Modular Algorithms for Transient Semiconductor Device Simulation, Part I: Analysis of the Outer Iteration

W. M. Coughran, Jr. Joseph W. Jerome*

Abstract

A general outer iteration, based upon linearization, is introduced at discrete time steps for the one-dimensional semiconductor device model. The iteration depends upon solving the semidiscrete device equations approximately, specifically, in such a way that the residual is of order Δt in an appropriate norm. It is shown that this maintains the order of the backward Euler method. A monitoring of the constants, including time-step requirements for solvability of the semidiscrete systems, as well as boundedness, smoothness and invertibility for the maps defining the Newton approximations, is carried out. An invariant-region principle provides an important theoretical basis, and a novel proof is provided. An essential component of the theory, providing the interface with the sequel to this paper, is that the Newton iterations are small in number, typically one or two, and may be realized as approximate Newton iterations. Continuation is employed as the time-stepping bridge.

1 Introduction

In Part I of this two part series we describe a modular development for a time-stepping algorithm, based upon the backward Euler method and associated linearization, as applied to the one-dimensional semiconductor device model. We provide a detailed mathematical analysis of the solvability of the semidiscrete equations and the associated invariant region; of the residual error tolerance in solving the semidiscrete systems approximately so as to maintain first order convergence; of an approximate Newton method, typically involving one or two iterations, serving as the outer iteration which defines the approximation strategy; and of the continuation procedure, linking the output of the previous time step to the starting iterate of the current time step. Part II of the series is concerned with a specific inner iteration, i.e., the fully discrete algorithm. In decoupling the outer and inner iterations in this way, we have established a framework in Part I for any Newton-based

1980 *Math. Subject Classification.* Primary 65M15,82A70. Secondary 65N40,35K45,47H17.
*Research supported by National Science Foundation grant DMS-8420192.

algorithm, and have demonstrated how the 'a priori' estimates lead to the control and estimation of fundamental constants required for a convergence analysis.

The devices studied here are one-dimensional, and the dependent variables selected are the electrostatic potential and the carrier concentrations. The model is described more fully in § 2. The mobility coefficients are selected to have a form similar to that used in computations for silicon devices so as to ensure the physically essential property of saturation; the Einstein relations are assumed only so that physically realistic representation of diffusion is possible. For the dependent variables selected here, the Einstein relations provide no equation simplification. The model assumes Shockley-Read-Hall recombination. As described here, the model is a special case of the spatially multidimensional model considered in [14] by Jerome, in particular, we deduce that the initial/boundary-value problem possesses a unique solution, globally in time. Uniqueness need not persist for the semidiscrete solutions without additional time-step restrictions, however, so that the linearized time-stepping must have an inherent tracking mechanism; this is provided by the continuation.

In this paper, we do not analyze second order time-stepping methods. We note, however, that one such method, possessing the property of L-stability, has been introduced in [1] by Bank et al. Issues of simulation and implementation are discussed at length in Part II. The ideas of Part I are developed fully in § 3 and § 4. In § 5 we provide a postscript, in which the paper is coalesced and a linkage to Part II is established.

2 The Transient Semiconductor Equations

Our interest here is in a transient initial/boundary-value problem, in one spatial dimension, that models the behavior of a simple semiconductor device; hence, the appropriate space-time domain is $\bar{\Omega} \times [0, T_0]$ where $\Omega = (a, b) \subset \mathbb{R}$ and T_0 is the final time of interest. The transient semiconductor equations, which hold for $(x, t) \in \Omega \times (0, T_0]$, are given by

$$-\nabla \cdot (\epsilon \nabla u) + q(n - p - N) = 0, \tag{1}$$

$$q\frac{\partial n}{\partial t} - \nabla \cdot J_n = -qR_n, \tag{2}$$

$$q\frac{\partial p}{\partial t} + \nabla \cdot J_p = -qR_p, \tag{3}$$

where

ϵ is the dielectric constant;

q is the electronic charge;

$u(x, t)$ is the electrostatic potential so that the electric field $E = -\nabla u$;

$n(x, t)$ and $p(x, t)$ are the electron and hole carrier concentrations, respectively;

$N(x)$ is the net impurity (doping) concentration;

$J_n(u,n)$ and $J_p(u,p)$ are the electron and hole current densities, respectively;

$R_n(n,p)$ and $R_p(n,p)$ are the carrier recombination-generation rates, respectively.

Note that (1) is the stationary Maxwell equation governing the electrostatic potential u while (2) and (3) are typical continuity equations governing n and p, respectively. In addition, we assume that appropriate initial data $n_0(x), p_0(x)$ and boundary data $\bar{u}(x,t), \bar{n}(x), \bar{p}(x)$ are given.

We write the current densities in the traditional drift-diffusion form (cf. Sze [19, p. 50])

$$J_n = -q\mu_n n\nabla u + qD_n\nabla n, \tag{4}$$
$$J_p = -q\mu_p p\nabla u - qD_p\nabla p, \tag{5}$$

where μ_n and μ_p are the field-dependent mobilities and D_n and D_p are the electron and hole diffusion coefficients, respectively.

One simple mobility model that includes velocity saturation has the form

$$\mu(\nabla u) = \mu_0 \left[1 + \left(\frac{\mu_0|\nabla u|}{v_{sat}}\right)^{\gamma}\right]^{-1/\gamma},$$

where μ_0 and v_{sat} are the low-field mobility value and temperature-dependent saturation velocity, respectively (cf. Caughey and Thomas [6] and Thornber [20]); usually, γ is taken to be 2 for electrons and 1 for holes. To avoid problems with smoothness, we will assume $\gamma \equiv 2$ for both carriers and write the mobility functions as

$$\mu_n = \mu_{0n}\left[1 + \left(\frac{\mu_{0n}|\nabla u|}{v_{sn}}\right)^2\right]^{-1/2}, \tag{6}$$
$$\mu_p = \mu_{0p}\left[1 + \left(\frac{\mu_{0p}|\nabla u|}{v_{sp}}\right)^2\right]^{-1/2}. \tag{7}$$

Of course, there are other possibilities for the mobilities, which are less smooth or are more complicated functions of the dependent variables. It is important to note, however, that the mobility models are largely phenomenological expressions, which attempt to incorporate a number of experimentally observed phenomena.

We supplement (4) and (5) by the Einstein relations

$$D_n = \frac{kT}{q}\mu_n, \quad D_p = \frac{kT}{q}\mu_p, \tag{8}$$

as in the Van Roosbroeck model [21]. Here, k and T are Boltzmann's constant and the temperature, respectively. To simplify matters, we select the natural units so that

$$\frac{kT}{q} \equiv 1; \tag{9}$$

in fact, we scale all of the equations in a fashion similar to that of DeMari [8, 7], except in our treatment of ϵ (see also the book by Markowich [16]).

Finally, we make use of the Shockley-Read-Hall recombination term [18, 10, 17]

$$R_n = R_p = \frac{np-1}{\tau_p(n+1)+\tau_n(p+1)} = \frac{np-1}{d} = R, \tag{10}$$

where τ_n and τ_p are electron and hole lifetimes, respectively. Equation (10) has this simple form because the carrier concentrations n and p are measured (scaled to be) in so-called effective intrinsic carrier concentration units. Our framework can be extended to include three-particle interactions, such as Auger recombination-generation, but we will not discuss these issues here.

3 Solvability and Approximate Solution of the Semidiscrete System

We begin by deriving a solvability result for the problem (1)–(10), which is based upon an invariant-region principle; since the complete verification of the solvability is quite detailed, we have presented the structure of the proof in § 3.1 and the details of the proof in § 3.2. The discussion here is related to earlier work described by Bank et al. [2] and Jerome [13]. A convergence result, where the residual is controlled in L^2, is presented in § 3.3, while the corresponding H^{-1} result is presented in § 3.4.

3.1 The Solvability Result

We shall find it convenient to expand the mobility terms in (2) and (3) by the product rule, substitute the second derivative terms via the potential equation (1), and make use of the Einstein relations (8) and the scaling (9). Thus, throughout much of § 3.1, we shall write the semidiscrete system via a fully implicit time discretization as

$$-\epsilon\nabla^2 u_k + n_k - p_k = N, \tag{11}$$

$$\frac{n_k - n_{k-1}}{\Delta t_k} - \nabla\cdot(\mu_n\nabla n_k) + \mu_n\nabla u_k\nabla n_k + U_{n,k} = 0, \tag{12}$$

$$\frac{p_k - p_{k-1}}{\Delta t_k} - \nabla\cdot(\mu_p\nabla p_k) - \mu_p\nabla u_k\nabla p_k + U_{p,k} = 0, \tag{13}$$

where

$$U_{n,k} = R_k + \epsilon^{-1}(\mu_n'\nabla u_k + \mu_n)n_k(n_k - p_k - N), \tag{14}$$

$$U_{p,k} = R_k - \epsilon^{-1}(\mu_p'\nabla u_k + \mu_p)p_k(n_k - p_k - N), \tag{15}$$

and $\mu_\star$ and $\mu_\star'$ are evaluated at ∇u_k.

The invariant-region principle referred to above represents a slight weakening of the usual such principle, since the region is permitted to increase linearly (not

exponentially) with time and includes the spatially dependent doping. Specifically, we define a number

$$\lambda_0 = \max(\|\bar{n}\|_{L^\infty}, \|\bar{p}\|_{L^\infty}, \|n_0\|_{L^\infty}, \|p_0\|_{L^\infty}), \tag{16}$$

and *functions* n_k^{max} and p_k^{max}, via the relations,

$$n_0^{max} = \lambda_0 + \mathrm{Step}(N^+), \tag{17}$$
$$p_0^{max} = \lambda_0 + \mathrm{Step}(N^-), \tag{18}$$
$$n_k^{max} - n_{k-1}^{max} = p_k^{max} - p_{k-1}^{max} = 4\rho\Delta t_k, \text{ for } 1 \le k \le L, \tag{19}$$

where

$$\rho = (\tau_p + \tau_n)^{-1}, \tag{20}$$
$$\mathrm{Step}(N^+) = \begin{cases} \sup N & \text{if } N > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{21}$$

It follows that, if

$$n^{max} = \lambda_0 + \sup N^+ + 4\rho T_0, \; p^{max} = \lambda_0 + \sup N^- + 4\rho T_0, \tag{22}$$

then $n^{max} \ge n_k^{max}$ and $p^{max} \ge p_k^{max}$. As before, $\bar{n}$ and $\bar{p}$ are linear extensions of the boundary data, $N = N^+ - N^-$, and T_0 is the maximum time.

We are now prepared for the first result.

Theorem 1 *Under hypotheses on the time step (cf. (37), (62), and (69) below), there is a solution of the Dirichlet boundary-value problem (11)–(13) with boundary values*

$$u_k(x_e) = \bar{u}(x_e, t_k), \; n_k(x_e) = \bar{n}(x_e), \; p_k(x_e) = \bar{p}(x_e) \text{ for } x_e = a, b. \tag{23}$$

The solution triple satisfies an invariant-region principle for the carrier concentrations and a generalized maximum principle for the potential:

$$0 \le n_k \le n_k^{max} \le n^{max}, \tag{24}$$
$$0 \le p_k \le p_k^{max} \le p^{max}, \tag{25}$$
$$|u_k| \le \|\bar{u}(\cdot, t_k)\|_{L^\infty} + \epsilon^{-1}(e^{b-a} - 1)(n^{max} + p^{max}) \stackrel{\mathrm{def}}{=} u_k^{max}. \tag{26}$$

The solution is unique under additional conditions on the time step, described in § 3.2, and the smoothness is determined by the doping function N; the components are minimally in $W^{2,\infty}$, however. Finally, if the recombination term, R_k, is not included in the model, then $\rho = 0$ may be selected in the definition of n^{max} and p^{max}, and the invariant region is independent of T_0, so long as $\bar{u}$ is bounded.

Proof Outline: For maximum clarity, we shall outline the steps of the proof here prior to amplification in § 3.2. The approach is inductive, and assumes the systems are well-defined for $j < k$, and satisfy (24), (25), and (26).

(I) Define a map T from a closed bounded convex subset K of $L^2(\Omega) \times L^2(\Omega)$ into K as follows. For

$$K \stackrel{\text{def}}{=} \{[\tilde{n}, \tilde{p}] : 0 \le \tilde{n} \le n_k^{max}, 0 \le \tilde{p} \le p_k^{max}\}, \tag{27}$$

where the inequalities are understood pointwise almost everywhere, define $u[\tilde{n}, \tilde{p}]$, with $[\tilde{n}, \tilde{p}] \in K$, as the unique solution of

$$-\epsilon \nabla^2 u + \tilde{n} - \tilde{p} = N, \ u(a) = \bar{u}(a, t_k), \ u(b) = \bar{u}(b, t_k). \tag{28}$$

With u specified by (28), set $T[\tilde{n}, \tilde{p}] = [n, p]$, where n and p are specified by the uncoupled equations

$$\begin{aligned} F(n) &= \frac{n - n_{k-1}}{\Delta t_k} - \nabla \cdot (\mu_n \nabla n) + \mu_n \nabla u \nabla n + \tilde{U}_n(\cdot, n, \tilde{p}) = 0, \quad (29) \\ & \qquad n(a) = \bar{n}(a), \ n(b) = \bar{n}(b), \\ G(p) &= \frac{p - p_{k-1}}{\Delta t_k} - \nabla \cdot (\mu_p \nabla p) - \mu_p \nabla u \nabla p + \tilde{U}_p(\cdot, \tilde{n}, p) = 0, \quad (30) \\ & \qquad p(a) = \bar{p}(a), \ p(b) = \bar{p}(b). \end{aligned}$$

Here $\tilde{U}_n$ and $\tilde{U}_p$ are represented by

$$\begin{aligned} \tilde{U}_n(\cdot, n, \tilde{p}) &= \tilde{R}(\cdot, n, \tilde{p}) + \epsilon^{-1} \left[f(\cdot, n) - n(\tilde{p} + N) \right] (\mu_n' \nabla u + \mu_n), \quad (31) \\ \tilde{U}_p(\cdot, \tilde{n}, p) &= \tilde{R}(\cdot, \tilde{n}, p) + \epsilon^{-1} \left[p(N - \tilde{n}) + g(\cdot, p) \right] (\mu_p' \nabla u + \mu_p). \quad (32) \end{aligned}$$

We shall define the functions $\tilde{R}$, f, and g and the motivation for altering $U_{n,k}$ and $U_{p,k}$. Note that there is no 'a priori' reason why n and p must have restricted range prior to the derivation of the invariant region principle. In order to facilitate the derivation, we introduce the function $\tilde{R}(\cdot, n, p)$, which, for each fixed $x \in [a, b]$, agrees with $R(n, p)$ for $0 \le n \le n_k^{max}(x)$, $0 \le p \le p_k^{max}(x)$, but is extended outside this rectangle, to the horizontal and vertical strips uniquely intersecting in the rectangle, in such a way that the partial derivatives with respect to n and p are constant. The functions f and g have the property that, for each fixed $x \in [a, b]$, $(\partial f / \partial n)(x, n)$ and $(\partial g / \partial p)(x, p)$ agree with $2n$ and $2p$ on $[0, n_k^{max}(x)]$ and $[0, p_k^{max}(x)]$, respectively, and are continuously extended to have constant values outside this range.

(II) Show that T is well-defined. This process involves the following.

- **(1)** The derivation of 'a priori' estimates for u and ∇u.
- **(2)** The existence of solutions for (29) and (30) and derivation of corresponding 'a priori' estimates.
- **(3)** The verification that $[n, p] \in K$.

(III) Show that T has a fixed point, $[n, p] = [n_k, p_k]$, which coincides with the solution of (11)–(13) and (23).

∎

3.2 Verification of Solvability

We provide the details of (II 1, 2, 3) and (III) in the proof of Theorem 1.

(II 1) A weak $H^2(\Omega)$ solution of (28), which is the "maximal" regularity (in fact, $W^{2,\infty}$ holds), given $\tilde{n}$ and $\tilde{p}$, satisfies the usual maximum principle

$$\|u\|_{L^\infty} \leq \|\bar{u}(\cdot, t_k)\|_{L^\infty} + \epsilon^{-1}(e^{b-a} - 1)\|\tilde{n} - \tilde{p} - N\|_{L^\infty}, \tag{33}$$

as can be seen by using the proof of Theorem 3.7 of Gilbarg and Trudinger [9], and an argument by contradiction replacing the local second derivative behavior at an extreme point with the weak formulation and corresponding first derivative behavior. Alternatively, one could employ a smoothing argument, and take limits of the classical result. Inequality (26) follows immediately from (33). The estimates,

$$\|\nabla u\|_{L^\infty} \leq C\|u\|_{H^2} \leq C' u_k^{max}, \tag{34}$$

follow from a one-dimensional Sobolev embedding, and from a standard isomorphism, respectively (cf. [9, Theorem 8.12] for the latter), as well as from the representation of u_k^{max}, given by the right-hand side of (26).

(II 2) The existence of a solution of (29) can be shown by Newton's method, as we now show. The linearized equation about fixed $n \in H^2(\Omega)$ becomes, for image element $-F(n)$,

$$\begin{aligned} -F(n) &= \frac{\delta n}{\Delta t_k} - \nabla \cdot (\mu_n \nabla \delta n) + \mu_n \nabla u \nabla \delta n + \frac{\partial \tilde{R}}{\partial n}(\cdot, n, \tilde{p})\delta n \\ &\quad + \frac{1}{\epsilon}\left(\frac{\partial f}{\partial n} - \tilde{p} - N\right)(\mu_n' \nabla u + \mu_n)\delta n, \end{aligned} \tag{35}$$

$$\delta n(a) = \delta n(b) = 0. \tag{36}$$

By direct calculation, $(\partial \tilde{R}/\partial n)(\cdot, n, \tilde{p}) \geq 0$ and $\epsilon^{-1}(\partial f/\partial n)(\mu_n' \nabla u + \mu_n) \geq 0$. Thus, for

$$\Delta t_k \leq \epsilon / \{2 \max(\mu_{0n}, \mu_{0p}) \left[\max(p^{max}, n^{max}) + \|N\|_{L^\infty}\right]\}, \tag{37}$$

the coefficient of δn is nonnegative; the same condition guarantees the analogous statement for δp in the linearized hole equation. Under this format, the structural hypotheses (8.5), (8.6), and (8.8) of [9] are satisfied, so that the boundary-value problem (35) and (36) possesses a solution (cf. [9, Theorem 8.3]). An estimate for the inversion operator associated with (35) and (36) of the form,

$$\|\delta n\|_{H^2} \leq C\|F(n)\|_{L^2}, \tag{38}$$

holds, with C independent of n. This follows from [9, Theorem 8.12] after a straightforward preliminary estimate of $\|\delta n\|_{L^2}$ in terms of $\|F(n)\|_{L^2}$.

In addition to the uniform inversion property, (38), the map F, defining (29), has a Lipschitz continuous derivative:

$$\|F'(n_1) - F'(n_2)\|_{H^2, L^2} \leq M\|n_1 - n_2\|_{H^2}, \tag{39}$$

where the constant M does not depend on n_1 or n_2. Indeed, a much stronger result holds as we shall now show. We have, for $\phi \in H^2$,

$$\begin{aligned}
&\|[F'(n_1) - F'(n_2)]\phi\|_{L^2} \\
&\quad \le \left\| \frac{\partial \tilde{R}}{\partial n}(\cdot, n_1, \tilde{p}) - \frac{\partial \tilde{R}}{\partial n}(\cdot, n_2, \tilde{p}) \right\|_{L^2} \|\phi\|_{H^1} \\
&\qquad + \frac{1}{\epsilon} \left\| \left(\frac{\partial f}{\partial n}(\cdot, n_1) - \frac{\partial f}{\partial n}(\cdot, n_2) \right) (\mu_n' \nabla u + \mu_n) \right\|_{L^2} \|\phi\|_{H^1}
\end{aligned} \tag{40}$$

where we have used a standard ring property (cf. Kato [15]). By direct computation we have

$$\left\| \left(\frac{\partial f}{\partial n}(\cdot, n_1) - \frac{\partial f}{\partial n}(\cdot, n_2) \right) (\mu_n' \nabla u + \mu_n) \right\|_{L^2} \le 4\mu_{0n} \|n_1 - n_2\|_{L^2}, \tag{41}$$

while, if $n_1 \ge 0$, $n_2 \ge 0$, we have,

$$\begin{aligned}
&\left| \frac{\partial \tilde{R}}{\partial n}(\cdot, n_1, \tilde{p}) - \frac{\partial \tilde{R}}{\partial n}(\cdot, n_2, \tilde{p}) \right| \\
&\quad = \left| \frac{(\tilde{p}+1)(\tilde{p}\tau_n + \tau_p)}{[\tau_p(n_1+1) + \tau_n(\tilde{p}+1)]^2} - \frac{(\tilde{p}+1)(\tilde{p}\tau_n + \tau_p)}{[\tau_p(n_2+1) + \tau_n(\tilde{p}+1)]^2} \right| \\
&\quad \le 2(\tilde{p}+1) \left| \frac{\tau_p(n_2 - n_1)}{[\tau_p(n_1+1) + \tau_n(\tilde{p}+1)][\tau_p(n_2+1) + \tau_n(\tilde{p}+1)]} \right| \\
&\quad \le \left(\frac{2(\tilde{p}+1)\tau_p}{(\tau_p + \tau_n)^2} \right) |n_2 - n_1|,
\end{aligned}$$

which makes use of the elementary identity,

$$\frac{1}{a^2} - \frac{1}{b^2} = \left(\frac{1}{a} - \frac{1}{b} \right) \left(\frac{1}{a} + \frac{1}{b} \right).$$

This final inequality also holds on the set where $n_1 < 0$ or $n_2 < 0$, by the way R was redefined in this region. It follows that

$$\left\| \frac{\partial \tilde{R}}{\partial n}(\cdot, n_1, \tilde{p}) - \frac{\partial \tilde{R}}{\partial n}(\cdot, n_2, \tilde{p}) \right\|_{L^2} \le \left(\frac{2(p^{max}+1)\tau_p}{(\tau_p + \tau_n)^2} \right) \|n_1 - n_2\|_{L^2} \tag{42}$$

so that (39) follows 'a fortiori' from (40), (41), and (42).

Now the existence result for (29) can be deduced from the existence of a homotopy solution set, as developed by the second author in [12], applied to the map $F(\lambda, n)$, which folds F by multiplying the term $\tilde{U}_n$ in (29) by λ, where λ is a homotopy parameter, $0 \le \lambda \le 1$. The existence result for $\lambda = 0$ follows from [9, Theorem 8.3], and we accordingly begin the homotopy solution set at the solution of the linear equation, $F(0, n) = 0$. Although the result of [12, Theorem 4.1] appears

to be local, in relation to the radius r of the H^2 ball B_r, on which F is defined, as compared to the length of the homotopy interval, in fact, we may select r as large as required. This is permissible since the estimates (38) and (39) readily transfer to the homotopy map, in a manner independent of r. This completes (II 2), since the boundary-value problem (30) is handled analogously, with no additional restriction on the time step.

(II 3) The choice of the functions p_k^{max} and n_k^{max} ensures the validity of the lower bounds in the inequalities

$$\begin{aligned}
\tilde{U}_n(\cdot, n, \tilde{p}) &\leq |n| \left(\sigma + 2\epsilon^{-1}\mu_{0n} p^{max}\right), && n \leq 0, && (43)\\
\tilde{U}_n(x, n, \tilde{p}) &\geq -\rho - \mu_{0n}\epsilon^{-1} n \|N\|_{L^\infty}, && n \geq n_k^{max}(x), && (44)\\
\tilde{U}_p(\cdot, \tilde{n}, p) &\leq |p| \left(\sigma + 2\epsilon^{-1}\mu_{0p} n^{max}\right), && p \leq 0, && (45)\\
\tilde{U}_p(x, \tilde{n}, p) &\geq -\rho - \mu_{0p}\epsilon^{-1} p \|N\|_{L^\infty}, && p \geq p_k^{max}(x), && (46)
\end{aligned}$$

where

$$\sigma = 2\epsilon^{-1} \max(\mu_{0n}, \mu_{0p}) \|N\|_{L^\infty}. \tag{47}$$

We proceed now to verify the inequalities,

$$0 \leq n \leq n_k^{max},\ 0 \leq p \leq p_k^{max}. \tag{48}$$

Our verification of (48) will be based upon a fully discrete approximation, satisfying inequality (48), except possibly on an asymptotically thin set. Passage to the limit will yield the result. Initially, we assume that $\tilde{p}$ and $\tilde{n}$ are continuous. The equally spaced mesh points are denoted x_j and $x_{j+1/2} \stackrel{\text{def}}{=} (x_j + x_{j+1})/2$. Now we apply the box method to that part of the differential equation (29), exclusive of the convective term, i.e., we integrate terms in (29) over fixed subintervals $(x_{j-1/2}, x_{j+1/2})$, $1 \leq j \leq M$, with the following identifications:

$$\begin{aligned}
\int_{x_{j-1/2}}^{x_{j+1/2}} -\nabla \cdot (\mu_n \nabla n)\, dx &\approx \mu_n|_{x_{j-1/2}} \frac{n_j^h - n_{j-1}^h}{h} - \mu_n|_{x_{j+1/2}} \frac{n_{j+1}^h - n_j^h}{h}, && (49)\\
\int_{x_{j-1/2}}^{x_{j+1/2}} \frac{n - n_{k-1}}{\Delta t_k}\, dx &\approx \frac{h}{\Delta t_k}[n_j^h - n_{k-1}(x_j)], && (50)\\
\int_{x_{j-1/2}}^{x_{j+1/2}} \tilde{U}_n(\cdot, n, \tilde{p})\, dx &\approx h\tilde{U}_n\left(x_j, n_j^h, \tilde{p}(x_j)\right). && (51)
\end{aligned}$$

Here, h denotes the uniform grid spacing and n_j^h the "unknowns". The convective term, however, prior to "integration," is discretized by the following upwinding scheme:

$$\mu_n \nabla u \nabla n(x_j) \hookrightarrow (\mu_n \nabla u)(x_j)(n_j^h - n_{j-1}^h)/h, \tag{52}$$

if $(\mu_n \nabla u)(x_j) \geq 0$, while

$$\mu_n \nabla u \nabla n(x_j) \hookrightarrow (\mu_n \nabla u)(x_j)(n_{j+1}^h - n_j^h)/h, \tag{53}$$

if $(\mu_n \nabla u)(x_j) < 0$. The important properties of this discretization are now summarized:

(i) Let the discrete system be written

$$(I + L^h)n^h = n_{k-1} + B - \Delta t_k \tilde{U}_n, \tag{54}$$

where n_{k-1} and $\tilde{U}_n$ are column vectors, with components determined from (49), (50), and (51), and where B involves the boundary terms. Thus, if $\nabla u(x_1) > 0$,

$$0 \le b_1 = \alpha \mu_n \left(\nabla u(x_{1/2})\right) \bar{n}(a) + \beta(\mu_n \nabla u)(x_1)\bar{n}(a)$$

is the first component of B, and subsequent components b_j satisfy $b_j = 0$, $2 \le j \le M-1$, until

$$b_M = \alpha \mu_n \left(\nabla u(x_{M+1/2})\right) \bar{n}(b) + \beta(\mu_n \nabla u)(x_M)\bar{n}(b) \ge 0,$$

if $\nabla u(x_M) \le 0$. The second terms in b_1 and b_M are zero if $\nabla u(x_1) \le 0$, or $\nabla u(x_M) \ge 0$, respectively. Here we have used the notation $\alpha = \Delta t_k / h^2$ and $\beta = \Delta t_k / h$.

(ii) L^h has the property that its diagonal components are positive, and its off diagonal nonpositive. By the choice of discretization (cf. (49)–(53)), it follows that L^h is weakly diagonally dominant; in fact, the upwinding has been selected to preserve the weak diagonal dominance of the box method applied to $-\nabla(\mu_n \nabla n)$. It follows (cf. [22]) that $\theta I + L^h$ is a diagonally dominant M-matrix for every $\theta > 0$, and, in particular, that $\theta I + L^h$ is invertible, with $(\theta I + L^h)^{-1} \ge 0$. It is also true that $\|(I + L^h)^{-1}\|_{\ell^\infty, \ell^\infty} \le 1$, a fact which is less familiar, and therefore proved in Lemma 2, at the conclusion of (II 3). In fact, a more general result is proved.

In the light of properties (i) and (ii), we establish the inequality $n^h \ge 0$ and, up to a set Ω^h, $|\Omega^h| \to 0$ as $h \to 0$, $n^h \le n_k^{max}$. The desired inequality of (48) is achieved in the limit, as $h \to 0$, as we later indicate. The set Ω^h is concentrated about the finite number of points where N changes sign. A remark is in order, which will be used in both cases, i.e., if D is a nonnegative diagonal matrix, and

$$(\theta I + D + L^h)y = w, \tag{55}$$

$$(\theta I + L^h)z = w, \tag{56}$$

where $w \ge 0$ is assumed, as well as the property that $\theta I + L^h$ is a diagonally dominant M-matrix, then

$$0 \le y \le z. \tag{57}$$

If D is simply diagonal, then the lower bound holds when $\theta I + D + L^h$ is a diagonally dominant M-matrix. The lower bound is well known (see Berman and Plemmons [5]). One verifies the upper bound in (57) by simple subtraction and inversion of (55) and (56). We now proceed to verify $n^h \ge 0$. We distinguish two cases. For j, such that $n_j^h > 0$, we have

$$-[\tilde{U}_n(\cdot, n^h, \tilde{p})]_j \ge -\left(\frac{2\mu_{0n}}{\epsilon}\right)\left(\frac{f(x_j, n_j^h)}{n_j^h} + \|N\|_{L^\infty}\right) n_j^h,$$

while for $n_j^h \le 0$, (43)–(46) hold. If

$$D \stackrel{\text{def}}{=} \text{diag}[\delta_j],$$

where

$$\delta_j \stackrel{\text{def}}{=} \begin{cases} 0, & n_j^h \le 0, \\ \Delta t_k \left[2(\epsilon n_j^h)^{-1}\mu_{0n} f(x_j, n_j^h) + \sigma\right], & n_j^h > 0, \end{cases}$$

then (54) and the above yield

$$(\theta I + D + L^h)n^h \ge n_{k-1} + B \ge 0,$$

for

$$\theta \stackrel{\text{def}}{=} 1 - (2\epsilon^{-1}\mu_{0n}p^{max} + \sigma)\Delta t_k.$$

Implicit in this display is that $\theta > 0$; however, if Δt_k satisfies (62) to follow, then $\theta > 0$ does hold. Upon use of (55) and (57), and the M-matrix property of $\theta I + L^h$, we deduce that $n^h \ge 0$. In a similar way, $p^h \ge 0$ and the time-step restriction is also implied by (62).

The proof that the jth component of n^h does not exceed $n_k^{max}\ (x_j)$, except in a thin set Ω^h, begins with the vector equation

$$\begin{aligned} &(I + L^h)[n^h - n_{k-1}^{max} + \text{Step}(N^+)] \\ &\quad = \left(n_{k-1} - n_{k-1}^{max}\right) + \left(B - L^h[n_{k-1}^{max} - \text{Step}(N^+)]\right) \\ &\qquad + \text{Step}(N^+) - \Delta t_k \tilde{U}_n(\cdot, n^h, \tilde{p}). \end{aligned} \tag{58}$$

For j, such that $n_j^h < n_k^{max}(x_j)$, we have, upon omission of $\tilde{p}(x_j)$ and $N^+(x_j)$, for some $0 < c < 1$,

$$\begin{aligned} -[\tilde{U}_n(\cdot, n^h, \tilde{p})]_j \quad \le \quad & -c\epsilon^{-1}\mu_{0n}\left\{\left[\left(n_j^h - n_{k-1}^{max}(x_j)\right) + \left(n_{k-1}^{max}(x_j) - n_k^{max}(x_j)\right)\right]\right. \\ & \cdot \left(n_j^h + n_k^{max}(x_j) + N^-(x_j)\right) + \left(f(\cdot, n_k^{max}) - n_k^{max}N^+\right)(x_j) \\ & \left. - \left(n_k^{max}\tilde{p} - n_k^{max}N^-\right)(x_j)\right\}, \end{aligned}$$

while for j, such that $n_j^h \ge n_k^{max}(x_j)$, we have (43)–(46). If

$$D \stackrel{\text{def}}{=} \text{diag}[\delta_j],$$

where

$$\delta_j \stackrel{\text{def}}{=} \begin{cases} -\Delta t_k \epsilon^{-1}\mu_{0n}\|N\|_{L^\infty}, & n_j^h \ge n_k^{max}(x_j), \\ \Delta t_k \epsilon^{-1}\mu_{0n}\left[c\left(n_j^h + n_k^{max}(x_j) + N^-(x_j)\right) - \|N\|_{L^\infty}\right], & n_j^h < n_k^{max}(x_j), \end{cases}$$

then we obtain from (58), and the above, the estimate, where, by (62), $\delta_j \ge -1/2$,

$$\begin{aligned} &(I + D + L^h)[n^h - n_{k-1}^{max} + \text{Step}(N^+)] \\ &\quad \le B - L^h[n_{k-1}^{max} - \text{Step}(N^+)] + (I + D)\,\text{Step}(N^+) + \text{col}(\rho)\Delta t_k \\ &\qquad + 2\epsilon^{-1}\mu_{0n}(n_k^{max} - n_{k-1}^{max})(n^{max} + \sup N^-)\Delta t_k. \end{aligned} \tag{59}$$

The final two terms in (59) result from distinguishing the cases $n_j^h \geq n_k^{max}(x_j)$ and $n_j^h < n_k^{max}(x_j)$, respectively. Note that we have also been able to employ the induction hypothesis and the inequality,

$$\left(f(\cdot, n_k^{max}) - n_k^{max}\,\mathrm{Step}(N^+)\right)(x_j) \geq \left(n_k^{max}\tilde{p} - n_k^{max}\,\mathrm{Step}(N^-)\right)(x_j),$$

in deriving (59) from (58). In order to treat the term involving B, we note that, by design,

$$n_{k-1}^{max} - \mathrm{Step}(N^+) = \mathrm{col}(c),\ c = c(k), \tag{60}$$

for a constant $c \geq \lambda_0$, where λ_0 is defined in (16). Now the composite difference scheme annihilates constants, in the sense that

$$L^h\,\mathrm{col}(c) = \tilde{B},$$

where $\tilde{B}$ is formed from B by replacing the boundary values by c. It follows from (19) and (60), since $\tilde{B} \geq B$, that

$$\begin{aligned} n^h - n_{k-1}^{max} + \mathrm{Step}(N^+) \quad \leq \quad & Q^h \Delta t_k [I + 8\epsilon^{-1}\mu_{0n}\Delta t_k (n^{max} + \sup N^-)]\,\mathrm{col}(\rho) \\ & + Q^h (I + D)\,\mathrm{Step}(N^+), \end{aligned}$$

where we used (16)–(21), and where $Q^h \stackrel{\mathrm{def}}{=} (I + D + L^h)^{-1} \geq 0$. This expression can be simplified by use of Lemma 2. Indeed, we obtain

$$\begin{aligned} n^h - n_{k-1}^{max} + \mathrm{Step}(N^+) \quad \leq \quad & 2\Delta t_k [I + 8\epsilon^{-1}\mu_{0n}\Delta t_k (n^{max} + \sup N^-)]\,\mathrm{col}(\rho) \\ & + Q^h (I + D)\,\mathrm{Step}(N^+). \end{aligned}$$

After same simplification, we obtain

$$n^h \leq n_{k-1}^{max} - \mathrm{Step}(N^+) + Q^h (I + D)\,\mathrm{Step}(N^+) + 4\Delta t_k\,\mathrm{col}(\rho), \tag{61}$$

provided

$$8\epsilon^{-1}\Delta t_k \max(\mu_{0n}, \mu_{0p})[\max(n^{max}, p^{max}) + \|N\|_{L^\infty}] \leq 1. \tag{62}$$

Note that (62) also accommodates the hole equation analysis. Now, Lemma 2 implies that

$$\|Q^h (I + D)\,\mathrm{Step}(N^+)\|_{\ell^\infty} \leq \sup N^+,$$

and a direct analysis shows that, for a string of zeros in $\mathrm{Step}(N^+)$, the corresponding positions in $Q^h(I + D)\,\mathrm{Step}(N^+)$ are zero, except possibly for the positions of the first and last zeros. Thus, Ω^h is defined as the union of intervals, specified by the partition points on either side of the (finitely many) points where N changes sign. This analysis shows that, for $x_j \notin \Omega^h$,

$$n^h(x_j) \leq n_{k-1}^{max}(x_j) + 4\rho\Delta t_k = n_k^{max}(x_j).$$

In particular, the upper bound has been demonstrated for the discrete solutions, except on Ω^h, since a similar argument, and result, hold for p^h, with a possible enlargement of Ω^h.

If we knew that solutions of (29) and (30) were smooth, then well-known arguments associated with the convergence of the box method, augmented by a straightforward analysis of the upwinding approximation, would give the required bounds (48), and would permit $\tilde{U}_n$ and $\tilde{U}_p$ to be replaced by U_n and U_p in (29) and (30). Since C^3 solution regularity is sufficient for $O(h)$ convergence, we may assume that the result holds for this class. To achieve this, we could mollify the nonnegative L^2 functions $\tilde{n}$ and $\tilde{p}$ to obtain smooth nonnegative functions for which the corresponding C^3 solutions of (29) and (30) are nonnegative. However, these functions need not satisfy the upper bounds. Instead, we define, for $\delta > 0$,

$$n_\delta(x) = \begin{cases} \tilde{n}(x), & x \notin \Omega^\delta, \\ 0, & x \in \Omega^\delta, \end{cases}$$

where Ω^δ is a set of measure proportional to δ and centered about the points where N changes sign. On each of the complementary subintervals, n_k^{max} reduces to a constant upper bound, though the constant itself changes from interval to interval. The individual function components of n_δ can be mollified, with the support of the mollified components contained within the subinterval, in such a way that the bounds are satisfied, i.e., the mollified pair is in K. In this way, one obtains an L^2 convergent sequence of approximations to $\tilde{n}$ (and $\tilde{p}$). It is a standard result, and, in fact, follows from a slight generalization of the analysis of (III) to follow, that this yields L^2 convergence of the solutions of the mollified problems to the fixed solutions of (29) and (30). This completes the verification of parts (1), (2), and (3) of (II), except for Lemma 2, which we present now.

Lemma 2 *Let L^h be the weakly diagonally dominant matrix of (54), with positive diagonal entries, and nonpositive off-diagonal entries. Let D be any diagonal matrix, $D = \mathrm{diag}[\delta_j]$, with $\delta_j > -1$. Then*

$$\|(I + D + L^h)^{-1}(I + D)\|_{\ell^\infty, \ell^\infty} \leq 1. \tag{63}$$

If $\delta_j \geq -1/2$, then

$$\|(I + D + L^h)^{-1}\|_{\ell^\infty, \ell^\infty} \leq 2. \tag{64}$$

Proof: Let y and z satisfy

$$(I + D + L^h)y = (I + D)z,$$

and let j denote an index for which $|y_j| = \|y\|_{\ell^\infty}$. Without loss of generality we assume $y_j \geq 0$. Then,

$$(1 + \delta_j)y_j \leq (1 + \delta_j)z_j + y_j \left(-l_{jj} + \sum_{m \neq j} l_{jm} \right),$$

where we have denoted the diagonal elements of L^h by l_{jj} and the off-diagonal elements by $-l_{jm}$. Since weak diagonal dominance assures that

$$-l_{jj} + \sum_{m \neq j} l_{jm} \leq 0,$$

we conclude that

$$0 \le y_j \le z_j \le \|z\|_{\ell^\infty}.$$

This establishes the first statement of the lemma, since 'a priori' considerations assure that $(I + D + L^h)^{-1}$ exists. The proof of the second statement is similar. ∎

(III) We shall directly estimate the Lipschitz constant of $T : K \to K$. Thus, let

$$u_1 = u_1[\tilde{n}_1, \tilde{p}_1],\ u_2 = u_2[\tilde{n}_2, \tilde{p}_2]. \tag{65}$$

The first step consists of the estimate,

$$\begin{aligned} \|u_1 - u_2\|_{L^2}^2 &\le \left(\frac{b-a}{\pi}\right)^2 \|u_1 - u_2\|_{H^1}^2 \\ &\le \left(\frac{2(b-a)^4}{\pi^4 \epsilon}\right) \|[\tilde{n}_1 - \tilde{n}_2, \tilde{p}_1 - \tilde{p}_2]\|_{L^2 \times L^2}^2 , \end{aligned} \tag{66}$$

which follows upon subtracting the respective formulations of (28) for $[\tilde{n}_1, \tilde{p}_1]$ and $[\tilde{n}_2, \tilde{p}_2]$, using $u_1 - u_2$ as a test function in the weak version, estimating inner products in terms of sums of squares, and, finally, estimating the L^2 norm in terms of the H^1 norm. The next step uses a variant in which (29) and (30) are collapsed back so that $\nabla \cdot (\mu_n \nabla u\, n)$ and $-\nabla \cdot (\mu_p \nabla u\, p)$ are the relevant drift terms. In this format, we subtract the relevant equations involving n_1 and n_2 as well as p_1 and p_2 if

$$[n_1, p_1] = T[\tilde{n}_1, \tilde{p}_1],\ [n_2, p_2] = T[\tilde{n}_2, \tilde{p}_2]. \tag{67}$$

Using the monotonicity of the recombination terms, in each variable separately, and using $n_1 - n_2$, $p_1 - p_2$ as test functions, respectively, we obtain, for $\lambda > 0$,

$$\begin{aligned} &\|n_1 - n_2\|_{L^2}^2 + \Delta t_k \left\| \sqrt{\mu_n(\nabla u_1)} \nabla (n_1 - n_2) \right\|_{L^2}^2 \\ &\quad \le \Delta t_k \left\{ \frac{\lambda}{2} \|\nabla n_2\|_{L^\infty}^2 \|\mu_n(\nabla u_1) - \mu_n(\nabla u_2)\|_{L^2}^2 + \frac{1}{2\lambda} \|\nabla (n_1 - n_2)\|_{L^2}^2 \right. \\ &\qquad + \frac{\lambda}{2} \|n_2\|_{L^\infty}^2 \|\mu_n(\nabla u_1)\nabla u_1 - \mu_n(\nabla u_2)\nabla u_2\|_{L^2}^2 + \frac{1}{2\lambda} \|\nabla (n_1 - n_2)\|_{L^2}^2 \\ &\qquad \left. + \frac{\lambda}{2} \|\mu_n(\nabla u_1)\nabla u_1\|_{L^\infty}^2 \|n_1 - n_2\|_{L^2}^2 + \frac{1}{2\lambda} \|\nabla (n_1 - n_2)\|_{L^2}^2 \right\}. \end{aligned} \tag{68}$$

A similar inequality holds for the p equation. The functions $\xi \mapsto \mu_n(\xi)$ and $\xi \mapsto \xi \mu_n(\xi)$ are Lipschitz continuous, with constants μ_{0n}^2 / v_{sn} and $2\mu_{0n}$ so that, combined with the 'a priori' estimates for ∇u_1, n_2, and ∇n_2 this yields, for $\lambda = 2/\inf \mu_n(\nabla u_1)$,

$$\|n_1 - n_2\|_{L^2}^2 \le \Delta t_k \left[\left(\frac{v_{sn}^2}{\inf \mu_n} \right) \|n_1 - n_2\|_{L^2}^2 + C \|u_1 - u_2\|_{H^1}^2 \right],$$

where C does not depend on $\tilde{n}_1$, $\tilde{p}_1$, $\tilde{n}_2$, $\tilde{p}_2$. A similar inequality holds for $\|p_1 - p_2\|_{L^2}$. These computations show that T is Lipschitz continuous if

$$\Delta t_k < \frac{\min(\inf \mu_n, \inf \mu_p)}{\max(v_{sn}^2, v_{sp}^2)}. \tag{69}$$

Rather than make an unreasonable further restriction to guarantee that T is a strict contraction, we may apply the Schauder fixed-point theorem to T to conclude the argument. The necessary H^1 bounds are obtained from the second term on the left-hand side of (68). Uniqueness, however, is not obvious without the further restriction that T be a strict contraction.

3.3 Approximate Solvability and L^2 Residual Control

A typical computing procedure involves solving the system (11)–(13) only approximately. In the next major result, we present and analyze a criterion for approximate solvability in terms of the residuals of the individual systems at successive time steps. The computed triple, $[u_k, n_k, p_k]$, need not satisfy the invariant-region property as presented in Theorem 1. Since pointwise bounds are essential to the theory, they are built into the hypothesis structure directly at the outset.

Theorem 3 *Suppose that the triple $[u_k, n_k, p_k]$ approximately solves (11)–(13), i.e., the boundary conditions (23) are satisfied exactly and*

$$-\epsilon\nabla^2 u_k + n_k - p_k - N = r_1, \tag{70}$$

$$\frac{n_k - n_{k-1}}{\Delta t_k} - \nabla\cdot(\mu_n\nabla n_k) + \mu_n\nabla u_k\nabla n_k + U_{n,k} = r_2, \tag{71}$$

$$\frac{p_k - p_{k-1}}{\Delta t_k} - \nabla\cdot(\mu_p\nabla p_k) - \mu_p\nabla u_k\nabla p_k + U_{p,k} = r_3, \tag{72}$$

where the residuals r_1, r_2, and r_3 satisfy

$$\|r_1\|_{L^2}^2 \le C_u(\Delta t_k)^2,\ \|r_2\|_{L^2} \le C_n\Delta t_k,\ \|r_3\|_{L^2} \le C_p\Delta t_k, \tag{73}$$

for $k = 1, \ldots, L$. Define the errors,

$$v_k \stackrel{\text{def}}{=} u(\cdot, t_k) - u_k, \tag{74}$$

$$e_k \stackrel{\text{def}}{=} n(\cdot, t_k) - n_k, \tag{75}$$

$$q_k \stackrel{\text{def}}{=} p(\cdot, t_k) - p_k. \tag{76}$$

Suppose that pointwise bounds,

$$\|u_k\|_{L^\infty} \le c_u,\ \|n_k\|_{L^\infty} \le c_n,\ \|p_k\|_{L^\infty} \le c_p, \tag{77}$$

exist for the computed semidiscrete approximations, together with the lower bounds,

$$1 + n_k \ge 0,\ 1 + p_k \ge 0. \tag{78}$$

Define

$$\Delta t \stackrel{\text{def}}{=} \max_k \Delta t_k. \tag{79}$$

If $1 - c_1\Delta t > 0$ with c_1 specified in (96) below, then these errors converge with optimal order in L^2:

$$\sup_{1\le M\le L}(\|e_M\|_{L^2}^2 + \|q_M\|_{L^2}^2) \tag{80}$$
$$+\min(\inf\mu_n, \inf\mu_p)\sum_{k=1}^{L}(\|\nabla e_k\|_{L^2}^2 + \|\nabla q_k\|_{L^2}^2)\Delta t_k \le C_1(\Delta t)^2,$$
$$\sup_{1\le M\le L}\|\nabla v_M\|_{L^2}^2 \le C_2(\Delta t)^2, \tag{81}$$

where C_1 and C_2 are certain positive constants, given explicitly in (98) and (99) below. The lower bounds, (78), are unnecessary if $R \equiv 0$ is selected. The result remains valid if n^2 is replaced by $(n_+)^2$ and p^2 by $(p_+)^2$ in $U_{n,k}$ and $U_{p,k}$, respectively.

Proof: Subtract (71) from the second equation, (2), of the device system to obtain

$$\begin{aligned}(\Delta t_k)^{-1}(e_k - e_{k-1}) &- \nabla\left(\mu_n(\nabla u_k)\nabla e_k\right)\\ &+\nabla\left(\left[\mu_n(\nabla u_k) - \mu_n\left(\nabla u(\cdot,t_k)\right)\right]\nabla n(\cdot,t_k)\right) + \mu_n(\nabla u_k)\nabla u_k\nabla e_k\\ &-\left[\mu_n(\nabla u_k)\nabla u_k - \mu_n\left(\nabla u(\cdot,t_k)\nabla u(\cdot,t_k)\right)\right]\nabla n(\cdot,t_k)\\ &= -r_2 + V_{n,k} + (\Delta t_k)^{-1}\int_{t_{k-1}}^{t_k}(t_{k-1}-s)\frac{\partial^2 n}{\partial t^2}(s)\,ds,\end{aligned} \tag{82}$$

where

$$V_{n,k} = -[U_n\left(u(\cdot,t_k), n(\cdot,t_k), p(\cdot,t_k)\right) - U_{n,k}(u_k, n_k, p_k)]. \tag{83}$$

In the estimation to follow, we shall write

$$\begin{aligned}V_{n,k} \;=\; &[U_n(u,n,p_k) - U_n(u,n,p)] - [R(n,p_k) - R(n_k,p_k)]\\ &-\epsilon^{-1}\left[(\mu_n'(\nabla u)\nabla u + \mu_n(\nabla u)) - (\mu_n'(\nabla u_k)\nabla u_k + \mu_n(\nabla u_k))\right]\\ &\quad\cdot[n^2 - p_k n - Nn].\\ &-\epsilon^{-1}[\mu_n'(\nabla u_k)\nabla u_k + \mu_n(\nabla u_k)][(n^2 - n_k^2) - p_k(n-n_k) - N(n-n_k)]\end{aligned} \tag{84}$$

Multiplication of (82) by $e_k\Delta t_k$, integration over Ω, and summation on $k = 1,\dots,M$ for $1 \le M \le L$ yield, after some simplification,

$$\begin{aligned}&\frac{1}{2}\|e_M\|_{L^2}^2 + \inf\mu_n\sum_{k=1}^{M}\|\nabla e_k\|_{L^2}^2\Delta t_k\\ &\quad\le\left\{\left(\frac{\mu_{0n}^2}{v_{sn}}\right)\|\nabla n\|_{L^\infty}\sum_{k=1}^{M}\|\nabla v_k\|_{L^2}\|\nabla e_k\|_{L^2} + v_{sn}\sum_{k=1}^{M}\|\nabla e_k\|_{L^2}\|e_k\|_{L^2}\right.\\ &\qquad\left.+2\mu_{0n}\|\nabla n\|_{L^\infty}\sum_{k=1}^{M}\|\nabla v_k\|_{L^2}\|e_k\|_{L^2} + C_n\sum_{k=1}^{M}\|e_k\|_{L^2}\Delta t_k\right.\end{aligned} \tag{85}$$

$$+\sum_{k=1}^{M}\|e_k\|_{L^2}\|f_k\|_{L^2}+C_{1n}\sum_{k=1}^{M}\|e_k\|_{L^2}^2$$
$$+C_{2n}\sum_{k=1}^{M}\|q_k\|_{L^2}\|e_k\|_{L^2}+C_{3n}\sum_{k=1}^{M}\|\nabla v_k\|_{L^2}\|e_k\|_{L^2}\Bigg\}\Delta t_k.$$

Here,

$$f_k=(\Delta t_k)^{-1}\int_{t_k-1}^{t_k}(t_{k-1}-s)\frac{\partial^2 n}{\partial t^2}(s)\,ds, \tag{86}$$

and we have noticed that $\xi\mapsto\mu_n(\xi)$, $\xi\mapsto\xi\mu_n(\xi)$, and $\xi\mapsto\xi\mu_n'(\xi)$ have respective Lipschitz constants of μ_{0n}^2/v_{sn}, $2\mu_{0n}$, and $3\mu_{0n}^2/v_{sn}$. Also we have used the monotonicity, in (84), of $R(\cdot,p_k)$ and of the quadratic in n. Finally, C_{1n}, C_{2n}, and C_{3n} are given here by

$$C_{1n} = \epsilon^{-1}\mu_{0n}(c_p+\|N\|_{L^\infty}), \tag{87}$$
$$C_{2n} = \frac{\tau_p+\tau_n}{\tau_p^2}+\epsilon^{-1}\mu_{0n}\|n\|_{L^\infty}, \tag{88}$$
$$C_{3n} = \left(\frac{4\mu_{0n}^2}{v_{sn}\epsilon}\right)\left[\|n\|_{L^\infty}^2+\|n\|_{L^\infty}(c_p+\|N\|_{L^\infty})\right]. \tag{89}$$

The first term defining C_{2n} embodies an upper bound for $\partial R/\partial p$, via the mean value theorem, and this was obtained by use of (78).

A similar inequality utilizing (72) and (3) holds for q_k. By use of (70) and (1) we deduce the important relation,

$$\|\nabla v_k\|_{L^2}^2\le\left(\frac{4(b-a)^2}{\pi^2\epsilon}\right)\left(\|e_k\|_{L^2}^2+\|q_k\|_{L^2}^2+\|r_1\|_{L^2}^2\right). \tag{90}$$

Note that we have employed here an elementary version of Rayleigh's inequality, and made judicious use of the inequality,

$$\alpha\beta\le\frac{1}{2}(\lambda\alpha^2+\lambda^{-1}\beta^2), \tag{91}$$

as applied to the estimated weak version of the subtracted equations. The term, $C_n\|e_k\|_{L^2}(\Delta t_k)^2$, is expanded as

$$C_n\|e_k\|_{L^2}(\Delta t_k)^2\le\frac{1}{2}C_n\|e_k\|_{L^2}^2\Delta t_k+\frac{1}{2}C_n(\Delta t_k)^3, \tag{92}$$

with a similar inequality for the C_p term. The estimate,

$$\|f_k\|_{L^2}^2\le\Delta t_k\left\|\frac{\partial^2 n}{\partial t^2}\right\|_{L^2(\Omega\times(t_{k-1},t_k))}^2, \tag{93}$$

is also used, with its counterpart involving p. The inequality (91) is employed to estimate the remaining terms. With proper choice of α and β, the choice, $\lambda=1$, is

made, except for those terms involving $\|\nabla e_k\|$, where $\beta = \|\nabla e_k\|_{L^2}$, $\lambda = 2/\inf \mu_n$ are appropriate choices as applied to the inequality (91). Similar choices hold for the companion inequality. The Gronwall inequality requires the absorption of the terms, $\|e_M\|^2_{L^2}$ and $\|q_M\|^2_{L^2}$, when arising on the right-hand side of the summed inequalities, to be transferred to the left-hand side. This imposes a restriction on Δt_k. Altogether, we have, upon adding the inequalities for e_k and q_k,

$$\|e_M\|^2_{L^2} + \|q_M\|^2_{L^2} + \min(\inf \mu_n, \inf \mu_p) \sum_{k=1}^{M} (\|\nabla e_k\|^2_{L^2} + \|\nabla q_k\|^2_{L^2}) \Delta t_k \tag{94}$$
$$\leq (1 - c_1 \sup \Delta t_k)^{-1} \left[c_0 (\Delta t)^2 + c_1 \sum_{k=1}^{M-1} (\|e_k\|^2_{L^2} + \|q_k\|^2_{L^2}) \Delta t_k \right],$$

where

$$c_0 = \kappa C_u + (C_n + C_p) T_0 + \|n_{tt}\|^2_{L^2(\Omega \times (0,T_0))} + \|p_{tt}\|^2_{L^2(\Omega \times (0,T_0))}, \tag{95}$$

$$c_1 = \kappa + 2 + 2 \max(C_{1n}, C_{1p}) + \max(C_n, C_p) + 2 \max(C_{2n}, C_{2p}) \tag{96}$$
$$+ \max(C_{3n}, C_{3p}) + \frac{2 \max(v^2_{sn}, v^2_{sp})}{\min(\inf \mu_n, \inf \mu_p)},$$

$$\kappa = \left(\frac{4(b-a)^2}{\pi^2 \epsilon} \right) \Big\{ C_{3n} + C_{3p} + 2(\|\nabla n\|^2_{L^\infty} + \|\nabla p\|^2_{L^\infty}) \tag{97}$$
$$\cdot \left[\frac{\max(\mu^4_{0n}/v^2_{sn}, \mu^4_{0p}/v^2_{sp})}{\min(\inf \mu_n, \inf \mu_p)} + 2 \max(\mu^2_{0n}, \mu^2_{0p}) \right] \Big\}.$$

Since the inequality (94) holds for each $M = 1, \ldots, L$, it is a consequence of the Gronwall inequality (cf. Jerome [11, pp. 52–54]) that (80) and (81) hold with

$$C_1 = \frac{c_0 \exp[c_1 T_0/(1 - c_1 \Delta t)]}{1 - c_1 \Delta t}, \tag{98}$$

$$C_2 = \left(\frac{4(b-a)^2}{\pi^2 \epsilon} \right) (C_u + C_1). \tag{99}$$

∎

Remark 1 *The advantage of the replacement of n^2 and p^2 by $(n_+)^2$ and $(p_+)^2$ is that it leads to uniform boundedness of approximate solutions (cf. Lemma 7).*

3.4 H^{-1} Residual Control

We shall develop a necessary parallel result for the case where the residual is estimated in H^{-1}. This has important implications for classes of algorithms, such as the transport diffusion algorithm, or those using second order numerical methods, which lead naturally to H^{-1} residual estimation. In order to provide as efficient

an estimation procedure as possible, we introduce the family of Riesz maps T_a, associated with the linear differential operators,

$$-\nabla a \nabla : H_0^1(\Omega) \to H^{-1}(\Omega). \tag{100}$$

T_a satisfies the relation, for $\ell \in H^{-1}(\Omega)$,

$$\langle \ell, v \rangle = (a \nabla T_a \ell, \nabla v)_{L^2}, \quad v \in H_0^1(\Omega). \tag{101}$$

Here, $a \geq a_0 > 0$ is an essentially bounded function and will be identified with μ_n or μ_p. We note the fundamental relation,

$$(f, T_a f)_{L^2} \geq \inf a (f, T_e f)_{L^2} \geq (\inf a / \sup a)(f, T_a f)_{L^2}, \tag{102}$$

where $e(x) \equiv 1$, which follows from (101). Here, we have interpreted the duality pairing $\langle \cdot, \cdot \rangle$ as the L^2 pivot space inner product for $f \in L^2(\Omega)$. We choose, for the (equivalent) H^{-1} norm,

$$\|\ell\|_{H^{-1}} = \langle \ell, T_e \ell \rangle^{1/2}. \tag{103}$$

The maps T_a, when restricted to $L^2(\Omega)$, are positive definite, self-adjoint, compact operators; they are also (pointwise) nonnegative operators. Moreover, if g is Lipschitz continuous,

$$\|g(v) - g(w)\|_{H^{-1}} \leq \|g\|_{Lip} \| |v - w| \|_{H^{-1}}, \quad v, w \in L^2(\Omega). \tag{104}$$

In the extension of Theorem 3 to the H^{-1} case, one is compelled to make systematic use of (101)-(104) and the properties just mentioned, together with their implications, as well as one additional result, contained in (106). To describe this, let

$$c_a \geq \sup a. \tag{105}$$

Now we have, for $f \in H^1(\Omega)$ and $g \in W^{1,\infty}(\Omega)$ with $\|g\|_{L^\infty}^2 + \|\nabla g\|_{L^\infty}^2 \leq \gamma^2$,

$$\|T_a^{1/2}(g \nabla f)\|_{L^2} \leq c_a \|T_e^{1/2}(g \nabla f)\|_{L^2} \leq C_a \|f\|_{L^2}, \tag{106}$$

as can be seen from the following computations. Since T_e is the Dirichlet solution operator for $-\nabla^2$,

$$T_e^{1/2}[g \nabla f(x)] = \sum_{m=1}^{\infty} \frac{a_m}{\sqrt{\lambda_m}} \sin\left(\frac{m\pi(x-a)}{b-a} \right),$$

where

$$a_m = \frac{2}{b-a} \int_a^b f'(s) g(s) \sin\left(\frac{m\pi(s-a)}{b-a} \right) ds$$

and $\lambda_m = [m\pi/(b-a)]^2$ denotes the mth eigenvalue of $-\nabla^2$. We have,

$$\begin{aligned} \left(\frac{b-a}{2} \right) a_m &= -\sqrt{\lambda_m} \int_a^b f(s) g(s) \cos\left(\frac{m\pi(s-a)}{b-a} \right) ds \\ &\quad - \int_a^b f(s) g'(s) \sin\left(\frac{m\pi(s-a)}{b-a} \right) ds \\ &\stackrel{\text{def}}{=} \left(\frac{b-a}{2} \right) \left(-\sqrt{\lambda_m} b_m - c_m \right). \end{aligned}$$

Now, if N_0 denotes the Neumann operator for $-\nabla^2$, i.e.,

$$-\nabla^2 N_0 f = f - \frac{1}{b-a}\int_\Omega f,\ \nabla N_0 f(a) = \nabla N_0 f(b) = 0,\ \int_\Omega N_0 f = \int_\Omega f,$$

it follows from the above that

$$N_0\left(fg - \frac{1}{b-a}\int_\Omega fg\right) = \sum_{m=1}^{\infty}\frac{b_m}{\lambda_m}\cos\left(\frac{m\pi(x-a)}{b-a}\right),$$

hence that (cf. [11, pp. 17,18])

$$\begin{aligned}
\|T_e^{1/2}(g\nabla f)\|_{L^2}^2 &= \frac{b-a}{2}\sum_{m=1}^{\infty}\frac{a_m^2}{\lambda_m} \\
&\le (b-a)\left(\sum_{m=1}^{\infty} b_m^2 + \sum_{m=1}^{\infty}\frac{c_m^2}{\lambda_m}\right) \qquad (107)\\
&= 2\left\|fg - \frac{1}{b-a}\int_\Omega fg\right\|_{L^2}^2 + 2\|fg'\|_{H^{-1}}^2 \\
&\le 2\left(1 + \frac{(b-a)^2}{\pi^2}\right)\gamma^2\|f\|_{L^2}^2.
\end{aligned}$$

Inequality (107) immediately yields (106) with $C_a = c_a\sqrt{2}[1+(b-a)^2/\pi^2]^{1/2}\gamma$.

Theorem 4 *Suppose the hypothesis, (73), is restated in terms of H^{-1} norms of the individual residuals:*

$$\|r_1\|_{H^{-1}}^2 \le C_u(\Delta t_k)^2,\ \|r_2\|_{H^{-1}} \le C_n\Delta t_k,\ \|r_3\|_{H^{-1}} \le C_p\Delta t_k. \qquad (108)$$

Suppose that the pointwise bounds (77) and (78) hold and that r_1 is uniformly pointwise bounded. Then

$$\sup_{1\le M\le L}(\|e_M\|_{H^{-1}}^2 + \|q_M\|_{H^{-1}}^2) + \sum_{k=1}^{L}(\|\nabla e_k\|_{L^2}^2 + \|\nabla q_k\|_{L^2}^2)\Delta t_k \le C_1(\Delta t)^2, \quad (109)$$

$$\sup_{1\le M\le L}\|v_M\|_{L^2}^2 \le C_2(\Delta t)^2, \qquad (110)$$

where C_1 and C_2 are given explicitly below in (116) and (117).

Proof: Apply T_{μ_n} to (82) and compute the L^2 inner product with $e_k\Delta t_k$. We shall make use of the properties of T_a described earlier in order to estimate the resultant. The corresponding replacement for (85) is given by

$$\frac{1}{2}(\inf\mu_n)\|e_M\|_{H^{-1}}^2 + \sum_{k=1}^{M}\|e_k\|_{L^2}^2\Delta t_k$$

$$
\begin{aligned}
\leq \Bigg[& C_{\mu_n}\left(\frac{\mu_{0n}^{5/2}}{v_{sn}}\right)\|\nabla n\|_{L^\infty}\sum_{k=1}^{M}\|\nabla v_k\|_{L^2}\|e_k\|_{H^{-1}} + C_{\mu_n}\sqrt{\mu_{0n}}\sum_{k=1}^{M}\|e_k\|_{L^2}\|e_k\|_{H^{-1}} \\
& + 2\mu_{0n}^2\left(\frac{b-a}{\pi}\right)\|\nabla n\|_{L^\infty}\sum_{k=1}^{M}\|\nabla v_k\|_{L^2}\|e_k\|_{H^{-1}} \qquad (111)\\
& + C_n\mu_{0n}\sum_{k=1}^{M}\|e_k\|_{H^{-1}}\Delta t_k + \mu_{0n}\sum_{k=1}^{M}\|e_k\|_{H^{-1}}\|f_k\|_{H^{-1}} \\
& + C'_{1n}\mu_{0n}\left(\frac{b-a}{\pi}\right)\sum_{k=1}^{M}\|e_k\|_{L^2}\|e_k\|_{H^{-1}} + C_{2n}\mu_{0n}\left(\frac{b-a}{\pi}\right)\sum_{k=1}^{M}\|q_k\|_{L^2}\|e_k\|_{H^{-1}} \\
& + C_{3n}\mu_{0n}\left(\frac{b-a}{\pi}\right)\sum_{k=1}^{M}\|\nabla v_k\|_{L^2}\|e_k\|_{H^{-1}}\Bigg]\Delta t_k
\end{aligned}
$$

with $C'_{1n} = C_{1n} + \|n\|_{L^\infty} + c_n$. The reader will note that $\gamma = 1$ in the first appearance of C_{μ_n} above, whereas $\gamma = (v_{sn}^2 + 4\mu_{0n}^2\omega^2)^{1/2}$ in the second appearance; here, ω is an upper bound for $\|\Delta u_k\|_{L^\infty}$. The replacement for (90) is

$$
\|\nabla v_k\|_{H^{-1}}^2 = \|v_k\|_{L^2}^2 \leq \left(\frac{4(b-a)^2}{\pi^2\epsilon}\right)\left(\|e_k\|_{H^{-1}}^2 + \|q_k\|_{H^{-1}}^2 + \|r_1\|_{H^{-1}}^2\right). \qquad (112)
$$

The remainder of the proof proceeds similarly to the proof of Theorem 3. The new choices of c_0 and c_1 are given by

$$
\begin{aligned}
c_0 &= \kappa C_u + T_0\bar{\mu}\Big(C_n + C_p + \|(T^{1/2}n)_{tt}\|_{L^2(\Omega\times(0,T_0))}^2 \qquad (113)\\
&\qquad + \|(T^{1/2}p)_{tt}\|_{L^2(\Omega\times(0,T_0))}^2\Big),\\
c_1 &= \kappa + \bar{\mu}^2\left[1 + 6\max(C_{1n}'^2, C_{1p}'^2) + 6\left(\frac{b-a}{\pi}\right)^2\max(C_{2n}^2, C_{2p}^2)\right] \qquad (114)\\
&\qquad + \bar{\mu}\left[\max(C_n, C_p) + 6\sup C_\mu^2\right],\\
\kappa &= \frac{3\pi^2\epsilon}{2(b-a)^2}\Bigg\{\left[\max\left(\frac{\mu_{0n}^5}{v_{sn}^2}, \frac{\mu_{0p}^5}{v_{sp}^2}\right)\sup C_\mu^2 + 4\bar{\mu}^4\left(\frac{b-a}{\pi}\right)^2\right] \qquad (115)\\
&\qquad \cdot\left(\|\nabla n\|_{L^\infty}^2 + \|\nabla p\|_{L^\infty}^2\right) + \bar{\mu}^2\left(\frac{b-a}{\pi}\right)^2\left(C_{3n}^2 + C_{3p}^2\right)\Bigg\}
\end{aligned}
$$

with $\bar{\mu} = \max(\mu_{0n}, \mu_{0p})$ and $C'_{1p} = C_{1p} + \|p\|_{L^\infty} + c_p$. In (114) and (115), $\sup C_\mu^2$ refers to all choices of $\mu = \mu_n, \mu_p$, obtained from the semidiscrete system. The finiteness of this follows, ultimately, from the pointwise boundedness assumptions.

The new choices of C_1 and C_2 are given by

$$
\begin{aligned}
C_1 &= \frac{c_0\exp\{c_1T_0/[\min(\inf\mu_n, \inf\mu_p) - c_1\Delta t]\}}{\min(\inf\mu_n, \inf\mu_p) - c_1\Delta t}, \qquad (116)\\
C_2 &= \left(\frac{4(b-a)^2}{\pi^2\epsilon}\right)(C_u + C_1). \qquad (117)
\end{aligned}
$$

4 Approximate Newton Methods for the Semidiscrete System

We shall describe a rather general approximate Newton method, in which the approximate linear inversion is based upon an arbitrary inner iteration, arranged in such a way that the algorithm is quadratically convergent. Though the description is general, we shall aim for the development of an algorithm with a small number of outer iterations. The reader will recall from § 3 that it is the residual of the nonlinear map, defining the semidiscrete system, which is to be controlled of order Δt_k (cf. Theorem 3). It follows that the analysis of the approximate Newton method of this section will address this property, although, typically, quadratic convergence of the residuals occurs in tandem with quadratic convergence of the approximants themselves.

The plan of the section is as follows. An operator tableau for the linearized problem is presented in § 4.1, while the properties required of an approximate Newton method are presented in § 4.2. A study of these properties, as they affect the semiconductor model, is presented next (§ 4.3 and § 4.4), followed by a description of continuation from the previous discrete time in § 4.5.

4.1 The Linearized Problem

This is primarily a formal subsection, in which we briefly present the operator for the linearized problem. At the conclusion, we mention considerations of operator domain and range.

Suppose the system (11)–(13) is denoted by

$$F(u_k, n_k, p_k) = 0, \tag{118}$$

and the mth linear increment, i.e., the difference between the mth and the $(m-1)$th Newton iterates, is denoted by $[\phi_k^m, \psi_k^m, \omega_k^m]$. This increment satisfies

$$\begin{bmatrix} -\epsilon\nabla^2\star & \star & -\star \\ (F')_{21} & (F')_{22} & (\partial U_p/\partial p)\star \\ (F')_{31} & (\partial U_n/\partial n)\star & (F')_{33} \end{bmatrix} \begin{pmatrix} \phi_k^m \\ \psi_k^m \\ \omega_k^m \end{pmatrix} = - \begin{pmatrix} r_1^{m-1} \\ r_2^{m-1} \\ r_3^{m-1} \end{pmatrix} \tag{119}$$

where

$$(F')_{21} \stackrel{\text{def}}{=} -\nabla\left[\nabla n_k^{m-1}\mu_n'(\nabla u_k^{m-1})\nabla\star\right] + \nabla n_k^{m-1}\nabla u_k^{m-1}\mu_n'(\nabla u_k^{m-1})\nabla\star \tag{120}$$
$$+\nabla n_k^{m-1}\mu_n(\nabla u_k^{m-1})\nabla\star + \frac{\partial U_n}{\partial u}\nabla\star$$

$$(F')_{22} \stackrel{\text{def}}{=} \frac{\star}{\Delta t_k} - \nabla\left[\mu_n(\nabla u_k^{m-1})\nabla\star\right] + \mu_n(\nabla u_k^{m-1})\nabla u_k^{m-1}\nabla\star + \frac{\partial U_n}{\partial n}\star \tag{121}$$

$$(F')_{31} \stackrel{\text{def}}{=} -\nabla\left[\nabla p_k^{m-1}\mu_p'(\nabla u_k^{m-1})\nabla\star\right] - \nabla p_k^{m-1}\nabla u_k^{m-1}\mu_p'(\nabla u_k^{m-1})\nabla\star \tag{122}$$

$$-\nabla p_k^{m-1}\mu_p(\nabla u_k^{m-1})\nabla\star+\frac{\partial U_p}{\partial u}\nabla\star$$

$$(F')_{33} \stackrel{\text{def}}{=} \frac{\star}{\Delta t_k}-\nabla\left[\mu_p(\nabla u_k^{m-1})\nabla\star\right]-\mu_p(\nabla u_k^{m-1})\nabla u_k^{m-1}\nabla\star+\frac{\partial U_p}{\partial p}\star. \quad (123)$$

Moreover, r_1^{m-1}, r_2^{m-1}, and r_3^{m-1} are the residuals of $F(u_k^{m-1}, n_k^{m-1}, p_k^{m-1})$, given explicitly in (11)–(13); $[u_k^{m-1}, n_k^{m-1}, p_k^{m-1}]$ is the Newton iterate; and,

$$\begin{aligned}
\frac{\partial U_n}{\partial u} &\stackrel{\text{def}}{=} \frac{\partial U_n}{\partial u}(u_k^{m-1}, n_k^{m-1}, p_k^{m-1}) \\
&= \epsilon^{-1}[2\mu_n'(\nabla u_k^{m-1})+\mu_n''(\nabla u_k^{m-1})\nabla u_k^{m-1}] \qquad (124)\\
&\quad \cdot[(n_k^{m-1})^2 - n_k^{m-1}p_k^{m-1} - n_k^{m-1}N] \\
\frac{\partial U_p}{\partial u} &\stackrel{\text{def}}{=} \frac{\partial U_p}{\partial u}(u_k^{m-1}, n_k^{m-1}, p_k^{m-1}) \\
&= -\epsilon^{-1}[2\mu_p'(\nabla u_k^{m-1})+\mu_p''(\nabla u_k^{m-1})\nabla u_k^{m-1}] \\
&\quad \cdot[n_k^{m-1}p_k^{m-1} - (p_k^{m-1})^2 - p_k^{m-1}N] \qquad (125)\\
\frac{\partial U_n}{\partial n} &\stackrel{\text{def}}{=} \frac{\partial U_n}{\partial n}(u_k^{m-1}, n_k^{m-1}, p_k^{m-1}) \\
&= \frac{(\tau_p + p_k^{m-1}\tau_n)(p_k^{m-1}+1)}{(d_k^{m-1})^2} \qquad (126)\\
&\quad +\frac{1}{\epsilon}[\mu_n(\nabla u_k^{m-1})+\mu_n'(\nabla u_k^{m-1})\nabla u_k^{m-1}](2n_k^{m-1}-p_k^{m-1}-N) \\
\frac{\partial U_p}{\partial p} &\stackrel{\text{def}}{=} \frac{\partial U_p}{\partial p}(u_k^{m-1}, n_k^{m-1}, p_k^{m-1}) \\
&= \frac{(\tau_n + n_k^{m-1}\tau_p)(n_k^{m-1}+1)}{(d_k^{m-1})^2} \qquad (127)\\
&\quad -\frac{1}{\epsilon}[\mu_p(\nabla u_k^{m-1})+\mu_p'(\nabla u_k^{m-1})\nabla u_k^{m-1}](n_k^{m-1}-2p_k^{m-1}-N)
\end{aligned}$$

with

$$d_k^{m-1} \stackrel{\text{def}}{=} \tau_p(n_k^{m-1}+1)+\tau_n(p_k^{m-1}+1). \quad (128)$$

We shall find it advantageous to consider two distinct application frameworks for the mappings F and F'. Note that (119)–(123) is a delineation of the equation,

$$F'(u_k^{m-1}, n_k^{m-1}, p_k^{m-1})\begin{pmatrix}\phi_k^m\\ \psi_k^m\\ \omega_k^m\end{pmatrix} = -\begin{pmatrix}r_1^{m-1}\\ r_2^{m-1}\\ r_3^{m-1}\end{pmatrix}. \quad (129)$$

(Also note that (119) represents a considerable notational economy; a similar tableau was introduced by Bank et al. [4].) We shall permit both an L^2 and H^{-1} residual measurement framework because of the variety of applications. Thus, F will be defined from triples in the affine subspace of $H^s(\Omega)$ given by

$$X_s = [\bar{u}(\cdot, t_k), \bar{n}, \bar{p}] + Y_s, \quad s = 1, 2, \quad (130)$$

where

$$Y_s = \prod_1^3 H^s(\Omega) \cap H_0^1(\Omega), \quad s = 1, 2, \tag{131}$$

to the triple product of copies of $H^{s-2}(\Omega)$. $F'(z)$, for fixed z, is defined from Y_s to the same space. Any approximate inverse, denoted $G(z)$ in the next subsection, reverses domain and range. Finally, we shall find it convenient, and, in fact, necessary to restrict F to have local domain of definition.

4.2 A Class of Approximate Newton Methods

We begin by recalling the three fundamental properties required of an approximate Newton method, based on a family of linear maps, $G(z^m)$, where $\{z^m\}$ comprises the Newton sequence, defined by

$$z^m - z^{m-1} \stackrel{\text{def}}{=} -G(z^{m-1})F(z^{m-1}), \tag{132}$$

and where z, $F(z) = 0$, is sought as $z = \lim_{m\to\infty} z^m$. The three properties are:

$$\|G(z^{m-1})F(z^{m-1})\| \le M_1\|F(z^{m-1})\|, \tag{133}$$

$$\|[F'(z^{m-1})G(z^{m-1}) - I]F(z^{m-1})\| \le M_2\|F(z^{m-1})\|^2, \tag{134}$$

$$\|F'(x) - F'(y)\| \le 2M_3\|x - y\|, \tag{135}$$

together with a mechanism insuring that successive iterates lie within the domain of definition of F. Here, the constants M_1, M_2, and M_3 are independent of the elements in the domain of definition of the respective maps. This framework is reminiscent of that described by Bank and Rose [3].

In order to estimate $\|F(z^m)\|$, we write

$$F(z^m) = -[F'(z^{m-1})G(z^{m-1}) - I]F(z^{m-1}) + R(z^{m-1}, z^m) \tag{136}$$

where

$$R(z^{m-1}, z^m) = F(z^m) - F(z^{m-1}) - F'(z^{m-1})(z^m - z^{m-1}). \tag{137}$$

Notice that the first term in (136) may be dominated by $M_2\|F(z^{m-1})\|^2$, via (134), and the second term by

$$\|R(z^{m-1}, z^m)\| = \left\| \int_0^1 \left[F'\left(z^{m-1} + s(z^m - z^{m-1})\right) - F'(z^{m-1})\right](z^m - z^{m-1})\, ds \right\|$$

on use of a standard Taylor expansion. From these remarks we have the estimate

$$\|F(z^m)\| \le (M_1^2 M_3 + M_2)\|F(z^{m-1})\|^2. \tag{138}$$

We have proved the bulk of the following lemma.

Lemma 5 *Let F be a map defined on a closed ball B_r in an affine subspace x_0+X_0, contained in a Banach space X, with Fréchet derivative*

$$F'(z) : X_0 \mapsto W, \quad z \in B_r.$$

Here W is a Banach space containing the range of F. Suppose that $G(z) : W \mapsto X_0$ is defined for each $z \in B_r$, and that F, F', and G satisfy (133), (134), and (135). If z^0 satisfies

$$M_1\|F(z^0)\| \leq (1-\alpha)r, \tag{139}$$

where $0 \leq \alpha < 1$ is such that $z^0 \in B_{\alpha r}$, then $z^1 \in B_r$ and the residual $F(z^1)$ satisfies

$$\|F(z^1)\| \leq (M_1^2 M_3 + M_2)\|F(z^0)\|^2. \tag{140}$$

Remark 2 *The hypothesis (135) can be weakened from $x, y \in B_r$ to x, y on the line segment between z^{m-1} and z^m. This remark is especially applicable to hypotheses (170) and (171) to follow, which need only hold along such line segments, and, in fact, can only be expected to hold in such a local fashion as induced by numerical approximation schemes.*

4.3 Estimation of the Constants

We shall estimate the constant M_3 of the previous section, and determine an estimate for M_1 in the special case $G = (F')^{-1}$. In the following section we shall show how this provides constant estimates for general G. As a preliminary comment, we recall again that the carrier approximations, unlike the solutions, can assume negative values. Some mechanism in any computing process must be found to ensure that the recombination term is well-defined. In practice, one might wish to redefine this term for negative values of n and p, when the term appears explicitly (as opposed to implicitly). In any event, we shall assume that the denominators d_k^{m-1} of the recombination term are bounded away from zero, independently of m and k:

$$d_k^{m-1} \geq d > 0. \tag{141}$$

Hypothesis (141) is an intrinsic hypothesis associated with the model. We now estimate $(F')^{-1}$.

Proposition 6 *Let F be the map of (118), defined on a ball of radius r_0 in X_s. The inversion map of (129) satisfies, for $m \leq m_0$,*

$$\|[F'(u_k^{m-1}, n_k^{m-1}, p_k^{m-1})]^{-1} r^{m-1}\|_{H^s} \leq M\|r^{m-1}\|_{H^{s-2}}, \tag{142}$$

where M is a constant depending only on r_0, m_0, and a bound for $\|\nabla^2 u_k^0\|_{L^\infty}$. For $s = 2$, there are no further unspecified hypotheses, whereas for $s = 1$ the 'a posteriori' hypotheses (157) and (158) of Lemma 7 are assumed to hold. Alternatively, in lieu of (157) and (158), the result holds if

$$\|\nabla n_k^m\|_{L^\infty} \leq C_{\|\mathcal{P}\|}\|\nabla n_k^m\|_{L^2}, \quad m \leq m_0, \tag{143}$$

with $C_{\|\mathcal{P}\|}^2\|\mathcal{P}\| \to 0$ as $\|\mathcal{P}\| \to 0$.

Proof: We begin with the case $s = 2$. From the tableau (119)–(123) we obtain the estimates

$$\begin{aligned}(\Delta t_k)^{-1}\|\psi_k^m\|_{L^2}^2 &+ \frac{1}{2}\inf\mu_n\|\nabla\psi_k^m\|_{L^2}^2 \\ &\le \frac{1}{2}\|r_2^{m-1}\|_{L^2}^2 + \left(\frac{A^2}{\inf\mu_n}\right)\|\nabla\phi_k^m\|_{L^2}^2 \\ &\quad + \left(\frac{5}{4} + \frac{v_{sn}^2}{\inf\mu_n} + \sup\left|\frac{\partial U_n}{\partial n}\right|\right)\|\psi_k^m\|_{L^2}^2 + \frac{1}{2}\sup\left|\frac{\partial U_p}{\partial p}\right|^2\|\omega_k^m\|_{L^2}^2\end{aligned} \tag{144}$$

and

$$\|\nabla\phi_k^m\|_{L^2}^2 \le \left(\frac{4(b-a)^2}{\pi^2\epsilon}\right)(\|\psi_k^m\|_{L^2}^2 + \|\omega_k^m\|_{L^2}^2 + \|r_1^{m-1}\|_{L^2}^2). \tag{145}$$

Here,

$$A = \mu_{0n}\left[\left(\frac{\mu_{0n}}{v_{sn}}\right)\sup|\nabla n_k^{m-1}| + \sqrt{\inf\mu_n}\left(2\sup|\nabla n_k^{m-1}| + \sup\left|\frac{\partial U_n}{\partial u}\right|\right)\right]. \tag{146}$$

In conjunction with the estimate for ω_k^m, we obtain, from (144) and (145),

$$\|\psi_k^m\|_{L^2}^2 + \|\omega_k^m\|_{L^2}^2 \le c\|r^{m-1}\|_{L^2}^2, \tag{147}$$

where c is the positive number defined by

$$\begin{aligned}c^{-1} \;=\; &\left(\frac{1}{2} + \frac{4(b-a)^2}{\pi^2\epsilon}\right)^{-1}\left\{1 - \|\mathcal{P}\|\left[\frac{4A^2(b-a)^2}{\pi^2\epsilon\min(\inf\mu_n, \inf\mu_p)}\right.\right. \\ &\left.\left. + \max\left(\sup\left|\frac{\partial U_n}{\partial n}\right|^2, \sup\left|\frac{\partial U_p}{\partial p}\right|^2\right) + \frac{9}{4} + \frac{\max(v_{sn}^2, v_{sp}^2)}{\min(\inf\mu_n, \inf\mu_p)}\right]\right\}.\end{aligned} \tag{148}$$

Clearly, $A \le A_0$ if the H^2 norm is restricted as specified.

The estimate (147) implies

$$\|\nabla^2\phi_k^m\|_{L^2}^2 \le 4\epsilon^{-2}(\|r_1^{m-1}\|_{L^2}^2 + c\|r^{m-1}\|_{L^2}^2), \tag{149}$$

which, in conjunction with

$$\|\phi_k^m\|_{L^2}^2 \le 4\epsilon^{-1}\left(\frac{b-a}{\pi}\right)^4(\|r_1^{m-1}\|_{L^2}^2 + c\|r^{m-1}\|_{L^2}^2), \tag{150}$$

yields an H^2 estimate for ϕ_k^m. The bootstrapping can now proceed. One views the second equation defined by (119)–(123) as having the form

$$(\Delta t_k)^{-1}\psi_k^m - \mu_n\nabla^2\psi_k^m - \mu_n'\nabla^2 u_k^{m-1}\nabla\psi_k^m + \mu_n\nabla u_k^{m-1}\nabla\psi_k^m = -r_2^{m-1} + q_k^{m-1}, \tag{151}$$

where

$$\|q_k^{m-1}\|_{L^2} \le C\|r^{m-1}\|_{L^2}. \tag{152}$$

The latter inequality follows from (149) and (150) and the tableau display (119)–(123). Note that the constant C depends only on r_0.

If (151) is multiplied by ψ_k^m, with the second and third terms collapsed into divergence form, then integration by parts yields the estimate,

$$\|\nabla\psi_k^m\|_{L^2}^2 \leq (1+C^2)(\inf\mu_n)^{-1}\|r^{m-1}\|_{L^2}^2, \tag{153}$$

provided

$$1 - \|\mathcal{P}\| - \frac{1}{2}\|\mathcal{P}\|\max(v_{sn}^2, v_{sp}^2)/\min(\inf\mu_n, \inf\mu_p) > 0. \tag{154}$$

Here we have used (152). For the estimate (153), there is a corresponding estimate for $\nabla\omega_k^m$.

If (151) is multiplied by $-\nabla^2\psi_k^m$ and the resultant integrated over Ω, one has

$$\begin{aligned}(\Delta t_k)^{-1}\|\nabla\psi_k^m\|_{L^2}^2 + \frac{1}{2}(\inf\mu_n)\|\nabla^2\psi_k^m\|_{L^2}^2 \qquad (155)\\ \leq 2(\inf\mu_n)^{-1}(\mu_{0n}^2\|\nabla^2 u_k^{m-1}\|_{L^\infty}^2 + v_{sn}^2)\|\nabla\psi_k^m\|_{L^2}^2 + 2(1+C^2)\|r^{m-1}\|_{L^2}^2.\end{aligned}$$

The conjunction of (153), (155), and the inequality

$$\|\nabla^2 u_k^{m-1}\|_{L^\infty} \leq \|\nabla^2 u_k^0\|_{L^\infty} + (m_0-1)\sup_m\|\phi_k^m\|_{L^\infty} \tag{156}$$

for $m \leq m_0$, leads to (142) in the case $s = 2$. Note that (149) and (150) yield a bound for the supremum in (156).

With one major exception, the proof for $s = 1$ proceeds as above, where the Riesz maps are employed as in § 3.4. The exception relates to uniform gradient bounds for n_k^m and p_k^m, which, in the case $s = 2$, are assured by the action of the mapping F on a bounded set in H^2. The lemma to follow provides these bounds, but specifies additional residual control necessary to achieve the bounds. This explains the residual hypothesis in the case $s = 1$. The alternative hypothesis, which bounds the gradient pointwise according to (143), does not require Lemma 7, and enters in (148) so that $c > 0$. In the interest of brevity, we omit the remaining details. ∎

Lemma 7 *Suppose that, for $u_k = u_k^m$, $n_k = n_k^m$, and $p_k = p_k^m$, the L^2-norm of the residual associated with (70)–(72), denoted $r_k = [r_{1,k}, r_{2,k}, r_{3,k}]$, is square summable bounded, independently of the partition $\mathcal{P}$ and the level of the Newton iteration:*

$$\sum_k \|r_k\|_{L^2}^2 \leq \alpha \tag{157}$$

and that the residuals are uniformly bounded in L^∞:

$$\|r_k\|_{L^\infty} \leq \beta. \tag{158}$$

Suppose that the quadratic terms n^2 and p^2 are replaced by $(n_+)^2$ and $(p_+)^2$ in $U_{n,k}$ and $U_{p,k}$, respectively. Then the solution triples are bounded in $\prod_1^3 H^2$:

$$\|[u_k, n_k, p_k]\|_{H^2} \leq C. \tag{159}$$

Proof: Although it appears natural to use the system (70)–(72) to derive these estimates, the nonlinear diffusion coefficients present technical obstacles to this approach. Somewhat surprisingly, it is advantageous to use a perturbation argument employing the solution. The latter does, however, require preliminary L^∞ estimates, for which it is possible to use (70)–(72). In order to obtain a format conducive to integration by parts, we subtract the boundary values. More precisely, set

$$\nu_k \stackrel{\text{def}}{=} n_k - \bar{n}, \tag{160}$$

so that

$$\frac{\nu_k - \nu_{k-1}}{\Delta t_k} - \nabla(\mu_n \nabla \nu_k) = \nabla(\mu_n \nabla \bar{n}) - \mu_n \nabla u_k \nabla \nu_k - \mu_n \nabla u_k \nabla \bar{n} - U_{n,k} + r_{2,k} \tag{161}$$

is the equivalent version of (71). If (161) is multiplied by $\nu_k^{\gamma-1}$, for γ an even integer, then an application of the inequality

$$\alpha\beta \le \frac{\alpha^\gamma}{\gamma} + \frac{\beta^{1-1/\gamma}}{1-1/\gamma}, \quad \alpha \ge 0, \beta \ge 0, \gamma > 0 \tag{162}$$

yields, after integration over Ω,

$$\begin{aligned}
&\int_\Omega |\nu_k|^\gamma + (\gamma-1)\Delta t_k \int_\Omega \mu_n |\nabla \nu_k|^2 \nu_k^{\gamma-2} \\
&\quad \le v_{sn} \Delta t_k \int_\Omega |\nabla \nu_k| \, |\nu_k|^{\gamma-1} \\
&\qquad + \int_\Omega \left| \left(\nabla(\mu_n \nabla \bar{n}) - \mu_n \nabla u_k \nabla \bar{n} - \hat{U}_{n,k} + r_{2,k} \right) \Delta t_k + \nu_{k-1} \right| |\nu_k|^{\gamma-1} \\
&\quad \le \frac{v_{sn}^2 \Delta t_k}{2 \inf \mu_n (\gamma-1)} \int_\Omega |\nu_k|^\gamma + \frac{\gamma-1}{2} \Delta t_k \int_\Omega \mu_n |\nabla \nu_k|^2 \nu_k^{\gamma-2} \\
&\qquad + \frac{1}{\gamma} \int_\Omega \left| \left(\nabla(\mu_n \nabla \bar{n}) - \mu_n \nabla u_k \nabla \bar{n} - \hat{U}_{n,k} + r_{2,k} \right) \Delta t_k + \nu_{k-1} \right|^\gamma \\
&\qquad + \frac{\gamma-1}{\gamma} \int_\Omega |\nu_k|^\gamma.
\end{aligned}$$

Here we have recognized that $U_{n,k}(u_k, \cdot, p_k)$, where $\cdot$ stands for $\nu_k + \bar{n}$, can be written as a sum of two terms, one of which is $\hat{U}_{n,k}$, and the other of which has the same sign as ν_k, $\hat{U}_{n,k}$ has the property that it is of linear growth in its final two arguments (cf. (164)); the other summand, when multiplied by $\nu_k^{\gamma-1}$, may be neglected. A parallel statement holds for $U_{p,k}$.

Multiplication by γ and extraction of γth roots leads to (cf. (154))

$$\begin{aligned}
2^{-1/\gamma} \|\nu_k\|_{L^\gamma} \quad \le \quad & \|\nu_{k-1}\|_{L^\gamma} + c\Delta t_k \left[\left(\frac{\mu_{0n}^2}{v_{sn}} \right) \|\nabla^2 u_k\|_{L^\gamma} + v_{sn} |\Omega|^{1/\gamma} \right] \\
& + \Delta t_k \left[\|\hat{U}_{n,k}\|_{L^\gamma} + \|r_{2,k}\|_{L^\gamma} \right],
\end{aligned} \tag{163}$$

where $c = |\nabla \bar{n}|$ is a constant. We may now let $\gamma \to \infty$. There is a parallel expression for $\varpi_k \stackrel{\text{def}}{=} p_k - \bar{p}$. The linear growth of $\hat{U}_{n,k}$ and $\hat{U}_{p,k}$ and the equation (70) lead to expressions of the form

$$\|\hat{U}_{n,k}\|_{L^\infty} + \|\hat{U}_{p,k}\|_{L^\infty} + \|\nabla^2 u_k\|_{L^\infty} \le c_1(\|n_k\|_{L^\infty} + \|p_k\|_{L^\infty}) + c_0, \tag{164}$$

where c_0 does not depend on k. The Gronwall inequality now leads to an exponential bound for $\|\nu_k\|_{L^\infty}$ and $\|\varpi_k\|_{L^\infty}$.

We may now proceed to second-derivative estimates. Multiplication of (82) by $-\nabla^2 e_k$ gives the algebraic relation

$$\begin{aligned}
&(\Delta t_k)^{-1}(e_k - e_{k-1})(-\nabla^2 e_k) + \mu_n(\nabla^2 e_k)^2 \\
&\quad = -\mu_n' \nabla^2 u_k \nabla e_k \nabla^2 e_k + \mu_n \nabla u_k \nabla e_k \nabla^2 e_k \\
&\qquad + \nabla n \nabla^2 u_k [\mu_n'(\nabla u_k) - \mu_n'(\nabla u)] \nabla^2 e_k - \nabla n [\nabla^2 u - \nabla^2 u_k] \mu_n'(\nabla u) \nabla^2 e_k \\
&\qquad - \nabla n [\mu_n(\nabla u_k)\nabla u_k - \mu_n(\nabla u)\nabla u] \nabla^2 e_k + [\mu_n(\nabla u_k) - \mu_n(\nabla u)] \nabla^2 n \nabla^2 e_k \\
&\qquad - V_{n,k} \nabla^2 e_k + r_{2,k} \nabla^2 e_k - f_k \nabla^2 e_k.
\end{aligned} \tag{165}$$

Here, f_k and $V_{n,k}$ have the meaning of (86) and (83), respectively. Integrate (165) over Ω, integrating the first term by parts, and utilize (91) with $\lambda = 9/\inf \mu_n$. We sum over $k = 1, \ldots, L$ to obtain

$$\begin{aligned}
&(\Delta t_L)^{-1} \|\nabla e_L\|_{L^2}^2 + \frac{1}{2} \inf \mu_n \sum_{k=1}^{L} \|\nabla^2 e_k\|_{L^2}^2 \\
&\quad \le \frac{9}{2 \inf \mu_n} \sum_{k=1}^{L} [(v_{sn}^2 + \mu_{0n}^4 \sup |\nabla^2 u_k|^2 / v_{sn}^2) \|\nabla e_k\|_{L^2}^2 + C(\|e_k\|_{L^2}^2 + \|q_k\|_{L^2}^2) \\
&\qquad + \|r_{2,k}\|_{L^2}^2 + \|f_k\|_{L^2}^2 + B^2(\|e_k\|_{L^2}^2 + \|q_k\|_{L^2}^2 + \|r_{1,k}\|_{L^2}^2)]
\end{aligned} \tag{166}$$

where

$$\begin{aligned}
B \;=\; & \frac{2(b-a)\mu_{0n}}{\pi\sqrt{\epsilon}} \left(\frac{\mu_{0n}^2}{v_{sn}^2} \sup |\nabla n| \sup |\nabla^2 u_k| + 2 \sup |\nabla n| + \frac{\mu_{0n}}{v_{sn}} \sup |\nabla^2 n| \right) \\
& + \left(\frac{2\mu_{0n}^2}{\epsilon v_{sn}} \right) \sup |\nabla n|,
\end{aligned} \tag{167}$$

and C is a constant arising from the fact that $V_{n,k}$ can be written as a sum of terms exhibiting linear growth in e_k and q_k (cf. (75) and (76)). Note that we have used the Lipschitz properties of $\xi \mapsto \mu_n(\xi)$, $\xi \mapsto \mu_n'(\xi)$, and $\xi \mapsto \xi \mu_n'(\xi)$. Use of the auxiliary estimate,

$$\|\nabla^2 u_k\|_{L^\infty} \le \epsilon^{-1}(\|n_k\|_{L^\infty} + \|p_k\|_{L^\infty} + \|N\|_{L^\infty} + \|r_{1,k}\|_{L^\infty}), \tag{168}$$

and the hypotheses of the Lemma, now yield the inequality

$$\begin{aligned}
&\sum_k (\|\nabla^2 e_k\|_{L^2}^2 + \|\nabla^2 q_k\|_{L^2}^2) \\
&\qquad \le C_1 \sum_k \|r_k\|_{L^2}^2 + \|\mathcal{P}\| \left(\|n_{tt}\|_{L^2(\Omega \times (0,T_0))}^2 + \|p_{tt}\|_{L^2(\Omega \times (0,T_0))}^2 \right).
\end{aligned} \tag{169}$$

The conclusion (159) now follows from (168) and (169) and the specification of the boundary conditions. ∎

Our next result details the estimation of the Lipschitz constant for F'.

Proposition 8 *For the domain of definition of F restricted to a ball of radius r_0 in X_s, a constant M_3 exists, depending only on r_0, in the case $s = 2$, such that (135) holds. In the case $s = 1$, the hypotheses of Lemma 7 are also required, as well as a generalized inverse inequality:*

$$\begin{aligned} \|\nabla n_1 - \nabla n_2\|_{L^\infty} &\leq C_{\|\mathcal{P}\|}\|\nabla n_1 - \nabla n_2\|_{L^2}, & (170)\\ \|\nabla p_1 - \nabla p_2\|_{L^\infty} &\leq C_{\|\mathcal{P}\|}\|\nabla p_1 - \nabla p_2\|_{L^2}. & (171)\end{aligned}$$

In this case, M_3 depends upon $\|\mathcal{P}\|$. Inequalities (170) and (171) need only hold in the sense of Remark 2. In the case $s = 1$, hypothesis (143) may be substituted for the hypotheses of Lemma 7.

Proof: We consider first the case $s = 2$. Thus, we must estimate M_3 in

$$\|[F'(v_1) - F'(v_2)][\phi, \psi, \omega]\|_{L^2} \leq M_3\|v_1 - v_2\|_{H^2}\,\|[\phi, \psi, \omega]\|_{H^2}\,, \tag{172}$$

where

$$v_1 = [u_1, n_1, p_1],\ v_2 = [u_2, n_2, p_2]. \tag{173}$$

The first component of $F'(v_1) - F'(v_2)$ is zero, as is evident from (119)–(123). We estimate separately the terms in the second component of the element on the left-hand side of (172), and we begin by estimating

$$\begin{aligned} &\nabla\cdot[\nabla n_1\mu_n'(\nabla u_1)\nabla\phi] - \nabla\cdot[\nabla n_2\mu_n'(\nabla u_2)\nabla\phi]\\ &\quad= \nabla^2(n_1 - n_2)\mu_n'(\nabla u_1)\nabla\phi + \nabla^2 n_2\left[\mu_n'(\nabla u_1) - \mu_n'(\nabla u_2)\right]\nabla\phi\\ &\qquad+\nabla(n_1 - n_2)\mu_n''(\nabla u_1)\nabla^2 u_1\nabla\phi\\ &\qquad+\nabla n_2\left[\mu_n''(\nabla u_1)\nabla^2 u_1 - \mu_n''(\nabla u_2)\nabla^2 u_2\right]\nabla\phi\\ &\qquad+\nabla(n_1 - n_2)\mu_n'(\nabla u_1)\nabla^2\phi + \nabla n_2\left[\mu_n'(\nabla u_1) - \mu_n'(\nabla u_2)\right]\nabla^2\phi. \end{aligned} \tag{174}$$

We find the norm of this expression is bounded by

$$\begin{aligned} &\left(\frac{\mu_{0n}^2}{v_{sn}}\right)\Big\{\|n_1 - n_2\|_{H^2}\|\nabla\phi\|_{L^2} + \left(\frac{\mu_{0n}}{v_{sn}}\right)\|\nabla\phi\|_{L^\infty}\left[\|n_2\|_{H^2}\|\nabla(u_1 - u_2)\|_{L^2}\right.\\ &\qquad\left.+\|u_1\|_{H^2}\|\nabla(n_1 - n_2)\|_{L^2} + \|\nabla n_2\|_{L^2}\|u_1 - u_2\|_{H^2}\right] \qquad (175)\\ &\quad+\left(\frac{15\mu_{0n}^3}{v_{sn}^2}\right)\|\nabla n_2\|_{L^\infty}\|\nabla^2 u_2\|_{L^2}\|\nabla(u_1 - u_2)\|_{L^2}\|\nabla\phi\|_{L^\infty}\\ &\quad+\|\nabla(n_1 - n_2)\|_{L^2}\|\phi\|_{H^2} + \left(\frac{\mu_{0n}}{v_{sn}}\right)\|\nabla n_2\|_{L^2}\|\nabla(u_1 - u_2)\|_{L^\infty}\|\phi\|_{H^2}\Big\}. \end{aligned}$$

These terms are clearly of the form given by the right-hand-side estimate of (172). The constant M_3 is seen to depend on r_0. The next two terms to estimate are given by

$$\nabla n_1\nabla u_1\mu_n'(\nabla u_1)\nabla\phi - \nabla n_2\nabla u_2\mu_n'(\nabla u_2)\nabla\phi$$

and

$$\nabla n_1 \mu_n(\nabla u_1)\nabla\phi - \nabla n_2 \mu_n(\nabla u_2)\nabla\phi.$$

Each of these two differences can be represented analogously to (174). The first difference replaces ∇n_1 and ∇n_2 by n_1 and n_2, respectively, and $\xi \mapsto \mu_n'(\xi)$ by $\xi \mapsto \xi\mu_n'(\xi)$, in the first two terms of the right-hand side of (174). The latter replacement simply means that, in (175), a bound for μ_n' is replaced by a bound for $\xi\mu_n'$; this bound is simply μ_{0n}. The second difference is handled even more easily. The estimation of the $(2,1)$ block differences is completed by consideration of the terms

$$\begin{aligned}\epsilon^{-1}\left\{2\left[\mu_n'(\nabla u_1) - \mu_n'(\nabla u_2)\right] + \left[\mu_n''(\nabla u_1)\nabla u_1 - \mu_n''(\nabla u_2)\nabla u_2\right]\right\}\left(n_1^2 - n_1 p_1 - n_1 N\right)\\ +\epsilon^{-1}\left[2\mu_n'(\nabla u_2) + \mu_n''(\nabla u_2)\nabla u_2\right]\left[(n_1^2 - n_2^2) - (n_1 p_1 - n_2 p_2) - (n_1 - n_2)N\right].\end{aligned}$$

Estimates for these terms in the matrix product are provided by

$$\begin{aligned}&\left(\frac{5\mu_{0n}^3}{\epsilon v_{sn}^2}\right)\|\nabla(u_1-u_2)\|_{L^2}\|n_1^2 - n_1p_1 - n_1N\|_{L^\infty}\|\nabla\phi\|_{L^2}\\ &\quad+\left(\frac{3\mu_{0n}^2}{v_{sn}}\right)\|\nabla\phi\|_{L^2}\left(\|n_1-n_2\|_{L^2}\|n_1+n_2\|_{L^\infty} + \|n_1-n_2\|_{L^2}\|p_1\|_{L^\infty}\right.\\ &\qquad\left.+\|p_1-p_2\|_{L^2}\|n_2\|_{L^\infty} + \|n_1-n_2\|_{L^2}\|N\|_{L^\infty}\right).\end{aligned}$$

Again this fits into the format of (172).

From the $(2,2)$ block, we estimate the difference

$$\begin{aligned}&\nabla\left[\mu_n(\nabla u_1)\nabla\psi\right] - \nabla\left[\mu_n(\nabla u_2)\nabla\psi\right]\\ &\quad= \left[\mu_n(\nabla u_1) - \mu_n(\nabla u_2)\right]\nabla^2\psi + \mu_n'(\nabla u_1)\nabla^2 u_1\nabla\psi - \mu_n'(\nabla u_2)\nabla^2 u_2\nabla\psi\end{aligned}$$

as

$$\begin{aligned}\left(\frac{\mu_{0n}^2}{v_{sn}}\right)\left[\|\nabla(u_1-u_2)\|_{L^2}\|\psi\|_{H^2} + \left(\frac{\mu_{0n}^2}{v_{sn}}\right)\|\nabla(u_1-u_2)\|_{L^2}\|\nabla^2 u_1\|_{H^2}\|\nabla\psi\|_{L^2}\right.\\ \left.+\|u_1-u_2\|_{H^2}\|\nabla\psi\|_{L^2}\right],\end{aligned}$$

which is again of the form (172). The difference

$$\mu_n(\nabla u_1)\nabla u_1\nabla\psi - \mu_n(\nabla u_2)\nabla u_2\nabla\psi$$

is estimated by

$$2\mu_{0n}\|\nabla(u_1-u_2)\|_{L^2}\|\nabla\psi\|_{L^2},$$

which is of the form (172). The terms

$$\left[\frac{\partial U_n}{\partial n}(u_1,n_1,p_1) - \frac{\partial U_n}{\partial n}(u_2,n_2,p_2)\right]\psi$$

and

$$\left[\frac{\partial U_p}{\partial p}(u_1,n_1,p_1) - \frac{\partial U_p}{\partial p}(u_2,n_2,p_2)\right]\omega$$

are estimated by techniques similar to the above. This concludes the proof for $s = 2$.

In the case $s = 1$, the estimation of the linear functional,

$$(F'[u_1, n_1, p_1] - F'[u_2, n_2, p_2])[\phi, \psi, \omega],$$

proceeds by estimating terms from the $(2,1)$ block differences beginning with

$$\int_\Omega (\nabla n_1 - \nabla n_2)\mu_n'(\nabla u_1)\nabla\phi\nabla\psi_0 + \int_\Omega \nabla n_2 \left(\mu_n'(\nabla u_1) - \mu_n'(\nabla u_2)\right)\nabla\phi\nabla\psi_0,$$

where the functional acts on the test function $[\phi_0, \psi_0, \omega_0]$. This can be estimated by

$$\left(\frac{\mu_{0n}^2}{v_{sn}}\right)\|\nabla(n_1 - n_2)\|_{L^\infty}\|\nabla\phi\|_{L^2}\|\nabla\psi_0\|_{L^2} + \left(\frac{\mu_{0n}^3}{v_{sn}^2}\right)\|\nabla n_2\|_{L^\infty}\|\nabla(u_1 - u_2)\|_{L^\infty}\|\nabla\phi\|_{L^2}\|\nabla\psi_0\|_{L^2}.$$

This is the most delicate of the estimates, since it involves $\|\nabla(n_1 - n_2)\|_{L^\infty}$ and $\|\nabla n_2\|_{L^\infty}$. The latter can be estimated by Lemma 7, or alternatively (143). The former requires an estimate of the form

$$\|\nabla(n_1 - n_2)\|_{L^\infty} \le C_{\Delta t}\|\nabla(n_1 - n_2)\|_{L^2},$$

which is a property of the inverse hypothesis (170). The next two terms in the $(2,1)$ block difference do not require special hypotheses, while the final term in the $(2,1)$ block may make use of the estimate for $s = 2$. The terms in the $(2,2)$ and $(2,3)$ blocks are routinely estimated. ∎

4.4 Interface with the Inner Iteration

In Part II of this series, we shall examine decoupling inner iterations, giving rise to approximate inverses of the derivative maps $F'(u, n, p)$. Detailed analysis of typical decouplings of this type suggests that the approximate inverse satisfies an inequality of the form

$$\|[(F' \circ G)(u_k^{m-1}, n_k^{m-1}, p_k^{m-1}) - I]r^{m-1}\|_{H^{s-2}} \le C\Delta t_k\|r^{m-1}\|_{H^s}. \tag{176}$$

Once again, r^{m-1} denotes the residual, at the kth time step, of the $(m-1)$th Newton iteration. The conjunction of (176) and (142) suggests the inequality

$$\|G(u_k^{m-1}, n_k^{m-1}, p_k^{m-1})r^{m-1}\|_{H^s} \le M(1 + C\Delta t_k)\|r^{m-1}\|_{H^{s-2}}, \tag{177}$$

leading to the choice of M_1 in (133), dictated by

$$M_1 \stackrel{\text{def}}{=} M(1 + C\|\mathcal{P}\|). \tag{178}$$

Here M is defined in (142). To draw the circle completely around (133)–(135), we can identify M_2 of (134) with M_1 and M_3 by

$$M_2 \stackrel{\text{def}}{=} M_1^2 M_3, \tag{179}$$

and then link this choice of M_2 with that induced by (176):

$$M_1^2 M_3 \|r^{m-1}\|_{H^{s-2}} \geq C\Delta t_k. \tag{180}$$

The choice of M_2 in (179) is motivated by (140). Clearly, (180) is implied by

$$M^2 M_3 \|r^{m-1}\|_{H^{s-2}} \geq C\|\mathcal{P}\|, \tag{181}$$

for each $m \leq m_0$. Inequality (181) suggests that a small number of Newton iterations is desirable, to minimize the restriction on $\|\mathcal{P}\|$, provided this is consistent with the residual tolerance described by Theorems 3 and 4. If, as motivated by those theorems, we desire

$$\|r^m\|_{H^s} \leq c\Delta t_k, \tag{182}$$

for specified c, then (138) shows that

$$c\Delta t_k \geq (2M_1^2 M_3)^m \|r^0\|_{H^{s-2}}^{2^m} \tag{183}$$

must be satisfied.

Condition (183) imposes an implicit condition on $\|r^0\|_{H^{s-2}}$. For example, if only one Newton iteration is desired, then

$$\|r^0\|_{H^{s-2}} \leq \sqrt{\frac{c\Delta t_k}{2M_1^2 M_3}}. \tag{184}$$

One is tempted to use as a starting guess in the Newton iteration the accepted approximation at the previous time step. Such a selection may fail to satisfy conditions such as (183) or (184). In the next section we briefly discuss suggested procedures, specifically continuation, to achieve this.

4.5 Bridging Time Steps

In an actual implementation, one might select a second-order extrapolation method, based upon the accepted approximations at the two preceding time steps, as a way of determining the starting iterate. Since this depends on the fully discrete algorithm which we shall consider in the sequel to this paper (Part II), we prefer to describe a continuation procedure, which is amenable to the modularity inherent in our algorithmic analysis.

For simplicity, denote the map (118) at the $(k-1)$th time step by F_{k-1} and the map at the kth time step by F_k. By the criterion established in Theorem 3 or Theorem 4, if an approximation triple z_{k-1} is accepted at the $(k-1)$th time step,

then $F_{k-1}(z_{k-1})$ is of magnitude $c\Delta t_{k-1}$, in an appropriate norm. We shall allow F_{k-1} and F_k to act, more generally, on elements of the form

$$v_\lambda = v + (1-\lambda)\bar{z}_{k-1} + \lambda \bar{z}_k,$$

where $v \in Y_s$ (cf. (131)) and $\bar{z}_{k-1}$, $\bar{z}_k$ assume the boundary values. Define the homotopy map,

$$F(v,\lambda) = (1-\lambda)F_{k-1}(v_\lambda) + \lambda F_k(v_\lambda), \quad v \in Y_s,\ 0 \le \lambda \le 1. \tag{185}$$

Knowing that $F_{k-1}(z_{k-1})$ is of magnitude $c\Delta t_{k-1}$, we wish to vary λ, introducing intermediate approximations thereby, so that at termination, z_k^0 is obtained, where $F_k(z_k^0)$ satisfies, say, (184), or the more general residual condition (183). Here z_k^0 is a starting guess for a Newton outer iteration. A strategy consistent with the framework of the paper is a predictor/delayed-corrector method. Roughly, assuming $\Delta t_{k-1} = \Delta t_k = \Delta t$, we have $\lfloor N/2 \rfloor$ for the number of Euler predictor iterations performed at successive λ-values λ_i, before a series of m Newton corrector iterations is performed at some λ_i. Here

$$N^{-1} = (c\Delta t)^{1-1/2^m}(2M_1^2 M_3)^{m/2^m}. \tag{186}$$

The effect of the corrector iterations is to restore the residual to the order $c\Delta t$. We shall present the theory underlying these statements.

The map defined by (185) satisfies the algebraic relation

$$F(v,\lambda) = F(w,\mu) + F'_w(w,\mu)(v-w) + F'_\mu(w,\mu)(\lambda-\mu) + R(v,w;\lambda,\mu), \tag{187}$$

where

$$\begin{aligned} R(v,w;\lambda,\mu) \ \stackrel{\text{def}}{=}\ & (1-\mu)\{F_{k-1}(v_\lambda) - F_{k-1}(w_\mu) - F'_{k-1}(w_\mu)(v_\lambda - w_\mu)\} \\ & +\mu\{F_k(v_\lambda) - F_k(w_\mu) - F'_k(w_\mu)(v_\lambda - w_\mu)\} \\ & +(\mu-\lambda)\{[F_{k-1}(v_\lambda) - F_{k-1}(w_\mu)] - [F_k(v_\lambda) - F_k(w_\mu)]\}. \end{aligned} \tag{188}$$

If hypothesis (135) is satisfied, then (187) and (188) imply the smoothness estimate

$$\|R(v,w;\lambda,\mu)\| \le c_1\|v_\lambda - w_\mu\|^2 + c_2(\lambda-\mu)^2. \tag{189}$$

The hypotheses which are analogous to, but weaker than, (133) and (134) are:

$$\|H(v,\lambda)z\| \ \le\ M'_1\|z\|, \tag{190}$$

$$\|[F'_v(v,\lambda)H(v,\lambda) - I]z\| \ \le\ M'_2\|z\|, \tag{191}$$

for $z = F'_\lambda(v,\lambda)$, where $\{\, H(v,\lambda)\,\}$ comprises a set of approximate right inverses for $F'_v(v,\lambda)$. Note that $\|z\|$, not $\|z\|^2$, appears in (191); also, notice that $F'_\lambda(v,\lambda)$ has the representation

$$F'_\lambda(v,\lambda) = [F_k(v_\lambda) - F_{k-1}(v_\lambda)] + [\lambda F'_k(v_\lambda) + (1-\lambda)F'_{k-1}(v_\lambda)](\bar{z}_k - \bar{z}_{k-1}). \tag{192}$$

The control of $|\Delta\lambda|\,\|F'_\lambda(v,\lambda)\|$ is critical for the continuation, as v_{i-1} is replaced by a new predictor v_i. In practice, $\Delta\lambda$ is chosen adaptively, but here we assume for simplicity that the λ-points are equally spaced. We have the following.

Proposition 9 *Let*

$$\rho = \|F_{k-1}(z_{k-1})\| \leq c\Delta t_{k-1}, \tag{193}$$

suppose that $z_{k-1} \in B_{\alpha r_0}$, $0 \leq \alpha < 1$, *and let* $\Delta\lambda$ *be specified, with* $\lambda_i \stackrel{\text{def}}{=} i\Delta\lambda$. *Here* B_{r_0} *is a ball in* H^2 *of radius* r_0. *Suppose, for simplicity, that* $\Delta t_{k-1} = \Delta t_k = \Delta t$ *and that the successive Euler predictors are defined by* $v_0 \stackrel{\text{def}}{=} z_{k-1}$,

$$\frac{v_i - v_{i-1}}{\Delta\lambda} = -H(\hat{v}_{i-1}, \lambda_{i-1})F'_\lambda(\hat{v}_{i-1}, \lambda_{i-1}), \text{ for each } i. \tag{194}$$

Here, $\hat{v}_i$ *denotes the affine translate of* v_i *such that* $\hat{v}_i \in Y_s$. *Then, if* $\Delta\lambda$ *is such that (197) does not exceed* $c\Delta t$, *and such that*

$$|\Delta\lambda| M'_1 \|F'_\lambda(\hat{v}_{i-1}, \lambda_{i-1})\| \leq 2(1-\alpha)r_0/N, \tag{195}$$

where N *is given by (186), then it follows that* $v_1, \ldots, v_{\lfloor N/2 \rfloor}$ *are in* B_{r_0}, *and* $v_{\lfloor N/2 \rfloor}$ *satisfies the residual condition (183). In particular, the* m*th Newton iterate, starting with* $v_{\lfloor N/2 \rfloor}$ *as the zeroth iterate, again has residual of magnitude* $c\Delta t$.

Proof: By direct estimation, if w_0 denotes the center of B_r, we have, for $i = 1, \ldots, \lfloor N/2 \rfloor$,

$$\begin{aligned} \|v_i - v_0\| &\leq \|v_i - v_{i-1}\| + \|v_{i-1} - v_0\| + \|v_0 - w_0\| \\ &\leq |\Delta\lambda| \, \|H(\hat{v}_{i-1}, \lambda_{i-1})\| \, \|F'_\lambda(\hat{v}_{i-1}, \lambda_{i-1})\| + 2(i-1)(1-\alpha)r_0/N + \alpha r_0 \\ &\leq 2i(1-\alpha)r_0/N + \alpha r_0 \end{aligned}$$

if (195) and the induction hypotheses are invoked. It follows that $v_1, \ldots, v_{\lfloor N/2 \rfloor}$ are in B_{r_0}.

For the residual we have,

$$\begin{aligned} F(\hat{v}_i, \lambda_i) &= -\Delta\lambda[F'_v(\hat{v}_{i-1}, \lambda_{i-1})H(\hat{v}_{i-1}, \lambda_{i-1}) - I] \circ F'_\lambda(\hat{v}_{i-1}, \lambda_{i-1}) \\ &\quad + R(\hat{v}_i, \hat{v}_{i-1}; \lambda_i, \lambda_{i-1}) + F(\hat{v}_{i-1}, \lambda_{i-1}). \end{aligned} \tag{196}$$

The estimation of the first two terms is given by

$$|\Delta\lambda| M'_2 \|F'_\lambda(\hat{v}_{i-1}, \lambda_{i-1})\| \{1 + c_1 M'_2 |\Delta\lambda| \, \|F'_\lambda(\hat{v}_{i-1}, \lambda_{i-1})\|\} + c_2 |\Delta\lambda|^2. \tag{197}$$

By the hypothesis on $\Delta\lambda$, this quantity does not exceed $c\Delta t$. It follows from this analysis that the residual corresponding to $v_{\lfloor N/2 \rfloor}$ does not exceed $Nc\Delta t$. By direct computation, this satisfies (183). ∎

5 Postscript

This paper has described a modular algorithm development beginning with the semidiscrete systems determined by a fully implicit time discretization. Though first-order methods are analyzed here, the ideas are capable of extension to higher-order time discretizations.

The modular development proceeds from the fundamental observation, contained in Theorems 3 and 4, that the accuracy of Euler's method is maintained by approximate solvability, as specified by the residuals. The approximate solutions will not, in general, satisfy the invariant-region principle developed for the exact semidiscrete solution in Theorem 1, though this theory provides the appropriate analytical infrastructure, as well as a guide to the properties desired of a computational procedure.

Newton's method is our choice for approximate solvability. If the approximate solvability condition is expressed in terms of a residual, numerically bounded in norm by $c\Delta t_k$, and if m is the number of outer Newton iterations at t_k, then inequality (183) determines the necessary condition on the starting guess at the commencement of the outer iterations at $t = t_k$. The constants M_1 and M_3 in this inequality are associated with a class of generic approximate Newton methods. Their relation to the semiconductor model is drawn via (178) and Propositions 6 and 8. Although the constants appear not to be completely explicit, the existential components can be eliminated by a careful correlation between the radius of the local domain of definition of F and gradient estimates induced by Sobolev-type inequalities in one dimension.

It is natural to wish to use the accepted approximation at the previous time step as the starting guess at the current time step. Such a starting iterate, however, may fail to satisfy (183), if m is maintained as a small positive integer, e.g., $m = 1$ or $m = 2$; the potential boundary conditions are clearly not satisfied either. An approximate time-stepping bridge is furnished by a continuation method, specifically, a predictor-corrector method as described in Proposition 9. The analytical framework permits several Euler predictor steps, before approximate Newton steps are applied.

The function-space framework, in which the approximate Newton methods are analyzed at the differential equation level, admits both a classical and variational formulation. These are loosely organized via L^2 and H^{-1} settings. The latter requires a significant amount of technical supporting analysis; e.g., the $s = 1$ subcase of Proposition 6 requires the companion Lemma 7. Both the settings have been included with an eye to the fully discrete algorithms to be developed in the sequel to this paper, Part II.

We have deliberately refrained from specifying the specific form of the approximate Newton method. In fact, we view the exact Newton iterations as *outer iterations*, each realized in terms of a number of so-called inner iterations. The sum total of these inner iterations, corresponding to each outer iteration, represents the approximate Newton iteration; this process is repeated m times at each time step. The inner iterations are visualized as cumulative, in the sense that each succeeding iteration enhances the previous; also, it must be fully discrete in its final form. Our numerical studies thus far have indicated that the substitution giving rise to (14) and (15) creates a 'de facto' scale imbalance in the Jacobian; in order for a stable computation to proceed, some time-step restriction must be imposed. We shall deal with these issues in Part II, where a specific class of algorithms together with numerical studies, will be presented.

References

[1] R. E. Bank, W. M. Coughran, Jr., W. Fichtner, E. H. Grosse, D. J. Rose, and R. K. Smith. Transient simulation of silicon devices and circuits. *IEEE Trans. CAD*, CAD-4:436–451, 1985.

[2] R. E. Bank, J. W. Jerome, and D. J. Rose. Analytical and numerical aspects of semiconductor device modeling. In R. Glowinski and J.-L. Lions, editors, *Computing Methods in Applied Sciences and Engineering, V*, pages 593–597. North-Holland, 1982.

[3] R. E. Bank and D. J. Rose. Global approximate Newton methods. *Numer. Math.*, 37:279–295, 1981.

[4] R. E. Bank, D. J. Rose, and W. Fichtner. Numerical methods for semiconductor device simulation. *IEEE Trans. ED*, ED-30:1031–1041, 1983.

[5] A. Berman and R. J. Plemmons. *Nonnegative Matrices in the Mathematical Sciences*. Academic Press, 1979.

[6] D. M. Caughey and R. E. Thomas. Carrier mobilities in silicon empirically related to doping and field. *Proc. IEEE*, 55:2192–2193, 1967.

[7] A. DeMari. An accurate numerical one-dimensional solution of the P-N junction under arbitrary transient conditions. *Solid-State Electron.*, 11:1021–1053, 1968.

[8] A. DeMari. An accurate numerical steady-state one-dimensional solution of the P-N junction. *Solid-State Electron.*, 11:33–58, 1968.

[9] D. Gilbarg and N. S. Trudinger. *Elliptic Partial Differential Equations of Second Order*. Springer-Verlag, 2nd edition, 1983.

[10] R. N. Hall. Electron-hole recombination in germanium. *Physical Review*, 87:387, 1952.

[11] J. W. Jerome. *Approximation of Nonlinear Evolution Systems*. Academic Press, 1983.

[12] J. W. Jerome. Approximate Newton methods and homotopy for stationary operator equations. *Constr. Approx.*, 1:271–285, 1985.

[13] J. W. Jerome. Consistency of semiconductor modeling: An existence/stability analysis for the stationary van Roosbroeck system. *SIAM J. Appl. Math.*, 45:565–590, 1985.

[14] J. W. Jerome. Evolution systems in semiconductor device modeling: A cyclic uncoupled line analysis for the Gummel map. *Math. Methods Appl. Sci.*, to appear.

[15] T. Kato. The Cauchy problem for quasi-linear symmetric hyperbolic systems. *Arch. Rational Mech. Anal.*, 58:181–205, 1975.

[16] P. A. Markowich. *The Stationary Semiconductor Device Equations.* Springer-Verlag, 1986.

[17] R. P. Mertens, R. J. van Overstraeten, and H. J. de Man. Heavy doping effects in silicon. *Advances in Electronics and Electron Phys.*, 55:77–118, 1981.

[18] W. Shockley and W. T. Read. Statistics of the recombination of holes and electrons. *Physical Review*, 87:835–842, 1952.

[19] S. M. Sze. *Physics of Semiconductor Devices.* Wiley-Interscience, 2nd edition, 1981.

[20] K. K. Thornber. Relation of drift velocity to low-field mobility and high-field saturation velocity. *J. Appl. Phys.*, 51:2127–2136, 1980.

[21] W. Van Roosbroeck. Theory of flow of electrons and holes in germanium and other semiconductors. *Bell System Tech. J.*, 29:560–607, 1950.

[22] R. S. Varga. *Matrix Iterative Analysis.* Prentice-Hall, 1962.

Computing Mathematics Research Department
AT&T Bell Laboratories
Murray Hill, New Jersey 07974

Department of Mathematics
Northwestern University
Evanston, Illinois 60201