^2, W^2) - (V^1, W^1)\|^2_B. \end{aligned} \tag{3.26}$$

By a similar but easier argument,

$$\begin{aligned} \sum_T \mathcal{A}(T)\|\tilde\Psi^2 - \tilde\Psi^1\|^2_{L_\infty(T)} \leq C\sum_T \|\tilde\Psi^2 - \tilde\Psi^1\|^2_{L_2(T)} &= C\|\tilde\Psi^2 - \tilde\Psi^1\|^2_{L_2(\Omega)} \\ &\leq C|\tilde\Psi^2 - \tilde\Psi^1|^2_{H^1(\Omega)}, \end{aligned} \tag{3.27}$$

where the final step uses Poincarés inequality. Substituting (3.26), (3.27) into (3.25) yields

$$|\langle t_2, \tilde\Psi^2 - \tilde\Psi^1\rangle| \leq C\|(V^2, W^2) - (V^1, W^1)\|_B|\tilde\Psi^2 - \tilde\Psi^1|_{H^1(\Omega)}. \tag{3.28}$$

Now (3.23), (3.24) and (3.28) yield the required result. $\square$

THEOREM 3.6 *For each $M > 0$, there exists a constant C independent of h such that*

$$\|(\tilde{V}^1, \tilde{W}^1) - (\tilde{V}^2, \tilde{W}^2)\|_B \leq C \max\{|\underline{\alpha}|, |\overline{\alpha}|\}(1-\log(h))^{1/2}\|(V^1, W^1) - (V^2, W^2)\|_B, \quad (3.29)$$

for all $(V^1, W^1), (V^2, W^2) \in B$, provided $\max\{|\underline{\alpha}|, |\overline{\alpha}|\} \leq M$.

Proof To show that $(\tilde{V}^i, \tilde{W}^i) \in B$ we show that

$$\underline{\alpha} \leq \tilde{V}^i \leq \overline{\alpha}.$$

The corresponding result for $\tilde{W}^i$ is analogous. For this argument let $\overline{\alpha}$ denote the element of S_h which takes the value $\overline{\alpha}$ at every node on $\Omega \cup \partial\Omega$. Clearly then $\nabla\ \overline{\alpha} = \mathbf{0}$. So by (2.18),

$$(\exp(\overline{\tilde{\Psi}^i - V^i})\nabla(\tilde{V}^i - \overline{\alpha}),\ \nabla\phi_p) = 0, \quad i = 1, 2, \qquad (3.30)$$

for all $p \notin \partial\Omega_D$. Letting $\boldsymbol{x}, \boldsymbol{x}_D$ denote the values of $\tilde{V}^i - \overline{\alpha}$ at nodes in $\overline{\Omega}\backslash\partial\Omega_D$ and on $\partial\Omega_D$ respectively, we see that (3.30) is in the form

$$K\boldsymbol{x} + K_D\boldsymbol{x}_D \ = \mathbf{0},$$

where (cf. the discussion following (3.3)) K represents the coupling between the nodes on $\overline{\Omega}\backslash\partial\Omega_D$ induced by the first bilinear form in (2.18), whereas K_D represents the analogous coupling between the nodes on $\overline{\Omega}\backslash\partial\Omega_D$ and $\partial\Omega_D$. By the assumed properties of the meshes, we know that $K^{-1} > 0$ and that K_D contains only non-positive entries. Then, (since $\boldsymbol{x}_D \leq \mathbf{0}$), we have

$$\boldsymbol{x} = -K^{-1}\ K_D\ \boldsymbol{x}_D \leq \mathbf{0}.$$

This proves $\tilde{V}^i \leq \overline{\alpha}$. The proof of $\tilde{V}^i \geq \underline{\alpha}$ is analogous.

To prove the bound (3.29) we proceed again as in [18, Chapter 4]. Write

$$\begin{aligned}
&(\exp(\overline{\tilde{\Psi}^1 - V^1})\nabla(\tilde{V}^2 - \tilde{V}^1),\ \nabla(\tilde{V}^2 - \tilde{V}^1)) \\
&\qquad = (\{\exp\overline{\tilde{\Psi}^1 - V^1}) - \exp(\overline{\tilde{\Psi}^2 - V^2})\}\ \nabla\tilde{V}^2,\ \nabla(\tilde{V}^2 - \tilde{V}^1)) \\
&\qquad\quad + (\exp(\overline{\tilde{\Psi}^2 - V^2})\ \nabla\tilde{V}^2,\ \nabla(\tilde{V}^2 - \tilde{V}^1)) \\
&\qquad\quad - (\exp(\overline{\tilde{\Psi}^1 - V^1})\ \nabla\tilde{V}^1,\ \nabla(\tilde{V}^2 - \tilde{V}^1)). \qquad (3.31)
\end{aligned}$$

By (2.18) and (3.22), the last two terms on the right-hand side of (3.31) vanish. Hence, using the fact that $(V^1, W^1), (V^2, W^2) \in B$ and Theorem 3.3, the Cauchy-Schwarz inequality gives us

$$C|\tilde{V}^2 - \tilde{V}^1|^2_{H^1(\Omega)} \ \leq\ \|\exp(\overline{\tilde{\Psi}^1 - V^1}) - \exp(\overline{\tilde{\Psi}^2 - V^2})\|_{L_\infty(\Omega)}\ |\tilde{V}^2|_{H^1(\Omega)}\ |\tilde{V}^2 - \tilde{V}^1|_{H^1(\Omega)},$$

and hence

$$|\tilde{V}^2 - \tilde{V}^1|_{H^1(\Omega)} \ \leq\ C\|\exp(\overline{\tilde{\Psi}^1 - V^1}) - \exp(\overline{\tilde{\Psi}^2 - V^2})\|_{L_\infty(\Omega)}\ |\tilde{V}^2|_{H^1(\Omega)}. \qquad (3.32)$$

We shall bound the two terms on the right-hand side of (3.32) separately. Considering the second term, recall that $\tilde{V}^2$ is defined by (2.18) with $V = V^2$, $\tilde{\Psi} = \tilde{\Psi}^2$ and $r = 0$.

For each i, define u_i to be the solution of the weak problem: Find $u_i \in H^1(\Omega)$ with $u_i|_{\partial\Omega_{Dj}} = \delta_{ij}$ for each j and such that

$$(\exp(\overline{\tilde{\Psi}^2 - V^2})\nabla u_i,\ \nabla\phi) = 0, \quad \text{for all } \phi \in H^1,\ \ \phi = 0 \ \text{ on } \ \partial\Omega_D. \tag{3.33}$$

Let $U_i \in S_h$ be the usual finite element approximation of u_i. Then by standard theory, these finite element problems are well-posed and the solution is stable in the energy norm, i.e. $|U_i|_{H^1(\Omega)} \le C$, with C independent of h. Then by uniqueness and linearity, $\tilde{V}^2 = \sum_i \alpha_i U_i$, and hence we obtain the bound for the second term of (3.32):

$$|\tilde{V}^2|_{H^1(\Omega)} \le \sum_i |\alpha_i||U_i|_{H^1(\Omega)} \le C\max\{|\underline{\alpha}|, |\overline{\alpha}|\}, \tag{3.34}$$

with C independent of h and α_i for each i .

Considering now the first term of (3.32), we claim that it can be bounded by

$$\|\exp(\overline{\tilde{\Psi}^1 - V^1}) - \exp(\overline{\tilde{\Psi}^2 - V^2})\|_{L_\infty(\Omega)} \ \le\ C\{\|\tilde{\Psi}^1 - \tilde{\Psi}^2\|_{L_\infty(\Omega)} + \|V^1 - V^2\|_{L_\infty(\Omega)}\}. \tag{3.35}$$

This can be proved by observing that on each triangle T,

$$\begin{aligned}
&|\exp(\overline{\tilde{\Psi}^1 - V^1}) - \exp(\overline{\tilde{\Psi}^2 - V^2})| \\
&\qquad = \mathcal{A}(T)\left|\left\{\int_T \exp(V^1 - \tilde{\Psi}^1)\right\}^{-1} - \left\{\int_T \exp(V^2 - \tilde{\Psi}^2)\right\}^{-1}\right| \\
&\qquad = \mathcal{A}(T)\frac{|\int_T\{\exp(V^2 - \tilde{\Psi}^2) - \exp(V^1 - \tilde{\Psi}^1)\}|}{\{\int_T \exp(V^1 - \tilde{\Psi}^1)\}\{\int_T \exp(V^2 - \tilde{\Psi}^2)\}} \\
&\qquad \le C\mathcal{A}(T)^{-1}\int_T |\exp(V^2 - \tilde{\Psi}^2) - \exp(V^1 - \tilde{\Psi}^1)| \\
&\qquad \le C\|\exp(V^2 - \tilde{\Psi}^2) - \exp(V^1 - \tilde{\Psi}^1)\|_{L_\infty(T)}.
\end{aligned}$$

Now using the mean value theorem, the bound (3.35) follows.

In order to make use of Theorem 3.5 and to prove that the map $\mathcal{G}$ is a contraction, we would like now to bound the right-hand side of (3.35) in terms of the $H^1(\Omega)$ seminorms of $\tilde{\Psi}^1 - \tilde{\Psi}^2$ and $V^1 - V^2$. The Sobolev embedding theorem would allow us to do this if the norms on the right-hand side of (3.35) were in $L_p(\Omega)$ for some $p < \infty$, but this embedding fails when $p = \infty$. However there is a discrete Sobolev inequality which yields a mesh-dependent bound in the case $p = \infty$. This bound is well known in the domain decomposition literature (e.g. [11, Lemma 2]), and it states that if $X \in S_h$ with $X = 0$ at any point on $\overline{\Omega}$, then

$$\|X\|_{L_\infty(\Omega)} \ \le\ C(1 - \log(h))^{1/2}|X|_{H^1(\Omega)}.$$

Using this we obtain from (3.35) and using Theorem 3.5,

$$\|\exp(\overline{\tilde{\Psi}^1 - V^1}) - \exp(\overline{\tilde{\Psi}^2 - V^2})\|_{L_\infty(\Omega)} \ \le\ C(1 - \log(h))^{1/2}\|(V^1, W^1) - (V^2, W^2)\|_B,$$

with C independent of h. Using this and (3.34) in (3.32) yields

$$|\tilde{V}^2 - \tilde{V}^1|_{H^1(\Omega)} \le C\max\{|\underline{\alpha}|, |\overline{\alpha}|\}(1 - \log(h))^{1/2}\|(V^1, W^1) - (V^2, W^2)\|_B.$$

An analogous bound is obtained for $|\tilde{W}^2 - \tilde{W}^1|_{H^1(\Omega)}$, completing the proof. $\square$.

The following corollary is obtained using the contraction mapping theorem.

COROLLARY 3.7 *With $r = 0$ Gummel's method (2.17) – (2.19) converges for each fixed h provided* $\max\{|\underline{\alpha}|, |\overline{\alpha}|\}$ *is sufficiently small.*

This section shows overall that the convergence of Gummel's method only degrades logarithmically with h as the mesh is refined. Each iterate of Gummel's method requires the solution of a large sparse system of equations. All these systems are symmetric positive definite, but (especially in the case of (2.18), (2.19)) they are very poorly conditioned due to severe layers in the exponential coefficients. In the following section we describe a parallel method of solving these systems whose rate of convergence also degrades only logarithmically as the mesh is refined, and whose performance is independent of jump discontinuities across substructure boundaries of the coefficients of the underlying PDEs. In Section 5 we see examine experimentally how these methods perform in practice with coefficients exhibiting layer behaviour.

4 Parallel solution of linear systems

Each step of the iteration proposed in (2.17)-(2.19) (combined with the quasi-Newton method in §3.1 for (2.17)) amounts to the finite element approximation of a linear PDE problem in the following general form (with the same notation as in §2): Find $u \in H^1(\Omega)$ satisfying $u = g$ on $\partial\Omega_D$, such that

$$(a\nabla u, \nabla\phi) + (cu, \phi) = (f, \phi), \tag{4.1}$$

for all $\phi \in H^1(\Omega)$ with $\phi = 0$ on $\partial\Omega_D$. To cover all the systems arising in §2, we assume only that a is piecewise continuous and bounded above and below by positive constants, that c is continuous and non-negative, and f is piecewise continuous. In our applications all these functions may suffer severe (finite) jumps across narrow layers interior to the domain Ω. For convenience we shall assume in this section that $g = 0$. The introduction of $g \neq 0$ just yields a different right-hand side in the finite element equations corresponding to (4.1).

4.1 Basic domain decomposition technique

Recall the decomposition of Ω into substructures $\Omega^{(i)}$ and the corresponding coarse and fine grids that were defined in §2, and introduce the following notation:

- $\partial\Omega^{(i)}$ = boundary of $\Omega^{(i)}$.
- $\Gamma^{(i)} = \partial\Omega^{(i)} \backslash \partial\Omega_D$ = substructure boundaries on which a Dirichlet boundary condition is not present.
- $\Pi_h^{(i)}$ = nodes of the fine grid which lie on $\Gamma^{(i)}$.
- $\Pi_h = \cup_i \Pi_h^{(i)}$.
- $\Pi_H^{(i)}$ = nodes of the coarse grid which lie on $\Gamma^{(i)}$.

- $\Pi_H = \cup_i \Pi_H^{(i)}$.

For any set of nodes $\mathcal{N}$ a **nodal vector** on $\mathcal{N}$ is defined to be a vector with a unique entry for each node. The set of all nodal vectors on $\mathcal{N}$ is denoted $[\mathcal{N}]$.

Then the finite element method for (4.1) yields the linear system

$$\sum_i K^{(i)} \boldsymbol{x}^{(i)} = \sum_i \boldsymbol{b}^{(i)}. \tag{4.2}$$

Here $\boldsymbol{x}$ denotes the solution vector to be found and, for each i, $\boldsymbol{x}^{(i)}$ denotes its restriction to nodes in $\Omega^{(i)} \cup \Gamma^{(i)}$, where

$$K_{pq}^{(i)} = (a\boldsymbol{\nabla}\phi_p, \boldsymbol{\nabla}\phi_q)_{\Omega^{(i)}} + \langle c\phi_p, \phi_q\rangle_{\Omega^{(i)}}, \qquad b_p^{(i)} = \langle f, \phi_p\rangle_{\Omega^{(i)}}.$$

Here p, q ranges over all fine grid nodes in $\overline{\Omega}\backslash\partial\Omega_D$ and the subscripts on the inner products denote that they are taken over the substructure $\Omega^{(i)}$. (Here we have assumed that the zeroth order term in (4.1) is "mass lumped" and the load vector is calculated using quadrature, although the following development holds even if this is not done.) Each $K^{(i)}$ is a "local" matrix, having the same dimension as the number of fine mesh nodes in $\Omega^{(i)} \cup \Gamma^{(i)}$. In (4.2) each term in the summations are assumed to be extended (by zero) to a global nodal vector on the whole of $\overline{\Omega}\backslash\partial\Omega_D$ before the addition is performed. Then, in the usual way, (4.2) is equivalent to a global system $K\boldsymbol{x} = \boldsymbol{b}$ (with SPD matrix K).

We can partition the vector $\boldsymbol{x}^{(i)}$ into a part $\boldsymbol{x}_I^{(i)}$ containing its values in $\Omega^{(i)}$ and a part $\boldsymbol{x}_B^{(i)}$ containing its values on $\Gamma^{(i)}$. This induces a partition of the whole solution vector $\boldsymbol{x}$ into a part $\boldsymbol{x}_I$ of nodal values interior to substructures and a part $\boldsymbol{x}_B \in [\Pi_h]$. With obvious notation we can also partition the matrices $K^{(i)}$ and K and the vectors $\boldsymbol{b}^{(i)}$ and $\boldsymbol{b}$. Then (4.2) may be written

$$\sum_i \begin{bmatrix} K_{II}^{(i)} & K_{IB}^{(i)} \\ K_{IB}^{(i)T} & K_{BB}^{(i)} \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_I^{(i)} \\ \boldsymbol{x}_B^{(i)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{b}_I^{(i)} \\ \boldsymbol{b}_B^{(i)} \end{bmatrix}. \tag{4.3}$$

Since $K_{II}^{(i)}$ is SPD, it is then straightforward to show that $\boldsymbol{x}$ solves (4.3) if and only if $\boldsymbol{x}_I^{(i)}, \boldsymbol{x}_B^{(i)}$ solve the system:

$$\sum_i S^{(i)} \boldsymbol{x}_B^{(i)} = \sum_i \left\{ \boldsymbol{b}_B^{(i)} - K_{IB}^{(i)T} K_{II}^{(i)-1} \boldsymbol{b}_I^{(i)} \right\}, \tag{4.4}$$

$$K_{II}^{(i)} \boldsymbol{x}_I^{(i)} + K_{IB}^{(i)} \boldsymbol{x}_B^{(i)} = \boldsymbol{b}_I^{(i)} \quad \text{for each } i, \tag{4.5}$$

where $S^{(i)}$ is the "Schur complement" matrix $S^{(i)} = K_{BB}^{(i)} - K_{IB}^{(i)T} K_{II}^{(i)-1} K_{IB}^{(i)}$. We can now think of (4.4) as an equation to be solved for $\boldsymbol{x}_B \in [\Pi_h]$. Consequently, we can find each $\boldsymbol{x}_I^{(i)}$ by solving the (local) Dirichlet problems (4.5).

The matrix S is symmetric. Moreover, simple variational arguments on the positive definite quadratic form induced by K show that for all fixed $\boldsymbol{x}_B \in [\Pi_h]$,

$$\boldsymbol{x}_B^T S \boldsymbol{x}_B = \min_{\boldsymbol{x}_I} \left\{ [\boldsymbol{x}_I^T, \boldsymbol{x}_B^T] \begin{bmatrix} K_{II} & K_{IB} \\ K_{IB}^T & K_{BB} \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_I \\ \boldsymbol{x}_B \end{bmatrix} \right\}, \tag{4.6}$$

with the minimum being achieved at the unique vector $\boldsymbol{x}_I^0$ satisfying

$$K_{II}\boldsymbol{x}_I^0 + K_{IB}\boldsymbol{x}_B = \mathbf{0}. \tag{4.7}$$

Even though the underlying PDE may be much more general than Laplace's equation, the nodal vector $\boldsymbol{x}_I^0$ determined by (4.7) is usually called the *discrete harmonic extension* of $\boldsymbol{x}_B$, Since K is SPD, (4.6) implies that S is also SPD. In fact it is easily shown using (4.6), (4.7) that S is at least as well-conditioned as K. In some special circumstances S can be spectacularly well-conditioned. For example (following a remark in [2, p.1104]), suppose Ω is a square divided into two equal substructures and the coefficient c is zero while a takes different (constant) values on each of the substructures. Then, if the mesh is symmetric across the substructure interface, the condition number of S is independent of the jump in a. Other examples where S has this independence property are in [7]. However these examples seem quite special and in general the condition number of S has been observed to grow both with the number of degrees of freedom in $[\Pi_h]$ and with the jumps of the coefficients across substructure boundaries.

For convenience we suppress the subscripts in (4.4) and write it as

$$S\boldsymbol{x} = \boldsymbol{c}, \tag{4.8}$$

which we solve for $\boldsymbol{x} \in [\Pi_h]$ by the conjugate gradient method (CGM). The kernel of this algorithm (the matrix-vector products $S\boldsymbol{z}$ for any $\boldsymbol{z} \in [\Pi_h]$) are then naturally parallelised by the domain decompostion: To compute them simply break $\boldsymbol{z}$ up into parts $\boldsymbol{z}^{(i)}$, multiply these by the corresponding $S^{(i)}$ and then add together the contributions from neighbouring substructures across each substructure edge. S itself is never assembled. On massively parallel machines with thousands of processors it is then natural to assign a substructure to each processor so that this process is as parallel as possible. In this context then S can still be very large. We have found that preconditioning is essential especially for problems with highly discontinuous coefficients, even in 2D. Our preconditioning strategy is described in the next subsection.

4.2 Preconditioning

The following discussion uses the approach of Smith [30]. The effective solution of (4.8) by CGM requires the selection of an easily invertible SPD "preconditioner" $\hat{S}$. The rate of convergence of the associated preconditioned CGM is then governed by the condition number of the matrix $\hat{S}^{-1}S$. To estimate this we prove an inequality of the form

$$\underline{\kappa}\, \boldsymbol{x}^T \hat{S}\boldsymbol{x} \ \leq\ \boldsymbol{x}^T S\boldsymbol{x} \ \leq\ \overline{\kappa}\, \boldsymbol{x}^T \hat{S}\boldsymbol{x}, \quad \boldsymbol{x} \in [\Pi_h], \tag{4.9}$$

with constants $\underline{\kappa}, \overline{\kappa}$, from which it follows that $\kappa(\hat{S}^{-1}S) = \kappa(\hat{S}^{-1/2}S\hat{S}^{-1/2}) \ \leq\ \overline{\kappa}/\underline{\kappa}$. The construction of $\hat{S}$ and the corresponding proof of (4.9) are obtained using the decomposition of Ω into substructures. Observe that if $\boldsymbol{x} \in [\Pi_h]$ then

$$\boldsymbol{x}^T S\boldsymbol{x} = \boldsymbol{x}^T \sum_i S^{(i)}\boldsymbol{x}^{(i)} = \sum_i \boldsymbol{x}^{(i)^T} S^{(i)}\boldsymbol{x}^{(i)}, \quad \boldsymbol{x} \in [\Pi_h]. \tag{4.10}$$

In the procedure below we shall first define a preconditioner $\hat{S}^{(i)}$ for each $S^{(i)}$ and then set

$$\hat{S}\boldsymbol{x} = \sum_i \hat{S}^{(i)}\boldsymbol{x}^{(i)}, \quad \boldsymbol{x} \in [\Pi_h], \tag{4.11}$$

where, as before, summation means "extension by zero then summation". Thus

$$\boldsymbol{x}^T\hat{S}\boldsymbol{x} = \sum_i \boldsymbol{x}^{(i)^T}\hat{S}^{(i)}\boldsymbol{x}^{(i)}, \quad \boldsymbol{x} \in [\Pi_h], \tag{4.12}$$

and so if $\hat{S}^{(i)}$ is a "good" preconditioner for $S^{(i)}$, we expect that $\hat{S}$ will be a "good" preconditioner for S.

Consider any vector $\boldsymbol{x}$ defined at the fine grid nodes on all of $\cup_i\partial\Omega^{(i)}$. We can partition it into $\boldsymbol{x}^T = (\boldsymbol{x}_E^T, \boldsymbol{x}_H^T)$, with $\boldsymbol{x}_E$ containing the values of $\boldsymbol{x}$ at *interior* nodes of substructure edges and $\boldsymbol{x}_H$ containing values at substructure corners (i.e. coarse grid nodes). We can also express $\boldsymbol{x}$ in terms of the *hierarchical basis* obtained simply by taking the standard basis vectors at interior nodes of substructure edges and adding those vectors which are standard basis vectors on the coarse grid nodes and linear between them. If $\boldsymbol{y}$ is the vector of coordinates of $\boldsymbol{x}$ with respect to this new basis, then

$$\begin{bmatrix} \boldsymbol{x}_E \\ \boldsymbol{x}_H \end{bmatrix} = \begin{bmatrix} I & \mathcal{R}_H^T \\ 0 & I \end{bmatrix} \begin{bmatrix} \boldsymbol{y}_E \\ \boldsymbol{y}_H \end{bmatrix}. \tag{4.13}$$

The matrix appearing on the right-hand side is the change of basis matrix from standard to hierarchical basis. In fact $\mathcal{R}_H^T\boldsymbol{y}_H$ is the linear interpolant to $\boldsymbol{y}_H$ evaluated at interior nodes of substructure edges. The relationship (4.13) is for all $\boldsymbol{x}$ and $\boldsymbol{y}$ defined at the fine grid nodes on $\cup_i\partial\Omega^{(i)}$, but it can also be used for $\boldsymbol{x}$ and $\boldsymbol{y} \in [\Pi_h]$, by simply understanding $\boldsymbol{x}$ and $\boldsymbol{y}$ to be extended by zero at nodes on $\partial\Omega_D$. Analogously, for each i, a nodal vector $\boldsymbol{x}^{(i)}$ on $\partial\Omega^{(i)}$ may be partitioned as $\boldsymbol{x}^{(i)T} = (\boldsymbol{x}_E^{(i)T}, \boldsymbol{x}_H^{(i)T})$, and we have the corresponding *local* change to hierarchical basis:

$$\begin{bmatrix} \boldsymbol{x}_E^{(i)} \\ \boldsymbol{x}_H^{(i)} \end{bmatrix} = \begin{bmatrix} I & \mathcal{R}_H^{(i)T} \\ 0 & I \end{bmatrix} \begin{bmatrix} \boldsymbol{y}_E^{(i)} \\ \boldsymbol{y}_H^{(i)} \end{bmatrix}. \tag{4.14}$$

The matrix in (4.14) is simply the minor of that in (4.13), obtained by selecting only the rows and columns corresponding to fine grid nodes on $\partial\Omega^{(i)}$. With respect to the hierarchical basis, the bilinear form induced by the matrix $S^{(i)}$ is represented by

$$\begin{bmatrix} I & 0 \\ \mathcal{R}_H^{(i)} & I \end{bmatrix} \begin{bmatrix} S_{EE}^{(i)} & S_{EH}^{(i)} \\ S_{EH}^{(i)T} & S_{HH}^{(i)} \end{bmatrix} \begin{bmatrix} I & \mathcal{R}_H^{(i)T} \\ 0 & I \end{bmatrix}. \tag{4.15}$$

Since the change to hierarchical basis often produces a better conditioned stiffness matrix (see, for example [34]), (4.15) is approximated by discarding the off-diagonal blocks and replacing the block $S_{EE}^{(i)}$ by an approximation $\hat{S}_{EE}^{(i)}$ which neglects coupling between nodes on different edges of $\Omega^{(i)}$. This yields the block diagonal matrix:

$$\begin{bmatrix} \hat{S}_{EE}^{(i)} & 0 \\ 0 & \hat{S}_{HH}^{(i)} \end{bmatrix} := \begin{bmatrix} \hat{S}_{EE}^{(i)} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ \mathcal{R}_H^{(i)} & I \end{bmatrix} \begin{bmatrix} S_{EE}^{(i)} & S_{EH}^{(i)} \\ S_{EH}^{(i)T} & S_{HH}^{(i)} \end{bmatrix} \begin{bmatrix} 0 & \mathcal{R}_H^{(i)T} \\ 0 & I \end{bmatrix}, \tag{4.16}$$

which contains four independent diagonal blocks, one for the interior nodes of each of the edges of $\Omega^{(i)}$ plus a diagonal block which relates the values at the four corners of $\Omega^{(i)}$. Transforming back to standard basis we obtain the matrix

$$\hat{S}^{(i)} := \begin{bmatrix} I & 0 \\ -\mathcal{R}_H^{(i)} & I \end{bmatrix} \begin{bmatrix} \hat{S}_{EE}^{(i)} & 0 \\ 0 & \hat{S}_{HH}^{(i)} \end{bmatrix} \begin{bmatrix} I & -\mathcal{R}_H^{(i)T} \\ 0 & I \end{bmatrix}, \tag{4.17}$$

which we use as a preconditioner for $S^{(i)}$. Then, defining $\hat{S}$ by (4.11) yields

$$\hat{S} = \begin{bmatrix} I & 0 \\ -\mathcal{R}_H & I \end{bmatrix} \begin{bmatrix} \hat{S}_{EE} & 0 \\ 0 & \hat{S}_{HH} \end{bmatrix} \begin{bmatrix} I & -\mathcal{R}_H^T \\ 0 & I \end{bmatrix}.$$

Here $\hat{S}_{EE}$ is block diagonal with the blocks containing the restrictions of S to the interiors of each of the edges of the substructures and $\hat{S}_{HH}$ is just a coarse grid approximation to S, using linear interpolation and its adjoint as grid transfer operators. The inverse of $\hat{S}$ is, explicitly,

$$\hat{S}^{-1} = \begin{bmatrix} \hat{S}_{EE}^{-1} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & \mathcal{R}_H^T \\ 0 & I \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & \hat{S}_{HH}^{-1} \end{bmatrix} \begin{bmatrix} 0 & 0 \\ \mathcal{R}_H & I \end{bmatrix}. \tag{4.18}$$

Therefore each preconditioning step consists of independent local solves at interior nodes of substructure edges plus a global coarse grid solve. By a refinement of the arguments in [11], [30] and [31] we can now prove the following theorem.

THEOREM 4.1 *For each i there exists a constant $C^{(i)}$ such that*

$$\left(C^{(i)}\right)^{-1}(1+\log(H/h))^{-2}\boldsymbol{x}^{(i)T}\hat{S}^{(i)}\boldsymbol{x}^{(i)} \;\leq\; \boldsymbol{x}^{(i)T}S^{(i)}\boldsymbol{x}^{(i)} \;\leq\; 5\boldsymbol{x}^{(i)T}\hat{S}^{(i)}\boldsymbol{x}^{(i)}, \tag{4.19}$$

for all $\boldsymbol{x}^{(i)} \in [\Pi_h^{(i)}]$. The constants $C^{(i)}$ depend only on the restrictions of the coefficients a and c of (4.1) to the subdomain $\Omega^{(i)}$.

Proof
CASE I: If $\partial\Omega^{(i)} \cap \partial\Omega_D \neq \emptyset$, or if the coefficient c in (4.1) is positive at any fine grid node in $\Omega^{(i)}$, then $S^{(i)}$ and $\hat{S}^{(i)}$ are SPD and

$$\hat{S}^{(i)^{-1}} := \begin{bmatrix} \hat{S}_{EE}^{(i)^{-1}} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & \mathcal{R}_H^{(i)T} \\ 0 & I \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & \hat{S}_{HH}^{(i)^{-1}} \end{bmatrix} \begin{bmatrix} 0 & 0 \\ \mathcal{R}_H^{(i)} & I \end{bmatrix}. \tag{4.20}$$

Let E denote the interior of any of the edges of $\Gamma^{(i)}$ and define subspaces of $[\Pi_h^{(i)}]$ by

$$\begin{aligned} \mathcal{V}_E^{(i)} &= \{\boldsymbol{x}^{(i)} \in [\Pi_h^{(i)}] : \, x_p^{(i)} = 0, \; p \notin E\}, \\ \mathcal{V}_H^{(i)} &= \mathrm{Im}\begin{bmatrix} 0 & \mathcal{R}_H^{(i)T} \\ 0 & I \end{bmatrix}. \end{aligned}$$

Let $\mathcal{P}_E^{(i)}, \mathcal{P}_H^{(i)}$ denote the orthogonal projections (with respect to the inner product induced by $S^{(i)}$) of $[\Pi_h^{(i)}]$ onto $\mathcal{V}_E^{(i)}$ and $\mathcal{V}_H^{(i)}$. A short calculation then shows that $\hat{S}^{(i)^{-1}}S^{(i)} =$

$\sum_E \mathcal{P}_E^{(i)} + \mathcal{P}_H^{(i)}$, where the sum is over the edges E of $\Gamma^{(i)}$. The well-known additive Schwarz theory (e.g. [11], [12], [31]) can now be applied. Since $\mathcal{P}_E^{(i)}, \mathcal{P}_H^{(i)}$ are orthogonal projections, the largest eigenvalue of $\hat{S}^{(i)^{-1}} S^{(i)}$ is no more than 5. To bound the smallest eigenvalue we combine the estimate in [11, §5] with the procedure in [31, Theorem 4.1] (which transforms energies induced by $K^{(i)}$ to energies induced by $S^{(i)}$) to show that each $\boldsymbol{x}^{(i)} \in [\Pi_h^{(i)}]$ has the representation $\boldsymbol{x}^{(i)} = \sum_E \boldsymbol{x}_E^{(i)} + \boldsymbol{x}_H^{(i)}$, where $\boldsymbol{x}_E^{(i)} \in \mathcal{V}_E^{(i)}$, $\boldsymbol{x}_H^{(i)} \in \mathcal{V}_H^{(i)}$ and where

$$\sum_E \boldsymbol{x}_E^{(i)T} S^{(i)} \boldsymbol{x}_E^{(i)} + \boldsymbol{x}_H^{(i)T} S^{(i)} \boldsymbol{x}_H^{(i)} \le C^{(i)}(1 + \log(H/h))^2 \boldsymbol{x}^{(i)T} S^{(i)} \boldsymbol{x}^{(i)},$$

where $C^{(i)}$ depends on the variation of the coefficients a and c in $\overline{\Omega^{(i)}}$ and not on H or h. Then a simple application of P.L. Lions's lemma (e.g. [12, Lemma 2.1]) shows that the smallest eigenvalue of $\hat{S}^{(i)^{-1}} S^{(i)}$ is bounded below by $C^{(i)^{-1}}(1 + \log(H/h))^{-2}$. Expressing this using the Rayleigh quotient induced by $\hat{S}^{(i)}$ yields (4.19).

CASE II: If $\partial\Omega^{(i)} \cup \partial\Omega_D = \emptyset$ *and* $c \equiv 0$ on $\Omega^{(i)}$, then $K^{(i)}$ is the stiffness matrix corresponding to a pure Neumann problem for a PDE with only order second terms. Letting $\mathbf{1}_H^{(i)}$, $\mathbf{1}_h^{(i)}$ denote the unit vectors in $[\Pi_H^{(i)}]$, $[\Pi_h^{(i)}]$ respectively, it is easily seen that both $S^{(i)}$ and $\hat{S}^{(i)}$ are SPD on the restricted space $\mathrm{Im}(S^{(i)}) = \mathrm{Im}(\hat{S}^{(i)}) = \{\mathrm{span}(\mathbf{1}_h^{(i)})\}^{\perp}$. Similarly $\hat{S}_{HH}^{(i)}$ is SPD on $\{\mathrm{span}(\mathbf{1}_H^{(i)})\}^{\perp}$. The matrix on the right-hand side of (4.20) is well defined on $\mathrm{Im}(S^{(i)}) = \mathrm{Im}(\hat{S}^{(i)})$, and is a left-inverse for $\hat{S}^{(i)}$. The arguments of Case I can then be repeated for this case with $[\Pi_h^{(i)}]$, $\mathcal{V}_E^{(i)}$ and $\mathcal{V}_H^{(i)}$ being replaced by their orthogonal projections onto $\mathrm{Im}(S^{(i)})$. This yields inequalities (4.19) for all $\boldsymbol{x}^{(i)} \in \mathrm{Im}(S^{(i)})$. But since $\mathrm{Ker}(S^{(i)}) = \mathrm{Ker}(\hat{S}^{(i)}) = \{\mathrm{Im}(S^{(i)})\}^{\perp}$, (4.19) is also true in this case for all $\boldsymbol{x}^{(i)} \in [\Pi_h^{(i)}]$. $\square$

Now we can sum (4.19) over all substructures and recall (4.10), (4.12) to obtain the following corollary.

COROLLARY 4.2

$$\kappa(\hat{S}^{-1}S) \le 5 \max_i (C^{(i)})(1 + \log(H/h))^2.$$

Remark. The two-dimensional analogue of the results in [30] shows that $\kappa(\hat{S}^{-1}S)$ is independent of the coefficient jumps and also that it grows logarithmically with H/h. Corollary (4.2) provides a more precise statement of these facts, with both results provided by a single statement.

We now discuss the performance of this method on the semiconductor applications.

5 Computations on the MasPar MP–1

The Machine

A version of the algorithm described above has been implemented on a MasPar MP–1 parallel computer. This is a single instruction multiple data (SIMD) machine with (in general) a two dimensional array of an integer multiple of 1024 processing elements (PE's). The machine available to us has a 32×32 array of PE's each with an 1.8 MIPS control processor, forty 32–bit registers and 16 KBytes of RAM. The relatively small memory of each PE dictates that in applications storage should be kept to a minimum.

The PE array receives either data or instructions from an array control unit (ACU). The ACU broadcasts the same instruction to each PE simultaneously. The PE array is joined in a toroidal wrap so that each processor may communicate with any one of its 8 neighbours: This is X–Net communication, the fastest communication path in the system. Each processor may also communicate with others in the PE array via a global router. This, in general, is slightly slower than X–Net communication. Finally there is a UNIX subsystem "front end" to the machine which provides services to the data–parallel unit and performs all operations that cannot be implemented in parallel. Communication between the ACU and the PE array is the slowest in the system and thus inherently serial operations should be avoided.

Implementation

The implementation is in MasPar Fortran which is based on FORTRAN77 with array processing features added from the FORTRAN90 ANSI standard and other enhancements from DEC's VAX FORTRAN. As described at the end of §4.1, we assign a single substructure to each PE. As a preprocessing step, we compute the local Schur complement matrices $S^{(i)}$ on each processor. These are small (for our implementation) and can be computed in parallel, leading to a fast implementation of matrix-vector products involving S. The remaining implementation issue is the efficient application of the preconditioner (4.18). Rather than forming and explicitly inverting the matrices corresponding to the local edge and coarse grid problems (as suggested in [31]), we have chosen to solve these inner preconditioning problems iteratively. This is especially necessary in the case of the coarse grid problem which may still be very large. Although many choices of iterative method are possible, we have chosen to do these inner iterations again by the conjugate gradient method. This involves matrix-vector products with the matrices $\hat{S}^{(i)}_{EE}$ and $\hat{S}^{(i)}_{HH}$ which can be done in a massively parallel fashion as described below.

Edge Space Preconditioner.

Each substructure we require the solution of "edge problems" with coefficient matrix composed of the sum of at most two of the local matrices $\hat{S}^{(i)}_{EE}$ restricted to that edge. Hence each matrix-vector product in the iterative solution of each edge problem is simply the sum of two local matrix-vector products restricted to nodes on that edge. In the case of a regular grid this requires only X–Net communication. In the case of a pure Dirichlet problem on a quadrilateral domain with a regular grid we store information about the southern and eastern edges of a subdomain in each associated processor and hence we can complete all the local multiplications in two sweeps with each sweep containing, at most, two parallel steps. When we introduce some Neumann boundaries we may need one extra sweep to cope with the multiplications in the boundary substructures.

Coarse Grid Preconditioner.

In the preprocessing stage that generates the local Schur complement matrices $S^{(i)}$ on each processor we also derive (in parallel) the local 4×4 coarse grid matrices $\hat{S}^{(i)}_{HH}$. Using these we can compute the global action of $\hat{S}_{HH}$ by parallel multiplication in all the subdomains. This takes essentially the same time as a single 4×4 matrix-vector multiplication, plus communication needed for the local addition. In practice we found that since the coarse-grid problem itself may suffer from ill–conditioning due to its size and to the effect of coefficient discontinuities, it is necessary to use some preconditioner even for this inner solve. For this we use diagonal scaling. This again can be done at a local level with the diagonal of $\hat{S}_{HH}^{(i)}$ being stored as a 4×1 vector on each processor and the global diagonal then constructed by local addition.

Stopping Criterion.

In the course of implementation of the (outer) preconditioned CGM we have to compute the *preconditioned residual* $\boldsymbol{z} := \hat{S}^{-1}\boldsymbol{r}$, where $\boldsymbol{r} = S\boldsymbol{e}$ is the residual and $\boldsymbol{e}$ is the error due to the present iterate. We stop the iteration in practice when $||\boldsymbol{z}||_2$ is bounded by some user-defined stopping tolerance. Since $\|\boldsymbol{e}\|_2 = \|S^{-1}\hat{S}\boldsymbol{z}\|_2$, this should provide a robust estimate of the error provided $\hat{S}$ is a good preconditioner for S. After many numerical experiments with inexact preconditioning solves we concluded that for this algorithm it is best to do these exactly. Thus we have iterated all the inner problems until a tolerance of 10^{-12} is achieved for the residual. (This is taken to be the diagonally scaled residual in the case of the coarse grid solves.)

Numerical Results

So far, in our implementation of this algorithm, we have made some further simplifications mainly to reduce the storage overheads on our machine and also to ease the coding. Firstly we have divided the domain into $m \times m$ equal square substructures and assigned each substructure to a PE. Hence 32 is the maximum value of m we have run. The domain has then been triangulated with a uniform mesh consisting of $n \times n$ nodes internal to each substructure. Therefore h, the fine mesh diameter, is equal to $1/(m(n+1))$. This gives us an equal load balance on each of the PE's throughout the computation. Subsequently no processors are left idle, waiting for others to finish calculations. Although the theory of this paper assumes quasi-uniformity for the meshes, there is, in principle, no difficulty in implementating the scheme with non-uniform grids. This just takes time and extra coding effort to implement indirect addressing, etc. More details of our implementation and further numerical results will be in [7].

Example 5.1 Before showing results from a model semiconductor problem we shall consider the solution of

$$-\nabla.(a\nabla u) = 0$$

over the domain shown in Figure 1 with Dirichlet conditions on the parts of the boundary indicated in bold and homogeneous Neumann boundary conditions elsewhere. We take u to be 0 on the upper Dirichlet boundary and 0.2 on the lower Dirichlet boundary. The coefficient function a is taken to be piecewise constant with value c_1 in region **A** and c_2 in region **B**. This example will highlight some of the properties of the domain decomposition algorithm discussed in §4.

Case 5.1.1: $c_1 = c_2 = 1$. Here, as there are no jumps in a, the condition number of S varies only with the number of unknowns in the problem. In Table 1 consider first the results without preconditioning. For a fixed number of substructures (i.e. m fixed), the number of iterations to convergence grows as h decreases. This growth is consistent with

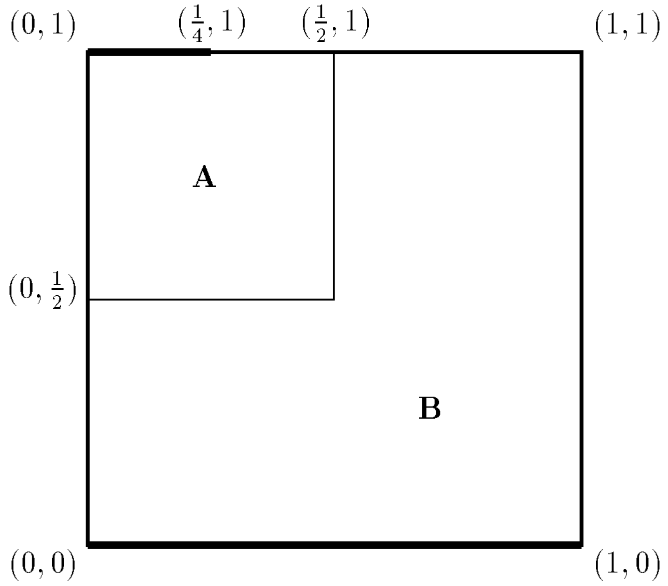

Figure 1: Domain of computation.

the linear growth of the condition number of S with the number of degrees of freedom in $[\Pi_h]$, as expected. Also, if we fix h and vary m, then convergence is quickest when m is smallest. This is because the number of degrees of freedom in $[\Pi_h]$ is then also smallest, but it hides the fact that we now have to solve larger problems to find the Schur complement. Thus each preprocessing step and also the calculation of the solution in the interior substructures will take correspondingly longer.

With preconditioning, we see that in all cases the number of iterations to convergence is greatly reduced. We also notice, as expected, a slight increase in the number of iterations as the ratio H/h increases. Finally note, that if we fix h at 1/64, say, then convergence occurs in fewest steps when m is largest. This is also due to the ratio H/h which is smallest when m is largest. However the important question is not how many preconditioned

m	n	H	h	**No preconditioner**	**Preconditioner**
8	1	1/8	1/16	43	6
8	3	1/8	1/32	65	7
8	7	1/8	1/64	93	8
16	1	1/16	1/32	83	7
16	3	1/16	1/64	123	7
16	7	1/16	1/128	173	8
32	1	1/32	1/64	158	7
32	3	1/32	1/128	231	7
32	10	1/32	1/352	385	9

Table 1: $c_1 = c_2 = 1$, Stopping tolerance $= 3.0 \times 10^{-4}$.

iterates the algorithm took (our results merely show that the theoretical estimates are sharp), but rather the efficiency of the algorithm in terms of computation time, We have timed the experiments in Table 1 and we found that, for these small "well–conditioned" problems, execution time is smallest when we choose not to precondition. This is due to the overhead in implementing the preconditioner. However when the conditioning of the system starts to deteriorate the overheads incurred in preconditioning become less influential in the comparison of the overall computation times. This can be seen in the next experiment.

Case 5.1.2: $c_1 \neq c_2$. We see from Table 2 that, for fixed h, as the ratio c_2/c_1 increases,

					No preconditioner		**Preconditioner**		Stopping
m	n	h	c_1	c_2	its	time	its	time	tolerance
32	1	1/64	10^{-1}	10^{+1}	583	42	7	43	3.0×10^{-4}
32	3	1/128	10^{-1}	10^{+1}	888	141	8	87	3.0×10^{-4}
32	7	1/256	10^{-1}	10^{+1}	1263	588	8	283	3.0×10^{-4}
32	1	1/64	10^{-2}	10^{+2}	2211	159	7	46	3.0×10^{-4}
32	3	1/128	10^{-2}	10^{+2}	3823	606	8	96	3.0×10^{-4}
32	1	1/64	10^{-3}	10^{+3}	2261	164	7	46	3.0×10^{-4}
32	3	1/128	10^{-3}	10^{+3}	5233	914	8	102	3.0×10^{-4}
32	1	1/64	10^{-4}	10^{+4}	6025	472	12	67	3.0×10^{-7}
32	1	1/64	10^{-5}	10^{+5}	7712	605	12	69	3.0×10^{-7}

Table 2: $c_1 \neq c_2$

then the CG method without preconditioning takes an increasing number of iterations to converge. The surprisingly low result of 2261 in the sixth row of this table is due to the fact that our stopping criterion (in this case with no preconditioning) is based on monitoring the unpreconditioned residual $||\boldsymbol{r}||_2$. For an ill-conditioned problem this is a poor indication of the error and the "converged" solution is likely to be poor.

We also see that the number of iterations required is enormously reduced when we precondition. Again H/h has a weak effect on the convergence rate. However it is now also obvious that the number of iterations is independent of the jumps in a across substructure boundaries when the preconditioner is applied (as predicted in Corollary 4.2). (Note that the jumps in a coincide with substructure boundaries here.) As the jumps become worse it becomes faster to precondition than not to (about 9 times faster when $c_2/c_1 = 10^{10}$). We are especially pleased that the execution times for the preconditioned algorithm also increase only modestly with the size of the jumps. Since the condition of the coarse grid matrix is heavily influenced by these jumps, there is a real danger that their presence will pollute the overall computation time for the preconditioned CGM. We have found the introduction of diagonal scaling for the coarse grid solves very influential in keeping this pollution to a minimum. Since we will have to solve many of these types of problems (with c_2/c_1 roughly $O(10^{16})$) recursively in a Gummel loop when dealing with the semiconductor equations, it is now clear that preconditioning will be essential for reasonable turnaround execution times for device modelling.

Example 5.2 Here we compute the solutions to the semiconductor system (see §2) on the domain again given in Figure 1. Dirichlet conditions are given at the contacts (now indicated by the bold portions of the boundary), and homogeneous Neumann conditions are assumed elsewhere on the boundary. We take the device to be a $p-n$ diode where the doping profile d takes the value -1 in region **A** and $+1$ in region **B**. We consider the case of reverse bias, which can be modelled by taking $v = w = 0$ at the upper contact and $v = w = \alpha$ at the lower contact where αU_T is the (positive) applied voltage. Following §2, this yields boundary conditions for ψ of $-\beta$ at the upper contact and $\beta + \alpha$ at the lower contact, where β is given in (2.9). We have used the statistics proposed in [27], namely

$$\begin{aligned} l &= 10^{-3}, \\ \epsilon &= 1.0359 \times 10^{-12}, \\ U_T &= 0.025852, \\ q &= 1.602189 \times 10^{-19}, \\ \tilde{D} &= 10^{18}, \\ n_i &= 1.22 \times 10^{10}. \end{aligned}$$

This gives the values $\lambda^2 = 1.6715 \times 10^{-7}$, $\delta = 1.22 \times 10^{-8}$ and $\beta = 18.2218$.

Numerical results are given in Tables 3-5. In practice we have found that the linear solves required for the quasi–Newton method, introduced in §3, are best performed without any preconditioning. This is because the systems involved are strongly diagonally dominant and hence only require a few CG iterations to converge. Therefore, as in Case 5.1.1 above, the quickest solution method is not to precondition. It was also found that after the first Gummel iteration the value of Ψ obtained was extremely close to the eventual solution and hence only a negligible amount of work was required in the quasi–Newton solver at subsequent Gummel iterates. On the other hand, it was found essential to use preconditioning on the solutions of the linearised continuity equations. These are particularly ill–conditioned problems. The Gummel iteration was continued until the solutions from the kth step gave a residual less than a specified tolerance at the $(k+1)$th step. This tolerance was set at 10^{-8} for the potential equation and 10^{-5} for both the electron and hole continuity equations.

Case 5.2.1: $\alpha = 0.77$ (Corresponds to a physical applied voltage of 0.02V). In Table 3 below we give results obtained with the recombination rate (r in the model in §2) switched off. The number of quasi–Newton iterates given is those required in the first Gummel iterate, after which the potential equation was essentially solved. As we can see the number is unaffected by the values of m or n. For the potential equation in subsequent Gummel iterations we use the converged upper and lower solutions from the previous Gummel iterate as starting guesses. Hence these could not be guaranteed upper and lower starting values for the new potential equation. However the quasi–Newton method appears to be robust enough to cope with this.

m	n	**Quasi–Newton its**	**Gummel its**	**Time (s)**
8	1	54	8	341
16	1	55	8	798
32	1	56	8	1666

Table 3: Applied Voltage = 0.02, recombination switched off.

We also see that the overall number of Gummel iterations required for convergence appears to be independent of H or h. The theoretical results of §3 tell us that, at worst, the convergence will deteriorate loharithmically with h. The times given are cummulative solve times for both the continuity equations (the potential equation basically contributes a one–off solve time in the first Gummel iteration). Although reasonably fast for such a difficult problem the times per linear solve are a little slower than the solve times from the model problem in Example 5.1. This is because the coefficients in the continuity equation do not now have simple jump discontinuities across substructure boundaries, but rather they exhibit a layer behaviour, where the variation may be smeared over several mesh widths. So the preconditioning described in §4 does not completely remove their influence. This effect becomes more pronounced as we increase the applied voltage. Table 4 shows results obtained for the same problem with the recombination rate switched

m	n	**Quasi–Newton its**	**Gummel its**	**Time (s)**
16	1	55	8	1065
16	2	56	8	1267

Table 4: Applied Voltage = 0.02, recombination switched on.

on. Obviously the number of quasi–Newton iterates required at the first Gummel step is unaffected, and again we see that this is independent of h. We also see that, for this example near equilibrium, the recombination rate has no effect on the total number of Gummel iterations required, but so far a proof of this fact has not been given. The extra work required to solve the systems arising from the continuity equations with non–trivial right hand side is reflected in increased execution times.

Case 5.2.2: $\alpha = 3.87$ (Corresponds to an applied voltage of 0.1V). Table 5 shows results obtained with an increased applied voltage. As predicted theoretically, this has the effect of increasing the contraction constant for the Gummel mapping. Hence we now require

m	n	**Quasi–Newton its**	**Gummel its**	**Time (s)**	r
16	1	61	25	2284	off
16	1	61	27	3006	on

Table 5: Applied Voltage = 0.1.

25 Gummel iterations for convergence when the recombination is switched off. We also see a very modest increase in the required quasi–Newton iterates, presumably because our initial starting guesses have changed. The overall increase in Gummel iterations has brought about an increased execution time. Finally we see that turning on the recombination rate increases the required number of Gummel iterates modestly and has extended the execution times.

Acknowledgements. We thank Florian Potra and Alastair Spence for useful discussions.

References

[1] I. BABUSKA and J.E. OSBORNE. Generalised finite element methods : Their performance and their relation to mixed methods. *SIAM. J. Numer. Anal.*, 20:510–536, 1983.

[2] P.E. BJØRSTAD and O.B. WIDLUND. Iterative methods for the solution of elliptic problems on regions partitioned into substructures. *SIAM J. Numer. Anal*, 23:1097–1120, 1986.

[3] J.H. BRAMBLE, J.E. PASCIAK, and A.H. SCHATZ. The construction of preconditioners for elliptic problems by substructuring. *Math. Comp.*, 47:103–134, 1986.

[4] F. BREZZI, L.D. MARINI, and P. CAPELO. Two-dimensional exponential fitting and application to drift-diffusion models . *SIAM J. Numer. Anal*, 26:1342–1355, 1989.

[5] F. BREZZI, L.D. MARINI, and P. PIETRA. Numerical simulation of semiconductor devices . *Comp. Meth. Appl. Mech. Engrg.*, 75:493–514, 1989.

[6] P.G. CIARLET. *The Finite Element Method for Elliptic Problems.* North Holland, Amsterdam, 1978.

[7] R.K. COOMER. *Parallel Iterative Methods in Semiconductor Device Modelling.* PhD thesis, School of Mathematical Sciences, University of Bath, 1993 in preparation.

[8] J.P. DARLING and I.D. MAYERGOYZ. Parallel algorithm for the solution of nonlinear Poisson equation of semiconductor device theory and its implementation on the MPP. *Journal of Parallel and Distributed Computing*, 8:161–168, 1990.

[9] P.M. DE ZEEUW. Nonlinear multigrid applied to a 1D stationary semiconductor model. *SIAM J. Sci. Stat. Comp*, 13:512–530, 1992.

[10] P. DEUFLHARD and F.A. POTRA. Asymptotic mesh independence of Newton-Galerkin methods via a refined Mysovskii theorem. *SIAM J. Numer. Anal.*, 29:1395–1412, 1992.

[11] M. DRYJA and O.B. WIDLUND. Some domain decomposition algorithms for elliptic problems. In Hayes, L. and Kincaid, D., editor, *Iterative methods for large linear systems.* Academic Press, Orlando, Florida, 1989.

[12] M. DRYJA and O.B. WIDLUND. Multilevel additive methods for elliptic finite element problems. In W. Hackbusch, editor, *Parallel algorithms for PDEs.* Vieweg, Braunschweig, Germany, 1991.

[13] H.K. GUMMEL. A self-consistent iterative scheme for one-dimensional steady state transistor calculations. *IEEE transactions on electronic devices*, ED-11:455–465, 1964.

[14] D. GUNASKERA, R.F. FOWLER, and C. GREENOUGH. Everest Solver Module, User Manual, Version 3.0 . Technical report, Rutherford Appleton Laboratory, November 1990.

[15] J.W. JEROME. Consistency of semiconductor modeling: an existence/stability analysis for the stationary Van Roosbroeck system. *SIAM. J. Appl. Math.*, 45(4):565–590, 1985.

[16] J.W. JEROME and T. KERKHOVEN. A finite element approximation theory for the drift diffusion semiconductor model. *SIAM J. Numer. Anal.*, 28(2):403–422, 1991.

[17] C. JOHNSON. *Numerical Solutions of Partial Differential Equations by the Finite Element Method.* Cambridge University Press, Cambridge, 1987.

[18] T. KERKHOVEN. *Coupled and decoupled algorithms for semiconductor simulation.* PhD thesis, Department of Computer Science, Yale University, 1985. (Yale Research Report YALEU/DCS/RR-429).

[19] T. KERKHOVEN. On the dependence of the convergence of Gummel's algorithm on the regularity of the solution. Technical Report YALEU/DCS/RR-366, Department of Computer Science, Yale University, March 1985.

[20] T. KERKHOVEN. On the effectiveness of Gummel's method. *SIAM J. Sci. Stat. Comput.*, 9(1):48–60, 1988.

[21] T. KERKHOVEN. A spectral analysis of the decoupling algorithm for semiconductor simulation. *SIAM J. Numer. Anal.*, 25(6):1299–1312, 1988.

[22] C.E. KORMAN and I.D. MAYERGOYZ. A globally convergent algorithm for the solution of the steady-state semiconductor device equations. *J. Appl. Phys.*, 68(3):1324–1334, 1990.

[23] J. MANDEL. Hybrid domain decomposition with unstructured subdomains. *Preprint*, 1993.

[24] P.A. MARKOWICH, C.A. RINGHOFER, and C. SCHMEISER. *Semiconductor Equations.* Springer, Wien - New York, 1990.

[25] P.A. MARKOWICH and M.A. ZLÁMAL. Inverse-average-type finite element discretizations of selfadjoint second-order elliptic problem. *Mathematics of Computation*, 51(184):431–449, 1988.

[26] J.M. ORTEGA and W.C. RHEINBOLDT. *Iterative Solution of Nonlinear Equations in Several Variables.* Academic Press, New York, 1970.

[27] S.J. POLAK, C. DEN HEIJER, W.H.A. SCHILDERS, and P. MARKOWICH. Semiconductor device modelling from the numerical point of view. *Internat. J. Numer. Meth. Engrg.*, 24:763–838, 1987.

[28] F.A. POTRA. Newton-like methods with monotone convergence for solving nonlinear operator equations. *Nonlinear Analysis, Theory, Methods & Applications*, 11:697–717, 1987.

[29] W.H.A. SCHILDERS. Initial guess strategies and extrapolation techniques for use in the Gummel method . *Preprint, Philips Research Laboratories, Eindhoven, Netherlands*, 1991.

[30] B.F. SMITH. A domain decomposition algorithm for elliptic problems in three dimensions. *Numer. Math.*, 60:219–234, 1991.

[31] B.F. SMITH. An optimal domain decomposition preconditioner for the finite element solution of linear elasticity problems. *Siam. J. Sci. Stat. Comput.*, 13, no.1:364–378, 1992.

[32] B.F. SMITH and O.B. WIDLUND. A domain decomposition algorithm using a hierarchical basis. *Siam. J. Sci. Stat. Comput.*, 11, no.6:1212–1220, 1990.

[33] R. S. VARGA. *Matrix Iterative Analysis.* Prentice Hall, Englewood Cliffs, 1962.

[34] H. YSERENTANT. On the multi-level splitting of finite element spaces. *Numer. Math.*, 49:379–412, 1986.

Appendix

A "Hybrid" Finite Element Scheme

To motivate the discretisations of the current continuity equations used in §2, consider the linear model problem

$$-\nabla.(a\nabla u) = f \quad \text{in } \ \Omega, \tag{A.1}$$

with boundary conditions $u = g$ on $\partial\Omega_D$ and $\partial u/\partial n = 0$ on $\partial\Omega_N$. Assume that a is bounded above and below by positive constants and (for simplicity) assume that f, g are smooth. Suppose that we are interested in approximations to the flux $a\nabla u$ which somehow model the conservation law (A.1). One strategy is to introduce some (as yet undetermined) piecewise constant functions $a_h, f_h \in \Sigma_h$ and to consider the (approximate) finite element method which computes $u_h \in S_h$ satisfying $u_h = g$ on $\partial\Omega_D$ and

$$\int_\Omega a_h \nabla u_h . \nabla\phi = \int_\Omega f_h \phi, \quad \phi \in S_h, \quad \phi = 0 \text{ on } \partial\Omega_D. \tag{A.2}$$

Consider the resulting piecewise constant function

$$\boldsymbol{J}_h := a_h \nabla u_h \in (\Sigma_h)^2 \tag{A.3}$$

which in some sense approximates the flux. It has certain weak conservation properties with respect to the (discontinuous piecewise linear) particular integral of (A.1) given on each triangle T by

$$\boldsymbol{J}_f(\boldsymbol{x}) = -\frac{1}{2}\left\{\boldsymbol{x} - \mathcal{A}(T)^{-1}\int_T \boldsymbol{y}\, \mathrm{d}\, y\right\}\{f_h|_T\}, \quad \boldsymbol{x} \in T.$$

Clearly for each T we have

$$-\nabla.\boldsymbol{J}_f = f_h \quad \text{on } T, \tag{A.4}$$

and

$$\int_T \boldsymbol{J}_f = \mathbf{0}. \tag{A.5}$$

Thus it follows that for any $\phi \in S_h$ with $\phi = 0$ on $\partial\Omega_D$

$$\sum_T \int_T (\boldsymbol{J}_h + \boldsymbol{J}_f).\nabla\phi = \sum_T \int_T \boldsymbol{J}_h.\nabla\phi = \sum_T \int_T f_h\phi. \tag{A.6}$$

Now use divergence theorem and (A.4) to obtain

$$\sum_T \int_{\partial T} (\boldsymbol{J}_h + \boldsymbol{J}_f).\boldsymbol{n}\phi = 0, \tag{A.7}$$

where for each T, $\boldsymbol{n}$ is the unit outward normal to its boundary ∂T. Now taking $\phi = \phi_p$ in (A.7) shows the local conservation property that the averages of the flux $\boldsymbol{J}_h + \boldsymbol{J}_f$ along all edges meeting at any node p is zero.

There is another quite different way of interpreting $\boldsymbol{J}_h$. Observe that the divergence theorem and (A.7) imply

$$\int_\Omega \boldsymbol{J}_h.\nabla\phi = -\sum_T \int_{\partial T} \{\boldsymbol{J}_f.\boldsymbol{n}\}\phi, \quad \phi \in S_h, \quad \phi = 0 \text{ on } \partial\Omega_D. \tag{A.8}$$

Moreover (A.3) and (A.5) also imply

$$\int_\Omega a_h^{-1}\{\boldsymbol{J}_h + \boldsymbol{J}_f\}.\boldsymbol{\tau} - \int_\Omega \nabla u_h.\boldsymbol{\tau} = 0, \quad \boldsymbol{\tau} \in (\Sigma_h)^2. \tag{A.9}$$

Equations (A.8), (A.9) constitute a “hybrid” discretisation of the mixed reformulation of (A.1) which seeks a pair ($\boldsymbol{J}$, u) such that $u = g$ on $\partial\Omega_D$, $(\boldsymbol{J}+\boldsymbol{J}_f).\boldsymbol{n} = 0$ on $\partial\Omega_N$, $(\boldsymbol{J}+\boldsymbol{J}_f)$ is continuous on Ω and

$$-\nabla.(\boldsymbol{J} + \boldsymbol{J}_f) = f, \tag{A.10}$$

$$a^{-1}\boldsymbol{J} + a^{-1}\boldsymbol{J}_f - \nabla u = \mathbf{0}. \tag{A.11}$$

Writing this in weak form, replacing a, f by a_h, f_h and seeking a solution in $(\Sigma_h)^2 \times S_h$ yields (A.8), (A.9). (This “hybrid” method is different from the usual mixed method, where more continuity is imposed on the space where $\boldsymbol{J}$ is sought and less on the space where u is sought. This method is proposed for a specific semiconductor equation (slightly different from the application in §2 of this paper) in [5].)

Comparing (A.11) with (A.9) suggests that we should choose $a_h \in \Sigma_h$ in (A.9) so that a_h^{-1} is a good approximation to a^{-1}. The obvious way to do this is to set

$$a_h = \left\{\frac{1}{\mathcal{A}(T)} \int_T a^{-1}\right\}^{-1}, \quad \text{on each } T. \tag{A.12}$$

This “harmonic average” approximation appears in the hybrid methods of [5] (see also [4]), but the method there is slightly different and corresponds to (A.2) but with a modified right-hand side.

Besides the connection with mixed methods, there are a number of other justifications for employing the approximation (A.12). Firstly if the one-dimensional analogue of (A.1) with $f = 0$ is solved by the finite element method, then the use of the harmonic average yields a scheme which is *exact* at the nodes. (This is known as the Allen-Southwell exponentially fitted scheme in the fluid dynamics literature and as the Scharfetter-Gummel discretisation in the semiconductor literature, although in these applications the exactness property is more usually presented for the equivalent convection-diffusion equations). Such an average also appears in [9], where schemes for equations of the form (A.1) are derived using a cell-centered finite volume method.

For two-dimensional versions of (A.1), harmonic averaging of coefficients (along element sides) leads to a piecewise constant approximation of the flux with accuracy which depends only on the smoothness of the flux and of the forcing term f [25]. Such methods are appropriate when the flux is smoother than a or u, as is (empirically) the case in semiconductor modelling. Related observations are made in the earlier work [1], where the use of the harmonic averaged coefficient (in one dimension) is shown to be equivalent to a "generalized finite element method" which is found to be more robust to jumps in a than the standard method.

Thus we propose to discretize (A.1) by (A.2) with a_h given by (A.12). In our application (equations (2.2),(2.3)) a_h can be easily computed in practice, For f_h in (A.2) the natural thing would be to use a standard average. However, this is not feasible in our application because of the complicated form of the zeroth order terms . Thus we approximate the average using the quadrature rule (2.13):

$$f_h = \frac{1}{3} \sum_{p_T} f(p_T) \quad \text{on each} \quad T. \tag{A.13}$$

It is a moderately easy exercise in finite element error analysis to show that method (A.2) yields a solution u_h which (as in the usual finite element method) converges to u with order h in the energy norm under appropriate smoothness assumptions on a , f and g.