NOTE: BIBLIOGRAPHY REFERENCES ARE MISSING

On Acceleration Methods for Coupled Nonlinear Elliptic Systems.

T. Kerkhoven* and Y. Saad†

September 25, 2002

Abstract

We compare both numerically and theoretically three techniques for accelerating the convergence of a nonlinear fixed point iteration arising from a system of coupled partial differential equations: Chebyshev acceleration, a second order stationary method, and a nonlinear version of the Generalized Minimal Residual Algorithm (GMRES) which we call NLGMR. All three approaches are implemented in 'Jacobian-free' mode, i.e., only a subroutine which returns $T(u)$ as a function of u is required. We present a set of numerical comparisons for the drift-diffusion semiconductor model. For the mapping T which corresponds to the nonlinear block Gauß-Seidel algorithm for the solution of this nonlinear elliptic system, NLGMR is found to be superior to the second order stationary method and the Chebychev acceleration. We analyze the local convergence of the nonlinear iterations in terms of the spectrum $\sigma[T_u(u^{(*)})]$ of the derivative T_u at the solution $u^{(*)}$. The convergence of the original iteration is governed by the spectral radius $\rho[T_U(u^{(*)})]$. In contrast the convergence of the two second order accelerations are related to the convex hull of $\sigma[T_u(u^{(*)})]$, while the convergence of the GMRES-based approach is related to the local clustering in $\sigma[I - T_u(u^{(*)})]$. The spectrum $\sigma[I - T_u(u^{(*)})]$ clusters only at 1 due to the successive inversions of elliptic partial differential equations in T. We explain the observed superiority of GMRES over the second order acceleration by its ability to take advantage of this clustering feature, which is shared by similar coupled nonlinear elliptic systems.

Key words: Compactness, Convergence, Nonlinear Maps, Conjugate gradients, Polynomial acceleration, matrix-free methods.

AMS(MOS) Classification: 65B, 65H10, 65N.

Acknowledgements: The first author is supported by the National Science Foundation under grant EET-8719100. The second author was supported in part by the National Science Foundation under Grants No. US NSF DCR84-10110 and US NSF DCR85-09970, by the US Department of Energy under Grant No. DOE DE-FG02-85ER25001, by the US Air Force under Contract AFSOR-85-0211, and by an IBM donation.

*University of Illinois at Urbana-Champaign, 1304 W. Springfield Ave.,Urbana, IL 61801

†Research Institute for Advanced Computer Science, NASA Ames Research Center, Moffett Field CA 94035.

1 Introduction.

In this paper we are concerned with the efficient solution of a system of coupled nonlinear elliptic Partial Differential Equations (PDE's.) For problems of this kind, it is often the case that efficient solution methods are available for each of the single equations, but not for the entire coupled system. The availability of efficient solvers for many single elliptic problems suggests using relaxation type techniques in which each of the elliptic PDE's is solved successively. Consistency is then obtained through an outer iteration. However, the convergence properties of the outer iteration may not be satisfactory and it may be desirable to apply an acceleration procedure. In this paper we show numerically, and theoretically, that for suitable elliptic systems, a Jacobian-free implementation of Newton's method which is based on a nonlinear version of the Generalized Minimum Residual algorithm GMRES [19] is an effective accelerator of the outer iteration.

Formally, the original iteration consists of a repeated application of a nonlinear fixed point mapping T which arises from the discretization of a mapping T_C, which operates on functions through the solution of certain elliptic PDE's. The fixed point $u^{(*)} = Tu^{(*)}$ of T corresponds to the simultaneous solution of the system of elliptic partial differential equations. In our test problem this mapping T is defined through the system of elliptic boundary value problems which corresponds to the drift-diffusion model for steady state semiconductors

$$\nabla\cdot(\exp(u-v)\nabla v) = 0, \tag{1.1}$$

$$\nabla\cdot(\exp(-u+w)\nabla w) = 0, \tag{1.2}$$

$$-\nabla^2 u + \exp(u-v) - \exp(w-u) - k_1 = 0, \tag{1.3}$$

subject to appropriate (possibly mixed) boundary conditions. This model is discussed further in Appendix A. A well-known technique, known as Gummel's iteration, consists of a nonlinear Gauß-Seidel iteration on the above system, or rather its discretized analogue. Thus, v is updated from its old value by solving (1.1) then w is updated by solving (1.2) and finally u is updated from the third equation. We call T the mapping that updates an old value of u to its new value $\tilde{u}$.

The idea of the techniques discussed in this paper is to accelerate the fixed point iteration $\tilde{u} = T(u)$.

We consider three such techniques: a nonlinear GMRES acceleration (called NLGMR), a second order stationary acceleration and a Chebyshev acceleration. All three take advantage, either explicitly or implicitly, of certain features in the eigenvalue spectrum of the Jacobian T_u of the nonlinear mapping T. Since T is defined from successive solutions of elliptic boundary value problems, each component of the output vector $T(u)$ will depend on all components of the input vector u. Therefore, neither T, nor its Jacobian T_u, are sparse. For this reason explicit representation of the Jacobian matrix is circumvented by exploiting iterative methods and finite difference approximations to compute Jacobian by vector products. Thus, all three approaches are Jacobian-free in that they require only a subroutine which takes as input an argument u and returns the vector $T(u)$.

The nonlinear GMRES approach amounts to applying an inexact Newton method [8] to solve the equation $u - T(u) = 0$, defined implicitly by successive solution of the elliptic equations. As is shown in §2, for a fixed point mapping T, which is defined in a similar way by inversion of elliptic differential operators, the spectrum of the Jacobian $I - T_u$, which must be inverted in Newton's method, clusters only at 1. As a result, GMRES can solve these systems of equations efficiently. The ellipticity of the equations in the system causes a clustering of the eigenvalues of the Jacobian at 1 for this choice of the mapping T. Therefore, in this approach the solution of Newton's equations $[I - T_u]du = -[u - T]$ by GMRES does not require preconditioning. In other words, this version of Newton's method for the solution of the nonlinear elliptic system is *already preconditioned* in that an implicit form of preconditioning is embedded in the operator T. Our techniques can be viewed alternatively as a preconditioned nonlinear Krylov subspace method where the preconditioning consists of one step of the nonlinear fixed point algorithm. This viewpoint was taken in [5] in a similar context.

In our experiments on the steady state drift-diffusion model (1.1-1.3), we found NLGMR to be more effective than the other two acceleration techniques. In §2, we present some results concerning the eigenvalue distributions in the spectra of compact mappings defined by the successive solution of elliptic boundary value problems. The compactness of the fixed point mapping T_C leads to a clustering of the eigenvalues of the derivative $I - T_{C,u}$ at 1. Conjugate gradient type algorithms such as the GMRES-based approach can take advantage of this clustering. This property, which is valid for the

inversion of similar elliptic operators, yields a clue on the superiority of the GMRES-based approach. This claim is based on Lemma 5.1 where we prove that the inversion of the *elliptic* operators in T_C results in clustering rates in $\sigma[T_{C,u}]$ which yield a superlinear convergence rate for suitable conjugate gradient algorithms. Because the clustering can be related to the inversion of elliptic operators in the mapping T_C at the PDE level, we expect the GMRES based approach to be an effective algorithm for the solution of similar coupled elliptic systems. However, it is not clear whether or not the same approach will be effective in accelerating fixed point mappings T that arise from systems of more general, not necessarily elliptic, partial differential equations.

Iteration with the nonlinear mapping T, used in our tests, converges rapidly far away from the solution $u^{(*)}$, but slows down once $u^{(*)}$ is approached. This effect is analyzed in [**?**] and explained in terms of the maximum principles which follow for the solution $u^{(*)}$ from the ellipticity of each of the coupled PDE's. Therefore, again specifically for elliptic systems, NLGMR acceleration which is based on Newton's method, is likely to be needed close to the solution only. However, near the solution Newton's method converges quadratically, and therefore the proposed acceleration will be most valuable precisely when the original fixed point iteration slows down. Global convergence strategies [**?**], can be incorporated to take care of the intermediate regions between the two regimes of convergence.

In §4, we examine the convergence for the second order schemes. Both second order methods are based on Chebychev acceleration. We start with the linear case and then examine the implications for nonlinear mappings. Theorem 4.1 and Corollary 4.1 are special cases of Theorems 3 and 2, respectively, in [**?**]. We present a simple proof for the special case that we are concerned with here. For greater generality the reader is referred to [**?**]. Our analysis of the second order methods depends on the convex hull of the spectrum of the derivative $T_u(u^*)$ at the fixed point $u^{(*)}$ only. The precise clustering properties of the eigenvalue spectrum are not directly used here. In related work, Hyman and Manteuffel [12] consider both first order and second order acceleration methods while Jesperson and Buning [15] consider first order acceleration schemes only. In both cases, the acceleration procedure is based on linearization arguments and no convergence proof was given for the nonlinear case. As in [**?**], we account for nonlinearities in the mapping T and prove convergence using a theorem from [**?**],

a somewhat more general and elaborate version of Ostrowski's Theorem.

One of the goals of the paper is to provide a theoretical foundation for the accelerations which were obtained for the steady state drift-diffusion model (1.1-1.3) with the considered techniques. However, a few of our arguments are conjectural in nature. When such is the case this is stated explicitly. With all three acceleration techniques we were able to decrease the total number of iterations and computation time considerably. In particular, as mentioned earlier, NLGMR was far more powerful than the two second order techniques, and has the advantage that it does not require the a-priori knowledge of the outmost eigenvalues of T_u which determine the convex hull of the spectrum. In the examples treated, NLGMR reduced the computational time by a factor larger than 8.

An alternative approach to the solution the nonlinear elliptic system (1.1-1.3) is the technique proposed by, e.g., Bank and Rose in [?] and the related joint paper with Fichtner [?]. This method is based on modified versions of Newton's method applied globally to the discretized PDE's. However, the solution of the Jacobian systems required at each iteration of Newtons's method, may pose a major computational task. In [?] it is stated that for large scale problems, especially those originating from three-dimensional models, it becomes imperative to solve these linear systems by means of iterative methods, specifically preconditioned Conjugate Gradient type methods*. The direct application of Newton's method relies heavily for its effectiveness on sparsity and computability of the Jacobian and the availability of of a good preconditioners. Depending on the application, both of these conditions may be difficult, if not impossible, to satisfy. This is the case, for example, when eigenvalue calculations are part of the coupled systems [?, ?]. In fact, for the quantum well problem in [?, ?], which includes the solution of an eigenvalue problem for Schrödinger's equation, we have shown the effectiveness of the proposed technique in [?]. We do not claim that the nonlinear GMRES is more effective than the global Newton approach, in cases where there are effective ways of generating the Jacobian and preconditioning it. On the other hand we do believe that the approach based on nonlinear GMRES acceleration of fixed point mappings, is far easier to implement in cases of significantly more complex equations, and that it has a wider range of applicability.

*However, notice that the cited paper deals with the simulation of transients by an algebraic-differential system, rather than solution of the steady state problem.

2 Fixed Point Convergence.

In this section we introduce some results which relate the speed of convergence of a nonlinear mapping T to the properties of the spectrum of its derivative $T_U(u^*)$ at its fixed point u^*. Moreover, we discuss the distribution of eigenvalues in the spectrum of a compact mapping T_C, or more specifically, in the spectrum of a mapping T which is defined by the successive inversion of elliptic operators.

First, we examine the iteration with the fixed point mapping T by itself. In this case the local speed of convergence depends on the spectral radius $\rho[T_u(u^{(*)})]$ at the solution $u^{(*)}$ only. An analysis for the original elliptic system is presented in [**?**]. Theorem **2.3** from [**?**] asserts convergence, and specified the convergence rate, in terms of the spectral radius of T.

Theorem 2.1 *Let the mapping T be Fréchet differentiable at its fixed point x^*. Let ρ_0 denote the spectral radius of the linear operator $T^{'}(x^*)$, and assume that $\rho_0 < 1$. Then for any positive scalar ϵ there is an x_0 sufficiently close to x^* and a scalar $c(x_0, \epsilon)$ such that the successive approximation*

$$x_{n+1} = Tx_n, \quad n = 0, 1, 2, \cdots$$

converges to x^*, and

$$\|x_n - x^*\| \leq c(x_0, \epsilon)(\rho_0 + \epsilon)^n. \tag{2.1}$$

The two second order recurrences are determined from the convex hull of the spectrum $\sigma[T_u(u^{(*)})]$. Therefore speed of the accelerated fixed point iteration depends on the convex hull as well. For the steady state model (1.1-1.3) the location of the convex hull of $\sigma[T_u]$ was determined theoretically in [**?**], and [**?**]. In particular, this convex hull is related to specific discretization techniques. The location in the complex plane of the convex hull of the spectrum $\sigma[T_u(u^{(*)})]$ is not directly related to the ellipticity of the PDEs in the system defining the mapping T,

The speed of convergence of the conjugate gradient method (on which the NLGMR method is partly based), on the other hand, depends on the distribution in the spectrum of $T_u(U^*)$ in a different way than either of the two algorithms mentioned above. GMRES is a conjugate gradient algorithm that minimizes the residual $b - Ax$ over expanding Krylov subspaces [19]. Therefore roughly speaking at each iteration NLGMR ought to be at least as effective as either of the two second order accelerations.

Moreover, the GMRES algorithm is used for solving Newton's equations $[I - T_u]du = -[u - T(u)]$. Therefore, the matrix A is equal to $I - T_u$ with eigenvalues $\mu_i = 1 - \lambda_i$, where λ_i is an eigenvalue of T_u. However, because the mapping T_C involves inversions of the Laplacean, the eigenvalues λ_i of T_u cluster at 0 only. For example, the eigenvalues of the Laplacean on the unit cube in N dimensions are given by

$$\lambda_{n_1,\cdots,n_N} = \pi^2[n_1^2 + \cdots + n_N^2].$$

Therefore, the number $N(R)$ of eigenvalues $\lambda_{n_1,\cdots,n_N}$ with size smaller than R is given asymptotically by $N(R) = c * R^{\frac{N}{2}}$. It is immediate that if $M(R)$ is the number of eigenvalues of the inverse operator which are bigger in size than R, then $M(R) = k*R^{-\frac{N}{2}}$. This implies that the eigenvalues of the inverted Laplacean accumulate at 0 only. The accumulation rate follows from the asymptotic expression above. This asymptotic clustering rate for the Laplacean is known to be valid for more general elliptic PDEs in divergence form (see, for instance, [**?**], p. 250.) However, there does not seem to exist a result in the literature that applies to our situation exactly.

In Lemma 5.1 of this paper we demonstrate that a residual minimizing algorithm like GMRES reduces the residual to 0 at least superlinearly for the clustering rates at 1 in the spectrum $\sigma[I - T_u(u^{(*)})]$ which we discussed above. Alternative analyses of the convergence of the conjugate gradient method depending on the distribution of the eigenvalues in the spectrum of a matrix A have been presented by Jennings [13], and by van der Sluis and van der Vorst [22]. Similar, but weaker, results for the GMRES algorithm are presented in [19]. Although Lemma 5.1 does imply superlinear convergence, it is difficult to estimate the actual convergence rate over subspaces of lower dimensions.

T_C represents a fixed point operator defined through successive solution of a system of coupled nonlinear elliptic Partial Differential Equations. This implies that T_C is compactly differentiable because the highest order operator is inverted for each of the PDE's. Even if not all of the results about clustering rates which we found in the literature can be applied directly to our model, we are able to provide a complete mathematical proof of clustering of the eigenvalues of T_u at 0 without specification of the clustering rate. This proof is based on the well-known result that the spectrum of a compact linear mapping can accumulate at 0 only. Before we state this theoretical result we

introduce the following formalism. We let

- X be a linear space,
- $L(X)$ be the space of bounded linear operators of X into itself,
- $\sigma(T)$ be the spectrum of T.

Next we state the following theorem from [6] p. 117 or [?] p. 281, about the spectral properties of compact operators.

Theorem 2.2 *Let $T \in L(X)$, and let T be compact. Then $\sigma(T)$ is a countable set with no accumulation point other than 0. Each nonzero $\lambda \in \sigma(T)$ is an isolated eigenvalue of T with finite algebraic multiplicity.*

In the appendix of [?] compact differentiability is proven for the mapping T_C, defined by inversion of the elliptic equations in the drift-diffusion model (1.1-1.3). Hence, 0 is the only accumulation point of $\sigma[T_{C,u}(u*)]$. Furthermore, it has been proven for projection methods in Theorems **5.5** and **5.10** from [6] that in the limit as the meshwidth $h \to 0$ the spectrum $\sigma[T_u(u^*)]$ converges to $\sigma[T_{C,u}(u^*)]$. Hence, the ellipticity of the PDEs in the system (1.1-1.3) implies that for a projection method, the spectrum $\sigma[T_u(u^*)]$ of the derivative of the discretized mapping T clusters around 0 as $h \to 0$. The discrete model in our numerical computations was obtained by a finite difference scheme which can be analyzed as a relaxed finite element method as developed in [?].

The implication of the above analytical results is that only a few isolated eigenvalues of $I - T_u$ can be close to 0 in the complex plane. As discussed in §6 on numerical results, the discrete spectrum which we determined computationally agrees well with this theoretical result.

3 Matrix Free Methods and Eigenvalue Estimations

In order to accelerate the fixed point iteration associated with the mapping T, by a technique such as Chebyshev acceleration, we need to compute eigenvalue estimates of the Jacobian of the mapping T. However, although finite difference or finite element discretization of the partial differential equations of semiconductor simulation defining T (equations (1.1-1.3)) results in a sparse representation of the

simulation problem, the Jacobian of the mapping T is dense. Therefore explicit calculations with, or representation of, this Jacobian is not practical. We are faced with the problem of computing eigenvalue estimates without computing the Jacobian matrix explicitly.

Many of the techniques for computing eigenvalue estimates only require that the user supplies a routine for performing a matrix by vector multiplication. An observation that has proved very useful principally in Ordinary Differential Equations methods [10, 3, 4, 2] is that the product of the Jacobian T_u by a vector v can be approximated by the difference formula

$$T_u v \approx \frac{T(u + \epsilon v) - T(u)}{\epsilon}, \tag{3.1}$$

where ϵ is a carefully chosen scalar. This approach has also been used by Eriksson and Rizzi to compute eigenvalues of operators arising in fluid dynamics [9].

Within this class of methods based on matrix by vector multiplication, one can choose between using either the subspace iteration [14, 20], or the nonsymmetric Lanczos algorithm [7], or Arnoldi's method [1, 17, 18]. The nonsymmetric Lanczos algorithm must be excluded because it also requires computing the product of the transpose of the Jacobian times a vector and this cannot be approximated by a formula similar to (3.1). Subspace iteration does not compute the outermost eigenvalues of T_u but those of largest modulii. However, as explained in §4, we accelerate the iteration by determining polynomials p_n for which the maximum of $|p_n|$ is minimized over a convex region E containing all the eigenvalues. Therefore we need a method which allows us to provide approximations to all eigenvalues located in the outermost part of the spectrum, not just those of largest moduli.

As a result the only alternative left to us is to use Arnoldi's method or one of its variations [17]. This algorithm, which is described in Appendix B, normally delivers approximations to the outermost eigenvalues of a given matrix A, which is precisely what is desired here. It has been successfully used to provide estimates of the outer spectrum of the Jacobian T_u. The matrix-vector multiplication in Arnoldi's method was performed with the help of (3.1) to avoid manipulating the Jacobian matrix explicitly. The coefficient ϵ in (3.1) was determined adaptively by estimating the norm of T_u and ensuring that the numerator in (3.1) is computed to within at least the square root of the machine epsilon relatively to the size of the term $T(u)$. It can be observed from the plots in Figure 2 (details

are explained in the beginning of §6 on numerical experiments) that the eigenvalue estimates compare well with the theoretical values of [**?**].

If Newton's method is employed, the formula (3.1) can be applied successfully as well for the solution of the linear sytems at each iteration by a conjugate gradient type method. This will be exploited in §5.

4 Second Order Acceleration

General acceleration techniques for speeding up convergence of fixed point iterations consist of forming a sequence of iterates which are linear combinations of the iterates in the original sequence [11]. Second order acceleration restricts the linear combination to satisfy a three-term recurrence of the form,

$$u_{n+1} = \rho_{n+1}[\gamma T(u_n) + (1-\gamma)u_n] + (1-\rho_{n+1})u_{n-1}. \tag{4.1}$$

The coefficients $\{\rho_n\}$ and γ are typically chosen so that in the linear case the residual vector $u_n - T(u_n)$ for the transformed iteration is much smaller than for the original one. The standard acceleration techniques used in iterative methods for solving linear systems can be adapted to accelerating any fixed point iteration and should be expected to be successful provided we are close enough to the solution that the regime of the original iteration is nearly linear [**?**]. The case of Chebyshev acceleration is described first.

4.1 Chebyshev acceleration

In what follows we recall the main ideas of Chebyshev acceleration. For a complete description see [**?**] or [16].

Initially, let us assume that the mapping T is linear. In this situation it can be seen that in an acceleration scheme of the form defined above, the fixed point residual vector $r_n = u_n - T(u_n)$ is of the form $r_n = p_n(T)r_0$ where p_n is a polynomial of degree n satisfying the consistency condition $p_n(1) = 1$. Hence, in the case where the operator T is diagonalizable the eigenexpansion $r_0 = \sum_{i=1}^{N} \alpha_i z_i$ is transformed into $r_n = \sum_{i=1}^{N} \alpha_i p(\lambda_i) z_i$. One would want to make the expansion coefficients $|\alpha_i p(\lambda_i)|$ as small as possible, by attempting to minimize the discrete norm $\max_{i=1,\ldots,N} |p_n(\lambda_i)|$ over all polynomials

of degree n satisfying the condition $p_n(1) = 1$, for example. However, this is difficult to implement in practice if only because it requires the knowledge of the whole spectrum. On the other hand, if we know that the spectrum of T is contained in some continuous region E, then the discrete norm can be replaced by the infinity norm on that region. For an ellipse E in the complex plane with real center d, semi-focal distance c, semi major axis a and semi minor axis b the normalized Chebyshev polynomial

$$p_n(z) \equiv \frac{C_n[(z-d)/c]}{C_n[(1-d)/c]}$$

is the unique polynomial of degree n which satisfies the consistency condition $p_n(1) = 1$ and for which

$$\max_{z \in E} |p_n(z)| \tag{4.2}$$

is minimal ([?] pp. 333-334.) Therefore, to obtain the coefficients γ and ρ_n corresponding to Chebyshev acceleration it suffices to compare the three-term recurrence for the residual vectors associated with the iteration (4.1) with that of the polynomial (4.2). It is easily found that

$$\gamma = \frac{1}{1-d} \tag{4.3}$$

$$\rho_n = 2\left(\frac{1-d}{c}\right)\frac{C_{n-1}[(1-d)/c]}{C_n[(1-d)/c]} \tag{4.4}$$

where C_k represents the Chebyshev polynomial of degree k of the first kind. In fact a simple relation links two successive values of ρ_n for $n > 1$ namely,

$$\rho_{n+1} = \frac{1}{1 - (\sigma^2/4)\rho_n} \tag{4.5}$$

in which $\sigma = c/(1-d)$.

The convergence rate of the iteration is determined by the rate of decrease to zero of the maximum modulus of p_n. It is known that the maximum modulus of p_n is reached at the point $z = d + a$ and therefore

$$\max_{z \in E} |p_n(z)| = \frac{C_(n)[a/c]}{C_n[(1-d)/c]} \approx \left(\frac{a/c + \sqrt{(a/c)^2 - 1}}{(1-d)/c + \sqrt{((1-d)/c)^2 - 1}}\right)^2. \tag{4.6}$$

Defining $\sigma_a = c/a$, the convergence ratio can then be written as

$$\rho_{th} = \frac{\sigma}{\sigma_a} \times \frac{1 + \sqrt{1 - \sigma_a^2}}{1 + \sqrt{1 - \sigma^2}}. \tag{4.7}$$

By a simple manipulation we can also recast the above formula in the following form:

$$\rho_{th}^2 = \left(\frac{1+\sqrt{1-\sigma_a^2}}{1-\sqrt{1-\sigma_a^2}}\right)\left(\frac{1+\sqrt{1-\sigma^2}}{1-\sqrt{1-\sigma^2}}\right)^{-1}. \tag{4.8}$$

We proceed from the linear case above to mappings T which are nonlinear. Because we employ the acceleration scheme only after rapid convergence has slowed down and we are close to the solution, we can consider linear models locally around the solution u^* to derive the correct acceleration schemes. If we let u^* be the solution to $T(u) = u$, and $du = u - u^*$, then, by the Fréchet differentiability of T,

$$|T(u) - T(u^*) - T_u(u^*)du| \leq C\epsilon|du|, \tag{4.9}$$

if $|du| \leq \delta$. Therefore, the three-term recurrence relation (4.1) and the triangle inequality imply

$$|du_{n+1}| \leq \rho_{n+1}\gamma C\epsilon_n|du_n| + |\rho_{n+1}[\gamma T_u(u^*)du_n + (1-\gamma)du_n] + (1-\rho_{n+1})du_{n-1}|. \tag{4.10}$$

In (4.10), the first term on the right hand side takes into account the nonlinearity of T, the second one is completely linearized and depends only on $T_u(u^*)$.

In (4.10) ϵ_n becomes arbitrarily small as n increases and $|u_n - u^*|$ decreases. Therefore, the Chebyshev acceleration of the nonlinear map employing the spectrum of the Jacobian $T_u(u^{(*)})$ at the fixed point should be a reasonable approach for n sufficiently large.

We now prove that the Chebyshev acceleration as defined by the parameters determined from the spectrum of the derivative $T_u[u^*]$ of T at the fixed point u^* converges locally for the nonlinear iteration.

Theorem 4.1 *Let u^* be the fixed point of the fixed point mapping T. Suppose that the sequence of iterates $\{u_n\}$ defined by the recursion $u_{n+1} = T[u_n]$ converges linearly to u^* with asymptotic convergence rate τ, then the accelerated sequence $\tilde{u}_n$ defined by the three term recurrence relation (4.1) with γ and the $\{\rho_n\}$ defined by (4.3) and (4.5) converges to the fixed point u^* with the same asymptotic convergence rate as would be obtained for the problem linearized around u^*.*

Proof. Chebyshev acceleration is defined through the three term recurrence relation (4.1) with γ and the $\{\rho_n\}$ defined through (4.3) and (4.5). This means that the Chebyshev accelerated scheme defines a fixed point mapping C_{nonl} from the vector $(u_{n-1}, u_n, \rho_{n+1}) \in \mathbb{R}^{2N+1}$ to its image $(u_n, u_{n+1}, \rho_{n+2})$.

Consider, however, the linear problem which is obtained from (4.1) by replacing the nonlinear mapping $T : u \to T(u)$ by its linear approximation at u^*. For this problem the nonlinear Chebyshev iteration is replaced by the linear Chebyshev iteration derived from (4.10) by ignoring the term which contains ϵ_n on the right hand side. For the resulting linear iteration we can define a fixed point mapping C_{lin} from $\mathbb{R}^{2N+1}$ to itself similarly to C_{nonl}. The mapping C_{lin} is nonlinear only in ρ. Moreover, the parameters of the Chebyshev iteration are such that this mapping is contracting at the limit point $(0, 0, \rho_\infty)$. In fact, the theoretical asymptotic convergence rate of the optimal Chebyshev iteration associated with the ellipse E enclosing the spectrum $\sigma(T)$ is given by the formula (4.8) which yields $\rho_{th} < 1$. The Jacobian of the mapping C_{lin} consists of a block matrix where the blocks are associated to the three variables u_{n+1}, u_n and ρ_{n+1}.

$$J = \begin{pmatrix} \frac{\partial u_{n+1}}{\partial u_n} & \frac{\partial u_{n+1}}{\partial u_{n-1}} & 0 \\ \frac{\partial u_n}{\partial u_n} & \frac{\partial u_n}{\partial u_{n-1}} & \frac{\partial u_n}{\partial \rho_n} \\ 0 & 0 & \frac{d\rho_{n+1}}{d\rho_n} \end{pmatrix},$$

and, substituting the derivatives,

$$J = \begin{pmatrix} \rho_{n+1}\gamma T_u(u_n) + (1-\gamma) & 1-\rho_{n+1} & 0 \\ I & \rho_n \gamma T_u(u_{n-1}) + (1-\gamma) & \gamma T_u(u_{n-1}) + (1-\gamma)u_{n-1} - u_n \\ 0 & 0 & (1-\sqrt{1-\sigma^2})/(1+\sqrt{1-\sigma^2}) \end{pmatrix}. \quad (4.11)$$

Observe that the last row of this matrix is zero except for its $(2N+1)\times(2N+1)$ element and as a result its spectrum is the union of the spectrum of the $(2N) \times (2N)$ upper submatrix and this bottom right element. The spectral radius of the upper submatrix is equal to ρ_{th} in (4.8). The bottom right element is equal to the second factor of (4.8) and is therefore not larger than ρ_{th}^2. Therefore, $\rho[C_{lin,u}] = \rho_{th} < 1$ at the fixed point $(0, 0, \rho_\infty)$. By definition the spectrum $\sigma[C_{nonl,u}]$ of the derivative of C_{nonl} at the fixed point $(u^{(*)}, u^{(*)}, \rho_\infty)$ is equal to the spectrum of the derivative $C_{lin,u}$. Therefore, at the fixed point $\rho[C_{nonl,u}] = \rho_{th} < 1$, which establishes local convergence of the Chebyshev accelerated nonlinear fixed point iteration with the same rate as for the corresponding linear problem through Theorem 2.1.

□

The above result was proved in the more general context of k-step methods in [**?**]. The above proof differs from that of Theorem 3 of [**?**] and can viewed as simplification of it for the specific case of Chebyshev acceleration.

The optimal $\{\rho_n\}$ and γ are determined from an ellipse E that contains the spectrum $\sigma[T_u(u^*)]$ of the derivative $T_u(u^*)$. To determine the ellipse E, we need estimates of the eigenvalues of the Jacobian. For the cases where the largest eigenvalues are close to 1 the eigenvalues that we found are by and large located in an ellipse in the complex plane. We employed the experimentally determined spectra to accelerate the convergence of the decoupling approach as discussed in [**?**]. The three-term recurrence (4.1) is then used with the coefficients $\{\rho_n\}$ and γ as determined by (4.3) and (4.4) in which the parameters d, c of the best ellipse E are known. The theoretical convergence rate of the optimal Chebyshev acceleration of the iteration defined through a mapping T for which the spectrum $\sigma(T) \subset E$ is still given by (4.8). As a result, it is possible to compare the observed rate of convergence with the theoretical one given by this formula.

4.2 Stationary Second Order Richardson Acceleration

It can be observed that the sequence of Chebyshev acceleration coefficients ρ_n converges as n tends to infinity. Hence, the idea of replacing the variable coefficient ρ_n in (4.1) by the constant coefficient ρ_∞ the limit of the sequence $\{\rho_n\}$. This scheme is referred to as the second order stationary Richardson acceleration process. It is straightforward from (4.5) that

$$\rho_\infty = \frac{2}{1 + \sqrt{1 - \sigma^2}} \tag{4.12}$$

In a sense the iteration resulting from this substitution seems a little more natural than the Chebyshev iteration because it leads to a transformed iteration of the form

$$u_{n+1} = \rho[\gamma T(u_n) + (1 - \gamma)u_n] + (1 - \rho)u_{n-1}$$

which is a simple linear transformation of the original fixed point iteration, known to converge to the same point as the original sequence when it converges. We were initially lead to consider this scheme because the Chebyshev iterates tend to increase the residuals at the beginning. The initial residuals for the stationary process do indeed increase less but this does not result in overall faster convergence. In fact, the behavior of the two schemes is rather similar as is illustrated in §6 on numerical experiments. Local convergence for this stationary nonlinear second order scheme follows by the same kind of argument as for the Chebyshev acceleration. The only difference is that we only

have to consider a fixed point mapping from $\mathbb{R}^{2N}$ to itself this time because ρ is fixed. This is stated in the following corollary.

Corollary 4.1 *Let u^* be the fixed point of the fixed point mapping T. Suppose that the sequence of iterates $\{u_n\}$ defined by the recursion $u_{n+1} = T[u_n]$ converges linearly to u^* with asymptotic convergence rate τ, then the accelerated sequence $\tilde{u}_n$ defined by the three term recurrence relation (4.1) with γ and the $\{\rho_n\}$ defined by (4.3) and (4.12) converges to the fixed point u^* with the same asymptotic convergence rate as would be obtained for the problem linearized at u^*.*

Proof. The proof is very similar to that of Theorem 4.1. The difference is that in this case the ρ_n are not defined recursively, but rather equal to a constant. As a result the last row and column in the matrix (4.11) are absent. The removal of this part of the matrix only simplifies the argument.

□

5 GMRES Acceleration.

GMRES is an algorithm developed in [19] for solving large nonsymmetric linear systems of equations. The algorithm is based on Arnoldi's process, described in Appendix B. More precisely, let us assume that we need to solve the linear system $Ax = b$, where A is an $N \times N$ nonsingular matrix. If we take $v_1 = b/\|b\|_2$ and run Arnoldi's procedure, then, at each step j, we obtain an $N \times j$ matrix V_j whose column vectors form an orthonormal basis of the Krylov subspace $K_j = span\{v_1, Av_1, ..., A^{j-1}v_1\}$. Moreover, at a given step m,

$$AV_m = V_{m+1}\bar{H}_m, \tag{5.1}$$

where the matrix $\bar{H}_m$ is an $(m+1) \times m$ Hessenberg matrix. Assume that we want to obtain the vector of $K_m = \text{span}\{v_1, Av_1, ..., A^{m-1}v_1\}$ which has the smallest residual norm for the linear system $Ax = b$. In other words we seek a vector of the form

$$x_m = V_m y,$$

with $y \in R^m$ which minimizes

$$\phi(y) \equiv \|b - AV_m y\|_2 = \|\beta v_1 - AV_m y\|_2,$$

where we have set $\beta \equiv \|b\|_2$. We can write

$$\begin{aligned} \phi(y) &= \|\beta v_1 - AV_m y\|_2 = \|V_{m+1}\beta e_1 - V_{m+1}\bar{H}_m y\|_2 \\ &= \|V_{m+1}[\beta e_1 - \bar{H}_m y]\|_2 = \|\beta e_1 - \bar{H}_m y\|_2. \end{aligned} \tag{5.2}$$

Here we have used the orthonormality of the vectors $v_i, i = 1, ..., m+1$. Thus, the best approximation to the solution from the Krylov subspace, in the sense of residual norms, can be obtained by solving a simple $(m+1) \times m$ least-squares problem. Moreover, it is easy to determine the residual norm of the solution that would be obtained at any step without actually computing it, provided the QR factorization of $\bar{H}_m$ is done progressively [19]. This may be used to prevent the algorithm from taking more steps than are required to satisfy the stopping criterion. The algorithm is typically used with a form of restarting which means that after an outer iteration of m steps is performed the algorithm is restarted as if we were to solve the new linear system for δ $A(x_m + \delta) = b$, i.e., the correction δ is obtained as the solution of $A\delta = r_m \equiv b - Ax_m$.

We can also define a NonLinear Generalized Minimal Residual method (NLGMR) to solve a system of nonlinear equations $F(u) = 0$ in the following way. Given the current iterate u_n, we seek a new iterate $u_{n+1} = u_n + \delta_n$, where δ_n is in some subspace to be defined shortly. Ideally we wish to minimize $\|F(u_n + \delta_n)\|_2$ over that subspace. Since this is a nonlinear optimization problem, it is easier to provide for an approximate solution by first linearizing it and seeking to minimize instead $\|F(u_n) + J_n\delta_n\|_2$ where J_n is the Jacobian matrix of F at the point u_n. If we were able to solve the linear system $J\delta_n = -F(u_n)$ exactly in the selected subspace this would just lead to a standard Newton step. A natural idea is to take an approximate Newton step which corresponds to selecting the Krylov subspace $K_m = \{v_1, J_n v_1, \cdots, J_n^{m-1} v_1\}$ where J_n is the Jacobian matrix of F at the point u_n and $v_1 = -F(u_n)/\|F(u_n)\|_2$. This amounts to an inexact Newton iteration where each linear system is approximated by m steps of the (linear) GMRES algorithm. In Appendix C we list one implementation of NLGMR based on these principles.

The idea of this algorithm has been used by Wigton et al. [23] and more recently in [**?**]. The backtracking strategy used in our code is based on a local quadratic model as described in [**?**]. Though simple, this technique seems sufficient to ensure global convergence and improve convergence properties

away from the solution. Another possibility explored in [**?**] is the more elaborate trust-region technique. We have not implemented this approach because the initial guess provided to NLGMR seems relatively close to the solution. Another detail worth mentioning about our implementation is that the maximum number m of steps in each iteration (2) is varied according to the level of nonlinearity as determined by the backtracking routine. When the process undergoes a severe backtracking this indicates that we are in a highly nonlinear region and therefore it is wasteful to take a large m. A value of m equal to 1 would simply correspond to a form of steepest descent step (in the linear case) and may be sufficient until we get closer to the solution.

5.1 Effect of clustering on the convergence of conjugate gradient algorithms

In this §, we examine the effect of the clustering at 1 of the eigenvalues of a linear mapping A on the asymptotic convergence rate of an algorithm $\mathcal{A}$ which is based on minimizing the residual norm over expanding Krylov subspaces. We will denote by $D(1,\epsilon)$ the disk of radius ϵ, centered at 1 in the complex plane. Let $\mathcal{A}$ be the algorithm which generates in k steps the approximation x_k which minimizes the 2-norm of the residual $\|b - Ax_k\|$ over the Krylov subspace $K_k \equiv span\{v_1, Av_1, \cdots, A^k v_1\}$. The following theorem implies that for a linear mapping for which the eigenvalues cluster suitably at 1 the convergence rate of this algorithm $\mathcal{A}$ will be superlinear.

Theorem 5.1 *Let $A : v \mapsto Av$ be a nonsingular, diagonalizable linear mapping with eigenvalues λ_i. Assume that for any $\epsilon > 0$ the number $M(\epsilon)$ of eigenvalues λ_i outside $D(1,\epsilon)$ is given by*

$$M(\epsilon) \leq K\epsilon^{-r}. \tag{5.3}$$

Consider an algorithm $\mathcal{A}$ which minimizes the residual over the Krylov subspace K_k at every step. Then the residual vector r_k at the k-th step satisfies

$$\|r_k\|_2 \leq c_k \left[\frac{K\rho(r+1)}{(1-\rho)k}\right]^{k/(r+1)} \tag{5.4}$$

where $c_k \to c$ as $k \to \infty$.

Proof. We start by ordering the eigenvalues λ_i of A such that $|\lambda_i - 1| \geq |\lambda_j - 1|$ if $i < j$. Hence, $|\lambda_i - 1|$ decreases monotonically with increasing i. Because the algorithm $\mathcal{A}$ minimizes the residual

norm over expanding Krylov subspaces,

$$\|r_k\|_2 = \min_{p \in P_k \ , p(0)=1} \|p(A) r_0\|_2 \tag{5.5}$$

where P_k is the space of polynomials of degree $\leq k$. Thus, for any polynomial p which satisfies the consistency condition $p(0) = 1$, the residual satisfies $\|r_k\|_2 \leq \|p(A) r_0\|_2$. As is usually done, we use the eigenbasis, i.e., A is expressed as $A = X \Lambda X^{-1}$ where Λ is the diagonal of eigenvalues. For any polynomial $p \in P_k \ , p(0) = 1$, we find that

$$\|r_k\|_2 \leq \tau(X) \|r_0\|_2 \max_{\lambda \in \sigma(A)} |p(\lambda)| \tag{5.6}$$

where $\tau(X)$ represents the spectral condition number of X.

A particular polynomial is constructed as follows: for each k we partition the spectrum $\sigma[A]$ into the subset of the first n_1 (here $n_1 < k$) eigenvalues λ_i which are furthest away from 1 and the subset of the other elements of $\sigma[A]$. The polynomial

$$p_k(z) = [1 - z]^{k - n_1} \prod_{i=1}^{n_1} \frac{\lambda_i - z}{\lambda_i}. \tag{5.7}$$

has n_1 roots r_i at the first n_1 eigenvalues λ_i and satisfies the consistency condition $p_k(0) = 1$. Moreover, (5.3) implies that all but the first n_1 eigenvalues in the spectrum of A are inside the circle or radius not exceeding $D(1, (K/n_1)^{1/r})$. Because $p_k(\lambda_i) = 0$ for $i = 1, \cdots, n_1$ the polynomial $p_k(z)$ is nonzero only on the part of the spectrum $\sigma[A]$ of A which is inside the circle $D(1, (K/n_1)^{1/r})$. Hence

$$\max_{\lambda_i \in \sigma[A]} |p_k(z)| \leq \max_{z \ : \ |z-1| \leq (K/n_1)^{1/r}} |p_k(z)|.$$

Within the circle $D(1, (K/n_1)^{1/r})$ the first factor in (5.7) satisfies $|[1 - z]^{k - n_1}| \leq (K/n_1)^{(k-n_1)/r}$.

The second factor in (5.7) is estimated as follows: First we choose a suitable $\rho < 1$. Then (5.3) implies that all but the first l eigenvalues λ_i (where $l = K \rho^{-r}$) are located in the disk $D(1, \rho)$. However, because we have assumed that the mapping A in nonsingular, $\lambda_i \neq 0 \forall i$. This implies that for $n_1 > l$ we can write

$$\prod_{i=1}^{n_1} \frac{\lambda_i - z}{\lambda_i} = \prod_{i=1}^{l} \frac{\lambda_i - z}{\lambda_i} \prod_{i=l+1}^{n_1} \frac{\lambda_i - z}{\lambda_i},$$

where the constant factor $F_l \equiv \prod_{i=1}^{l} \frac{\lambda_i - z}{\lambda_i}$ is finite. Using the triangle inequality, the numerator in each factor in the product $\prod_{i=l+1}^{n_1} \frac{\lambda_i - z}{\lambda_i}$ can be bounded as $|\lambda_i - z| \leq |\lambda_i - 1| + |1 - z|$. The denominator can be bounded as $|\lambda_i| \geq 1 - |\lambda_i - 1|$. Therefore, since $|1 - z| \leq (K/n_1)^{1/r}$, we get

$$\max_{\lambda_i \in \sigma[A]} |p_k(z)| \leq (K/n_1)^{(n-n_1)/r} F_l \prod_{i=l+1}^{n_1} \frac{|\lambda_i - 1| + (K/n_1)^{1/r}}{1 - |\lambda_i - 1|} \tag{5.8}$$

We can set n_1 (the number of eigenvalues inside $D(1, K/n_1)$) equal to $\lceil \alpha k \rceil$ (the smallest integer at least as big as αk) for $0 < \alpha < 1$. Then using the fact that for $i \geq l$ $|1 - \lambda_i| \leq \rho < 1$, (5.8) implies

$$\max_{\lambda_i \in \sigma[A]} |p_k(z)| \leq [K/\lceil \alpha k \rceil]^{(k - \lceil \alpha k \rceil)/r} F_l \left[\frac{\rho + [K/\lceil \alpha k \rceil]^{1/r}}{1 - \rho} \right]^{\lceil \alpha k \rceil - l}$$

Since $K/\lceil \alpha k \rceil]^{1/r}$ tends to zero as k tends to ∞, the term in the brackets in the right-hand-side is asymptotically equivalent to $[\rho/(1-\rho)]^{\alpha k - l}$. Similarly, the first term is clearly asymptotically equivalent to $[K/(\alpha k)]^{(1-\alpha)k/r}$. Therefore there is a positive sequence s_k which converges to 1 such that

$$\max_{\lambda_i \in \sigma[A]} |p_k(z)| \leq s_k F_l [K/(\alpha k)]^{(1-\alpha)k/r} [\rho/(1-\rho)]^{\alpha k - l}$$

We now take $\alpha = 1/(r+1)$ to obtain,

$$\max_{\lambda_i \in \sigma[A]} |p_k(z)| \leq F_l s_k \left[\frac{K\rho(r+1)}{(1-\rho)k} \right]^{k/(r+1) - l}$$

The proof follows immediately by combining this result with (5.6) and defining

$$c_k = s_k F_l \left[\frac{K\rho(r+1)}{(1-\rho)k} \right]^{-l} \tau(X) \|r_0\|_2.$$

□

Note that for small values of k the term inside the brackets of (5.4) is likely to be larger than one. As k tends to infinity, this convergence factor converges to zero, indicating a superlinear convergence. We also remark that the choice $\alpha = 1/(r+1)$ in the proof is somewhat arbitrary and is by no means optimal. It was made to yield a simple bound. Ideally, we would wish to minimize the asymptotic convergence rate

$$\eta(\alpha) = (1-\alpha)/r \log[K/(\alpha k)] + \alpha \log[\rho/(1-\rho)]$$

Physical Constants and Device Parameters	
parameter	numerical value
mobility	$\mu = 820 \frac{cm^2}{Vsec}$
intrinsic density	$n_i = 1.4 * 10^{10} cm^{-3}$
dielectric constant Si	$\epsilon_{Si} = 11.7$
dielectric constant Ox	$\epsilon_{Ox} = 3.78$
temperature	300 Kelvin
background doping	$3 * 10^{15} \mathrm{cm}^{-3}$
thickness oxide	250 Ångstrom
radius doping profile	.21 μ
length device	3 μ
depth device	2.8 μ
width source	.2 μ

Table 1: Physical parameters for test problem.

which would yield an optimal α that depends on k. In fact the theorem can be stated with an arbitrary value of α and the bound (5.4) can be replaced by

$$\|r_k\|_2 \leq c_k [K/(\alpha k)]^{(1-\alpha)k/r} [\rho/(1-\rho)]^{\alpha k}$$

for any α such that $0 < \alpha < 1$, still showing superlinear convergence.

6 Numerical Experiments.

Numerical tests were performed on an IBM 4381 computer in double precision arithmetic. The mapping T was defined by the successive solution of (1.1-1.3) for a semiconductor model as specified at the end of Appendix A. The depicted ellipse was employed to determine the parameters of both second order accelerations. Theoretically, the eigenvalues are expected to lie within the small circle centered at .5 . Further parameters are specified in Table 1. The derivative T_u of the fixed point mapping T is taken at the solution u^*.

We plotted the convergence of the iterates $\{u_n\}$ generated by the fixed point mapping T in terms of the L_∞ norm of the fixed point residual $r_n \equiv u_n - T(u_n)$ from §4.1. This residual is equal to the negative stepsize $du_n \equiv u_{n+1} - u_n$. The resulting accelerations are depicted in Figure 3 for the eigenvalues and the ellipse shown in Figure 2. The convergence of the accelerated schemes is linear, possibly slightly superlinear for NLGMR. In the case of Chebyshev acceleration the residual decreased

by a factor $3 * 10^{-4}$ in 100 iterations which coincides with a convergence factor $\rho_{ex} = .922$. The geometrical data for the ellipse in Figure 2 are $d = .475,\ c = .475,\ a = .505$. Using that $b = \sqrt{a^2 - c^2}$ and entering these data in the the theoretical expression (4.8) yields $\rho_{th} = .9036$. This discrepancy is probably due to a combination of two factors. First the convergence rate is sensitive to the variation of the parameters in expression (4.8). For example, for a slightly larger ellipse with $a = .509$ we obtain $\rho_{th} = .924$. Thus, a small error in the eigenvalue estimates may lead to a much larger variation in the actual convergence rate. The second factor is obviously that the actual iteration is nonlinear and we are using a linear model for it. The original (unaccelerated) scheme decreased the residual by a factor 10^{-1} in 100 iterations. Hence, the Chebyshev acceleration increased the speed of convergence by a factor of almost four. The observed convergence rate for the second order Richardson acceleration in Figure 3 shows hardly any difference with that of the Chebyshev acceleration.

The NLGMR algorithm was implemented with an adaptively expanding subspace over which Newton's equations are solved. We started with $m = 2$ in the algorithm presented in Appendix C and we doubled the size of the subspace up to a maximum of 25 whenever we found that the residual for the linearized equations was within a factor of 1.5 of the nonlinear residual. The size of the subspace was kept unchanged if the nonlinear residual was in between 1.5 and 5 times the linear residual. Otherwise the size of the subspace was halved.

The acceleration by the nonlinear version of GMRES outperforms the two previous accelerations considerably. In this case the residual decreased in 100 iterations by a factor $2 * 10^{-8}$ which corresponds to a linear convergence factor $\rho_{ex} = .838$. This substantial superiority over the results with the two second order recurrences can be attributed to the capacity of GMRES to take advantage of the clustering of the spectrum $\sigma[T_u]$ around the origin. As discussed in §2, this property of $\sigma[T_u]$ can be related to the compact differentiability of the continuous mapping T_C. Like other conjugate gradient type methods, GMRES is able to take advantage of a spectrum in which the rate of convergence is slowed down by a few isolated eigenvalues only. Thus the nonlinear version of GMRES can take advantage of exactly that characteristic of the spectrum which is due to the compact differentiability of T_C. The compact differentiability of T_C has no implications, on the other hand, for the overall

convex hull of the spectrum on which the spectrum of the second order recurrences depends.

Another important point is that the extra overhead required by NLGMR as compared to the smaller overhead incurred in the second order recurrences is almost negligible relatively to the expense in computing function evaluations Tu. Moreover, GMRES does not require the determination of the spectrum of the derivative which is rather costly in itself.

Nevertheless, efficient implementation of NLGMR for the acceleration of this problem requires careful choice of the parameters inherent to this algorithm. Most notably the finite difference approximations to the Jacobian must be done carefully and the size of the subspace over which we choose to solve Newton's equations must be monitored dynamically as was explained in §5.

7 Conclusion

We have shown that a nonlinear version of GMRES, called NLGMR, provides an effective means of accelerating the speed of convergence of a nonlinear fixed point mapping T, where T is defined through the successive inversion of a system of coupled nonlinear *elliptic* PDE's. In numerical tests on a fixed point mapping T, representing the nonlinear block Gauß-Seidel decoupling algorithm from semiconductor simulation, we have compared the performance of NLGMR with that of two second order recurrences. The observed convergence rates for the second order accelerations agree well with the theoretically predicted bounds for this approach. NLGMR is found to be substantially more effective in accelerating the fixed point iteration than the second order recurrences. Moreover, NLGMR does not require the determination of the spectrum of the derivative of the nonlinear mapping T which defines the nonlinear iteration. We have argued that the superior performance of NLGMR for this problem must be due to its capacity to exploit the clustering of the eigenvalue distribution of T_u around 0. This clustering is again related to the compact differentiability of the original mapping T_C and is valid for fixed point mappings T which involve solution of similar elliptic PDE's. Finally, we would like to emphasize the ease with which a simple code such as NLGMR can be implemented to speed-up a relaxation method for the solution of nonlinear elliptic systems.

A The Semiconductor Model

Mathematical analysis of the convergence of the semiconductor decoupling algorithm was first presented by Mock in [?]. Our formalism follows closely that exposed in [?, ?, ?], and [?]. The mathematical framework is based on the work by Jerome in [?, ?] and Seidman in [?].

In the drift-diffusion model (1.1-1.3) k_1 is the doping profile in units of the intrinsic density of the semiconductor. Assumed are a constant mobility, zero generation and recombination and that Einstein's relations are valid. Further, the physical density of the negatively charged conduction electrons n and the density of the positively charged holes p are expressed in terms of the "quasi-Fermi levels" v and w by $n = \exp(u - v)$ and $p = \exp(w - u)$. One iteration of the decoupling algorithm can be briefly described as follows. Given the dimensionless potential u from the previous iteration, we compute $\tilde{u} \equiv T_C(u)$ by solving the discretized equations obtained from the system (1.1-1.3). More precisely, application of the mapping $T_C : u \to \tilde{u}$ consists of first either solving one of the current continuity equations (1.1) or (1.2) for the intermediate variables v or w, or the solution of both equations (1.1) and (1.2) for v and w. Next, the potential equation (1.3) is solved for the image $\tilde{u}$. Iteration with the mapping T_C corresponds to the nonlinear block Gauß-Seidel decoupling algorithm in semiconductor simulation, otherwise known as "Gummel's method" [?].

Existence of a fixed point of the mapping T_C, which corresponds to a simultaneous solution of the equations in (1.1-1.3), follows from maximum principles and Schauder's fixed point theorem. For finite dimensional models Brouwer's Fixed Point theorem suffices.

We quote some of the main theoretical results which are derived in [?] for the eigenvalues of the derivatives of the continuous mapping T_C, and its submaps U from v and w to u, V from u to v and W from u to w.

Theorem A.1 *For the derivative V_u of the mapping V, defined implicitly by the current continuity equation for the electrons*

- *The eigenvalues λ_n are given by*

$$\lambda_n = \frac{1}{1 - ia_n}$$

where $a_n \neq 0$ and real. These λ_n are located on the circle in the complex plane, centered at $\frac{1}{2}$ and with radius $\frac{1}{2}$. However, 1 is not an eigenvalue.

- *The eigenfunctions ν_n are orthogonal with respect to the inner product*

$$\langle \nu, \omega \rangle \equiv \int_0^L \exp(u-v) \nabla\nu \cdot \nabla\omega^* dx.$$

The result for the mapping W is basically identical with minor modifications.

As mentioned in §2, the eigenpairs of the derivative T_u of the discretized mapping T converge to the eigenpairs of the derivative $T_{C,u}$ of the mapping T_C in the limit that the meshwidth $h \to 0$. In [**?**] it is shown that for finite meshwidth h eigenvalues of the mappings V and W, defined implicitly through the current continuity equations, are on the same circle in the complex plane as for the continuous problem if exponential upwinding is employed in the discretization of the drift-diffusion model. The eigenvalues are shown to be located in the interior of this circle if the current continuity equations are discretized employing the scheme of Scharfetter and Gummel [**?**].

We can show significantly more for the case of constant doping k_1 in one dimension. The mapping $T_C : u \to \tilde{u}$ is then defined by the system of equations

$$\begin{aligned} -[\exp(u-v)v_x]_x &= 0, \\ -[\exp(w-u)w_x]_x &= 0, \\ -\tilde{u}_{xx} + \exp(\tilde{u}-v) - \exp(w-\tilde{u}) - k_1 &= 0, \end{aligned} \tag{A.1}$$

on $[0, L]$, subject to Dirichlet boundary conditions.

With $C = \sinh^{-1}(K_1)$ the solution can be written $v = w = u - C$. The electron density $n = \exp(C)$, the hole density $p = \exp(-C)$. The boundary conditions are given by $u(0) = 0$, and $u(L) = VB$. In this case, the spectral radius ρ can be bounded in terms of the densities n and p at the solution and the device length as stated in the following theorem

Theorem A.2 *For the mapping $T_C : u \to \tilde{u}$, implicitly defined by the system of one dimensional two point boundary value problems (A.1) on a domain $\Omega \equiv [0, L]$, the spectral radius ρ of the derivative*

$T_{C,u}$ at the solution is bounded from above

$$\rho \leq \frac{N+P}{(\pi/L)^2+N+P}\frac{1}{\sqrt{1+(2\pi/VB)^2}}.$$

Hence, by Theorem 2.1, Gummel's method converges locally for arbitrary variation of the "bias potentials" to the solution of the corresponding system of boundary value problems.

The theorem establishes convergence for arbitrary variation of the "bias potentials." The resulting bound on the spectral radius depends on the length of the device because of Poisson's equation for the potential.

For the computations presented in this paper we employed a two dimensional model of an N type mosfet. The geometry of the device is presented in Figure 1. For the numerically computed spectrum presented in Figure 2 the applied bias potentials are mentioned in the caption: We chose the potential at the backgate as our reference at 0 Volt, the source potential at .5 Volt, the gate potential at 6.5 Volt, and the drain potential at 6.5 Volt. As mentioned we examined the mapping T which maps electrostatic potentials u to themselves, and both the current continuity equations for n and p were solved.

The equations were discretized in the way of Scharfetter and Gummel using standard charge conserving finite differences as described in [?] or [?]. The mesh was rectangular with 20 lines parallel to the interface and 27 lines perpendicular to it. In the finite difference formulation for the equation for the potential u the nonlinear terms were "lumped" (see [21]) on the diagonal of the finite difference matrix. (The lumping procedure is second order accurate and hence of the same accuracy as the discretization scheme for Poisson's equation.)

B Arnoldi's algorithm

Arnoldi's algorithm simply implements a projection technique (Galerkin) over the so-called Krylov subspace $K_m = \text{span}\{v_1, Av_1, .., A_1^{m-1}v_1\}$ and can be described as follows.

Algorithm

- 1. Start: Choose a dimension m and an initial vector v_1.
- 2. Iterate: For $j = 1, 2, ..., m$ do

$$\begin{aligned} w &:= Av_j && \text{(B.2)} \\ h_{i,j} &:= (w, v_j), i = 1, ..., j && \text{(B.3)} \\ w &:= w - \sum_{i=1}^{j} h_{ij} v_i && \text{(B.4)} \\ h_{j+1,j} &:= \|w\|_2 \\ v_{j+1} &:= w/h_{j+1,j} \end{aligned}$$

In other words, at each step of the algorithm a new vector v_{j+1} is built by multiplying the previous vector v_j by the matrix A and orthonormalizing the result against all previous vectors. Note that the algorithm is described for exact arithmetic for simplicity, but in practice the standard Gram-Schmidt process (B.3), (B.4) should be replaced by a modified Gram-Schmidt process. After m steps of the algorithm one obtains an orthonormal basis of the Krylov subspace K_m and a Hessenberg matrix H_m whose nonzero entries are the coefficients h_{ij}. It is easy to see that this matrix is nothing but the representation of the projection of the linear operator A to the Krylov subspace. In fact, if we denote by V_m the $N \times m$ matrix whose column vectors are the v_i's, then

$$H_m = V_m^T A V_m.$$

This scheme was proposed by Arnoldi for transforming the matrix A into Hessenberg form by choosing $m = N$. Arnoldi also hinted that since this scheme is a projection type method some of the eigenvalues of A are well approximated even when m is much smaller than N. In [17] and [18], some analysis was proposed to suggest that the method delivers good approximations to those eigenvalues located in the outmost part of the spectrum.

C Algorithm NLGMR

The NLGMR algorithm for solving $F(u) = 0$ can be formulated as follows. The Jacobian at the current approximation u_n is denoted by J_n.

Algorithm : NonLinear Generalized Minimal Residual (NLGMR)

(1) *Start:* Choose u_0 and a dimension m of the Krylov subspace. Set $n = 0$.

(2) *Arnoldi process:*

- Compute $\beta = \|F(u_n)\|_2$ and $v_1 = -F(u_n)/\beta$.
- For $j = 1, 2, .., m$ do:

$$\begin{aligned} h_{i,j} &:= (J_n v_j, v_i), \quad i = 1, 2, ..., j, \\ \hat{v}_{j+1} &:= J_n v_j - \sum_{i=1}^{j} h_{i,j} v_i \\ h_{j+1,j} &:= \|\hat{v}_{j+1}\|_2, \quad \text{and} \\ v_{j+1} &:= \hat{v}_{j+1}/h_{j+1,j}. \end{aligned}$$

Define $\bar{H}_m$ to be the $(m+1) \times m$ (Hessenberg) matrix whose nonzero entries are the coefficients $h_{ij},\ 1 \le i \le j+1, \quad 1 \le j \le m$ and define $V_m \equiv [v_1, v_2, \dots v_m]$

(3) *Form the approximate solution:*

- Find the vector y_m which minimizes the function $\phi(y) \equiv \|\beta e_1 - \bar{H}_m y\|_2$, where $e_1 = [1, 0, \dots 0]^T$, among all vectors y of R^m.
- Compute $\delta_n = V_m y_m$ and $u_{n+1} = u_n + \delta_n$.

(4) *Backtrack:* Choose a damping scalar $\lambda_n \le 1$ such that $\|F(u_n + \lambda_n \delta_n)\|_2$ shows a sufficient decrease with respect to $\|F(u_n)\|_2$.

(5) *Restart:* If satisfied stop, else set $u_n \leftarrow u_{n+1}$, $n \leftarrow n+1$, and goto (2).

References

[1] W. E. Arnoldi, *The principle of minimized iteration in the solution of the matrix eigenvalue problem*, Quart. Appl. Math., 9 (1951), pp. 17–29.

[2] P. N. Brown, *A local convergence theory for combined inexact-Newton/ finite difference projection methods*, SIAM Journal on Numerical Analysis, 24 (1987), pp. 407–434.

[3] P. N. Brown and A. C. Hindmarsh, *Matrix-free methods for stiff systems of ODEs*, SIAM Journal on Numerical Analysis, 23 (1986), pp. 610–638.

[4] ——, *Reduced-storage matrix methods in stiff ODE systems*, Tech. Rep. UCLR-95088, Comp. and Math. Res. Div. , L-316, Lawrence Livermore Lab., Livermore Ca., 1986.

[5] P. N. Brown and Y. Saad, *Hybrid Krylov methods for nonlinear systems of equations*, SIAM J. Sci. Stat. Comp., 11 (1990), pp. 450–481.

[6] T. F. Chan and K. R. Jackson, *Nonlinearly preconditioned Krylov subspace methods for discrete Newton algorithms*, SIAM J. Stat. Scien. Comput., 7 (1984), pp. 533–542.

[7] F. Chatelin, *Spectral Approximation of Linear Operators*, Academic Press, New York, 1984.

[8] J. Cullum and R. Willoughby, *A Lanczos procedure for the modal analysis of very large nonsymmetric matrices*, in Proceedings of the 23rd Conference on Decision and Control, Las Vegas, 1984.

[9] R. S. Dembo, S. C. Eisenstat, and T. Steihaug, *Inexact Newton methods*, SIAM J. Numer. Anal., 18 (1982), pp. 400–408.

[10] L. E. Eriksson and A. Rizzi, *Analysis by computer of the convergence of discrete approximations to the euler equations*, in Proceedings of the 1983 AIAA conference, Denver 1983, Denver, 1983, AIAA, pp. 407–442.

[11] C. W. Gear and Y. Saad, *Iterative solution of linear equations in ode codes*, SIAM J. Sci. Statist. Comput., 4 (1983), pp. 583–601.

[12] A. L. Hageman and D. M. Young, *Applied Iterative Methods*, Academic Press, New York, 1981.

[13] J. H. Hyman and T. A. Manteuffel, *Dynamic acceleration of nonlinear processes*, in Elliptic Problem Solvers II, G. Birkhoff and A. Schoenstadt, eds., New York, 1984, Academic Press, pp. 301–313.

[14] A. Jennings and M. A. Ajiz, *Influence of the eigenvalue spectrum on the convergence rate of the conjugate gradient method*, Journal of the Institute of Mathematics and its Applications, 20 (1977), pp. 61–72.

[15] A. Jennings and W. J. Stewart, *A simultaneous iteration algorithm for real matrices*, ACM, Trans. of Math. Software, 7 (1981), pp. 184–198.

[16] D. G. Jesperson and P. G. Buning, *Accelerating an iterative process by explicit annihilation*, SIAM Journal on Scientific and Statistical Computing, 6 (1985), pp. 639–651.

[17] T. A. Manteuffel, *An iterative method for solving nonsymmetric linear systems with dynamic estimation of parameters*, Tech. Rep. UIUCDCS-75-758, University of Illinois at Urbana-Champaign, Urbana, Ill., 1975. Ph. D. dissertation.

[18] Y. Saad, *Variations on Arnoldi's method for computing eigenelements of large unsymmetric matrices*, Linear Algebra and its Applications, 34 (1980), pp. 269–295.

[19] ——, *Projection methods for solving large sparse eigenvalue problems*, in Matrix Pencils, proceedings, Pitea Havsbad, B. Kagstrom and A. Ruhe, eds., Berlin, 1982, University of Umea, Sweden, Springer Verlag, pp. 121–144. Lecture notes in Math. Series, Number 973.

[20] Y. Saad and M. H. Schultz, *GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM Journal on Scientific and Statistical Computing, 7 (1986), pp. 856–869.

[21] G. W. Stewart, *Simultaneous iteration for computing invariant subspaces of non-Hermitian matrices*, Numerische Mathematik, 25 (1976), pp. 123–136.

[22] G. Strang and G. J. Fix, *An analysis of the finite element method*, Prentice Hall, Englewood Cliffs, NJ, 1973.

[23] A. van der Sluis and H. A. van der Vorst, *The rate of convergence of conjugate gradients*, Numerische Mathematik, 48 (1986), pp. 543–560.

[24] L. B. Wigton, D. P. Yu, and N. J. Young, *GMRES acceleration of computational fluid dynamics codes*, in Proceedings of the 1985 AIAA conference, Denver 1985, Denver, 1985, AIAA.