PARALLELIZABLE RESTARTED ITERATIVE METHODS FOR NONSYMMETRIC LINEAR SYSTEMS. PART I: THEORY

WAYNE D. JOUBERT and GRAHAM F. CAREY

The University of Texas at Austin

Large sparse nonsymmetric problems of the form $Au = b$ are frequently solved using restarted conjugate gradient-type algorithms such as the popular GCR and GMRES algorithms. In this study we define a new class of algorithms which generate the same iterates as the standard GMRES algorithm but require as little as half of the computational expense. This performance improvement is obtained by using short economical three-term recurrences to replace the long recurrence used by GMRES. The new algorithms are shown to have good numerical properties in typical cases, and the new algorithms may be easily modified to be as numerically safe as standard GMRES. Numerical experiments with these algorithms are given in Part II, in which we demonstrate the improved performance of the new schemes on different computer architectures.

KEY WORDS: GMRES, iterative method, linear equations, nonsymmetric, parallel, cache, Orthomin.

C.R. CATEGORIES: G.1.3 Numerical Linear Algebra, D.1.3 Concurrent Programming, J.2 Physical Sciences and Engineering.

1 INTRODUCTION

In this paper we are concerned with iterative methods for solving the linear system

$$Au = b, \tag{1}$$

where $A \in \mathbb{C}^{N \times N}$ is a square nonsingular matrix which is typically large and sparse and may be non-Hermitian. When A is Hermitian and positive definite (HPD), variants of the standard *conjugate gradient* (CG) method ([Hestenes/Stiefel 1952], [Ashby/Manteuffel/Saylor 1990]) may be used to solve (1). On the other hand, it is well-known that when A is not Hermitian, the solution of (1) by iterative methods is considerably more problematic (see e.g. [Faber/Manteuffel 1984, 1987], [Joubert/Young 1987]). Specifically, the CG method for the HPD case has no

generalization for general non-Hermitian A which possesses all of the useful properties of the CG method for the HPD case.

Several classes of generalized conjugate gradient methods have been developed for non-Hermitian problems (for an overview, see e.g. [Joubert/Manteuffel 1990]). One of these classes is the class of *restarted methods*, typified by the popular GCR and GMRES algorithms. These methods are derived from the full *minimal residual* method, which minimizes the l^2-norm of the residual over a shifted Krylov space. *Restarted methods* repeatedly apply k steps of this method to the problem (1), where k, the *restart frequency*, is a parameter supplied to the algorithm. The effect of restarting the full minimal residual method is to reduce the buildup of an excessive number of basis vectors which are typically very costly to the algorithm.

The iterates defined by the restarted minimal residual method may be generated by any one of a number of different algorithms. A chief distinguishing property of these algorithms is the choice of the Krylov basis used. For example, in the GCR and GMRES algorithms, a Gram Schmidt technique is used to generate the basis vectors, to insure numerical stability. However, other choices for the Krylov basis may be used which are also numerically stable.

In this paper, we define a new algorithm which uses bases derived from classical orthogonal polynomials. These bases are not only typically numerically well-conditioned but also require only three-term recurrences rather than the long recurrences of Gram Schmidt procedures. For large values of k, this replacement of the GMRES basis generation sequence can eliminate up to half the computational work of the algorithm.

In addition to being faster on scalar and vector machines, this algorithm (as well as several related algorithms also defined in this paper) is also more adaptable to various alternate computer architectures, including parallel distributed memory machines and machines with cache memory. Thus an even greater computational saving is accrued. The parallel aspects and implementation of these algorithms are described in detail in Part II.

We now give a brief outline of the current paper. In Section 2 we set forth a general framework and basic theory of restarted methods. This includes a review of restarted iterative methods as well as a discussion of the choice of basis for the relevant Krylov space. After this, in Section 3 we give a theoretical discussion of the condition of Krylov bases, based on polynomial theory. Following this in Section 4 we define particular algorithms, including the new algorithms based on Chebyshev and other polynomials. Then in Section 5 we compare the numerical performance of the algorithms. This includes a comparison of the CPU requirements on scalar and vector machines and also a set of numerical examples to illustrate some of the numerical properties of the bases generated by the different algorithms.

2   BASIC THEORY OF RESTARTED METHODS

In this section we lay the basic framework for restarted minimal residual methods. We then set forth notation for the choice of basis for implementing such methods and list examples and practical considerations for choosing such a basis.

2.1 Basic Definitions and Formulas

Here we consider *iterative methods* for solving (1), defined by an initial iterate $u^{(0)}$ followed by subsequent iterates $\{u^{(n)}\}_{n \geq 1}$ which approximate the true solution $u = A^{-1}b$ to (1). We denote the residual by $r^{(n)} = b - Au^{(n)}$.

An important class of iterative methods is the class of *polynomial methods*, which are methods which satisfy the property

$$u^{(n)} \in u^{(0)} + K_n(r^{(0)}, A) \tag{2a}$$

where $K_n(v, A) = \mathrm{span}\{A^i v\}_{i=0}^{n-1}$ is the associated *Krylov space*. The condition (2a) may be shown to be equivalent to

$$r^{(n)} = P_n(A)r^{(0)}, \quad \deg P_n \leq n, \quad P_n(0) = 1 \tag{2b}$$

for some polynomial P_n.

Many important iterative methods, including the standard conjugate gradient method, fall under the category of polynomial methods (see e.g. [Joubert/Manteuffel 1990]). Another important example of a polynomial method is the *minimal residual method*, defined by the properties

$$u^{(n)} \in u^{(0)} + K_n(r^{(0)}, A), \quad \|r^{(n)}\| \text{ minimized.} \tag{3a}$$

Here the Euclidean norm $\|v\| = \sqrt{(v, v)}$ is based on the standard inner product $(u, v) = \sum \bar{u}_i v_i$. The minimal residual method may be equivalently defined by the following properties (see e.g. [Saad/Schultz 1985]):

$$u^{(n)} \in u^{(0)} + K_n(r^{(0)}, A), \quad r^{(n)} \perp A K_n(r^{(0)}, A). \tag{3b}$$

We define the *degree* of a vector with respect to a matrix by $d(v, A) = \min\{\deg P: P \text{ monic}, P(A)v = 0\}$. It will be noted that for the minimal residual method, $u^{(n)}$ is always uniquely defined by (3a) or (3b), and furthermore $u^{(n)} = u$ if and only if $n \geq d(r^{(0)}, A)$ (see [Joubert/Manteuffel 1990]).

Now, let Q_n be any matrix of size $N \times m$ whose columns span the space $K_n(r^{(0)}, A)$. Note that Q_n is of full rank if and only if $n \leq d(r^{(0)}, A)$. Then it may be shown that for such n, the iterates defined by (3a, b) satisfy

$$u^{(n)} = u^{(0)} + Q_n[Q_n^* A^* A Q_n]^{-1} Q_n^* A^* r^{(0)}, \tag{4a}$$

$$r^{(n)} = r^{(0)} - A Q_n[Q_n^* A^* A Q_n]^{-1} Q_n^* A^* r^{(0)}, \tag{4b}$$

(see [Joubert 1990a]). Furthermore, the residual norm $\|r^{(n)}\|$ may be shown by simple algebra to satisfy

$$\|r^{(n)}\|^2 = \|r^{(0)}\|^2 - r^{(0)*} A Q_n[Q_n^* A^* A Q_n]^{-1} Q_n^* A^* r^{(0)}. \tag{5}$$

Here and throughout we assume that the superscript "*" denotes the complex conjugate of a scalar or the conjugate transpose of a vector or matrix.

On the other hand, if $n > d(r^{(0)}, A)$, then there exists $d \equiv d(r^{(0)}, A) < n$ such that $u^{(d)} = u$, $r^{(d)} = 0$ and Q_d is full rank. One could thus compute $u^{(n)} = u^{(d)} = u$ by using (4a, b) replacing n with d, since Q_d is of full rank. Alternatively, one could directly compute $u^{(n)} = u$ from Q_n by setting

$$u^{(n)} = u^{(0)} + Q_n y_n$$

where y_n is any vector satisfying the (singular) system

$$Q_n^* A^* A Q_n y_n = Q_n^* A^* r^{(0)}. \tag{6}$$

To do this, we may let

$$y_n = (Q_n^* A^* A Q_n)^+ Q_n^* A^* r^{(0)}$$

where $(Q_n^* A^* A Q_n)^+$ is the pseudo-inverse [Golub/Van Loan 1989, p. 243]. Then

$$u^{(n)} = u^{(0)} + Q_n[Q_n^* A^* A Q_n]^+ Q_n^* A^* r^{(0)},$$

$$r^{(n)} = r^{(0)} - A Q_n[Q_n^* A^* A Q_n]^+ Q_n^* A^* r^{(0)},$$

$$\|r^{(n)}\|^2 = \|r^{(0)}\|^2 - r^{(0)*} A Q_n[Q_n^* A^* A Q_n]^+ Q_n^* A^* r^{(0)}.$$

The pseudo-inverse may be used not only when Q_n is rank deficient but also when it is *numerically* rank deficient due to a poorly conditioned basis.

The accurate solution of (6) when $Q_n^* A^* A Q_n$ is nonsingular is based directly on the quantity $\text{cond}(Q_n^* A^* A Q_n)$. A numerically better scheme whose accuracy is based on the improved quantity $[\text{cond}(Q_n^* A^* A Q_n)]^{1/2}$ may be developed as follows. Let $A Q_n = \check{Q}_n R_n$ where $\check{Q}_n^* \check{Q}_n = I$. Note that R_n need not be square. Thus, finding y_n to satisfy (6) is equivalent to finding y_n to minimize $\|A Q_n y_n - r^{(0)}\|$, or equivalently $\|R_n y_n - \check{Q}_n^* r^{(0)}\|$. This small problem may be solved by standard least squares techniques [Golub/Van Loan 1989]. For examples of this approach, see [Saad/ Schultz 1986], [Walker 1988, 1989]. The use of this technique typically yields improved numerical properties; however, unless Q_n is orthogonal or some similar condition is satisfied, the formation of such $\check{Q}_n$ requires significant added computational work.

In order to save storage, it is common to let Q_{n-1} be a submatrix of Q_n for $n \geq 1$. We thus have a set of vectors $q^{(n-1)} \in K_n(r^{(0)}, A)$, so that $Q_n = [q^{(0)} \quad q^{(1)} \quad \cdots \quad q^{(n-1)}]$ for any such n, and for $1 \leq n \leq d(r^{(0)}, A)$ we in fact have $q^{(n-1)} \in K_n(r^{(0)}, A) \setminus K_{n-1}(r^{(0)}, A)$. This implies relations of the form

$$h_{1,0} q^{(0)} = r^{(0)}, \quad h_{n+1,n} q^{(n)} = A q^{(n-1)} - \sum_{i=0}^{n-1} h_{i+1,n} q^{(i)} \tag{7a}$$

for any $n > 0$, or

$$AQ_n = Q_{n+1} H_{n+1}, \tag{7b}$$

where $H_{n+1} = \{h_{i,j}\}$ is an $(n + 1) \times (n)$ upper Hessenberg matrix and $h_{i+1,i}$ is chosen nonzero for all i. The recursion (7a) may be used directly to compute the $q^{(i)}$'s; however, other schemes are possible ([Saad/Schultz 1985], [Walker 1988]). Thus (4a, b) may be written

$$u^{(n)} = u^{(0)} + Q_n y_n, \tag{8a}$$

$$r^{(n)} = r^{(0)} - Q_{n+1} H_{n+1} y_n \tag{8b}$$

where

$$y_n = [Q_n^* A^* A Q_n]^{-1} Q_n^* A^* r^{(0)} = [H_{n+1}^* Q_{n+1}^* Q_{n+1} H_{n+1}]^{-1} H_{n+1}^* Q_{n+1}^* r^{(0)}. \tag{8c}$$

Also,

$$\| r^{(n)} \|^2 = \| r^{(0)} \|^2 - r^{(0)*} Q_{n+1} H_{n+1} [H_{n+1}^* Q_{n+1}^* Q_{n+1} H_{n+1}]^{-1} H_{n+1}^* Q_{n+1}^* r^{(0)} \tag{9}$$

Thus it is possible to derive algorithms to calculate (4a, b) which do not store AQ_n in memory, but only Q_{n+1}.

What we have defined is the basic minimal residual method (3a, b) as well as the general form of any algorithm for computing the minimal residual iterates $u^{(n)}$. Since the cost to form $u^{(n)}$ from $u^{(0)}$ by (4a) or (8a,c) is generally of order n^2 (compared to order n for standard CG applied to the HPD case), minimal residual algorithms for most choices of $r^{(0)}$ and (non-Hermitian) A become costly in terms of work per iteration as well as storage for large n (for specific conditions when the work to compute $u^{(n)}$ from $u^{(0)}$ is of order n, see [Faber/Manteuffel 1984, 1987], [Joubert/Young 1987], [Joubert 1990b, Section 2.5]). Thus it is common to *restart* these algorithms periodically. Specifically, for some integer $k \geq 1$, the full minimal residual iterate $u^{(k)}$ defined by (4a) is taken as the initial iterate for a second k-step cycle of the minimal residual algorithm, and the process is repeated. The effect of restarting is to decrease the average work per iteration. However, since not all of the Krylov space is used, the restarted method is not always guaranteed to converge, e.g. if A is indefinite (for convergence theory for these methods, see e.g. [Eisenstat/Elman/Schultz 1983], [Joubert 1990b, Sections 3.1–3.3]).

When appropriate, we use the notation developed above to describe a single restart cycle of the restarted minimal residual method, for the sake of clarity. On the other hand, the actual iterates of the restarted method are defined as $u^{(mk+i)}$, for $m \geq 0$ and $0 \leq i \leq k$, where $u^{(mk+i)}$ is the application of i steps of the minimal residual method to $u^{(mk)}$.

2.2 The Choice of Basis

In order to define particular algorithms based on the above framework, it is necessary to make a specific choice for the basis matrix Q_n. We now list several of

● It is often desirable that it be economical to form $u^{(n)}$ and $r^{(n)}$ at every step of the iteration. This is done, for example, in the Orthomin algorithm. It is not imperative that these be formed at every step, however. This is especially true if the residual norm $\|r^{(n)}\|$ may be accurately estimated by (5) or (9) so that convergence may be tested without forming $u^{(n)}$ or $r^{(n)}$ explicitly. This approach is typically used for the GMRES algorithm, for example, to save computation time.

● It is desirable for the computations (7a, b) to be as economical as possible. When A happens to be HPD, the recurrences for important algorithms such as Orthomin and GMRES may be simplified to give short recurrences, though for general A this is not possible; see [Faber/Manteuffel 1984, 1987], [Joubert/Young 1987]. Alternatively, short recurrences may be obtained by using, for example, Chebyshev polynomials for generating Q_n. On the other hand, the need to calculate explicitly all the inner products contained in the matrix $Q_n^*A^*AQ_n$ for general A seems to be intrinsic and not remediable by a judicious choice of the basis; see [Joubert 1990b, Section 2.5].

● Finally, for computer architectures for which inner product computations require significant memory transfer startup times, or for machines for which it is desirable to keep data in cache memory as long as possible by not calculating inner products during the basis generation phase, it is desirable that the calculation of the coefficients contained in H_{n+1} require no inner product computations, or at least that the frequency of these computations is kept to a minimum. From this standpoint, a Gram Schmidt procedure is most expensive, and a set of coefficients derived without inner product information is least expensive.

The application of these ideas to particular algorithms is discussed more fully later in Section 4.

3 THE CONDITION OF KRYLOV BASES

In this section we give a more theoretical discussion of the condition of Krylov bases. We are interested here in the *condition* of the basis represented by the columns of the matrix AQ_n, which is defined as the ratio of extremal singular values of AQ_n, that is, the square root of the quantity

$$\text{cond}(AQ_n)^*(AQ_n) \equiv \|(AQ_n)^*(AQ_n)\| \cdot \|(AQ_n)^*(AQ_n)^{-1}\|.$$

The condition of the inner product matrix $(AQ_n)^*(AQ_n)$ may often be improved by a diagonal scaling. Thus, as indicated above, we are typically more concerned with the condition of the basis given by $A\tilde{Q}_n \equiv AQ_nD_n^{-1/2}$, where $(A\tilde{Q}_n)^*(A\tilde{Q}_n)$ is unit diagonal. The problem of finding the *best* scaling D_n to minimize the condition is in general a difficult problem; see e.g. [Businger 1967], [Golub/Van Loan 1989, p. 125].

For the Orthomin algorithm, the basis represented by $A\tilde{Q}_n$ is perfectly conditioned, i.e. has condition of 1. Furthermore, for the GMRES algorithm, since $H_{n+1}^*H_{n+1}$ is an orthogonal section of A^*A, the condition of the basis AQ_n is no worse than the condition of A.

For these two algorithms, the bases are guaranteed to be well-conditioned due to the use of a Gram Schmidt procedure. However, for bases derived directly from

polynomial theory, the estimation of the condition is more difficult. In what follows we give arguments to show that the Chebyshev basis is a natural choice which tends to keep the condition as small as possible.

The basic approach is the following. Certain results are known for the condition of polynomial bases, where the inner product on the polynomial space is defined to be some weighted integral over a region in the complex plane. If this region is some sense approximates the spectrum of the matrix A, then under appropriate circumstances the condition of the polynomial basis $\{q_i(z)\}$ approximates the condition of the related Krylov basis $\{q_i(A)Ar^{(0)}\}$. Thus, a well-conditioned set of polynomials (e.g. orthonormal polynomials) can be used to form a well-conditioned Krylov basis.

3.1 Basic Setting

We can now proceed to make this argument more specific. We let $\mathscr{P}_{n-1}$ be the space of polynomials over $\mathbb{C}$ of degree not greater than $n-1$. Given polynomials $\{q_i\}_{i=0}^{n-1}$ which form a basis for $\mathscr{P}_{n-1}$, we define the coefficient map $F_n: \mathbb{C}^n \to \mathscr{P}_{n-1}$ by $F_n(\alpha) = \sum_{i=1}^{n} \alpha_i q_{i-1}$, where $\alpha = [\alpha, \ldots, \alpha_n]^T$. We also define

$$G_n: \mathscr{P}_{n-1} \to AK_n(r^{(0)}, A) \subseteq \mathbb{C}^N$$

by $G_n(P) = P(A)Ar^{(0)}$.

Following [Gautschi 1984], we define the *condition* of a map $C: S_1 \to S_2$ between normed spaces S_1 and S_2 to be

$$\frac{\max_v \| C(v) \|_{S_2} / \| v \|_{S_1}}{\min_v \| C(v) \|_{S_2} / \| v \|_{S_1}},$$

where the relevant norms for the two spaces are used. If C is invertible, then this equals $\| C \|_{S_1, S_2} \cdot \| C^{-1} \|_{S_2, S_1}$, where the induced operator norms are used. By this definition, the condition of the basis represented by the columns of $A\tilde{Q}_n$ coincides with the condition of the map represented by this matrix between vector spaces endowed with the l^2 norm for the domain and range.

It is clear that F_n defined above is linear and in fact is an isomorphism. Also, G_n is linear, and it is one to one if and only if $n \leq d(r^{(0)}, A)$. The condition of these maps depends on the choice of the norms used for the spaces $\mathbb{C}^n$, $\mathscr{P}_{n-1}$ and $\mathbb{C}^N$. If the l^2 norms are used for $\mathbb{C}^n$ and $\mathbb{C}^N$, and the norms used for $\mathscr{P}_{n-1}$ are the same for both F_n and G_n, then we have the composition map $G_n \circ F_n: l^2(\mathbb{C}^n) \to l^2(\mathbb{C}^N)$, and it is easily seen that the condition of the map $G_n \circ F_n$ is exactly the condition of $A\tilde{Q}_n$, where we define $\tilde{Q}_n e_i = q_{i-1}(A)r^{(0)}$. Here $e_i = [0 \ \ldots \ 0 \ 1 \ 0 \ \ldots \ 0]^T$ is the standard unit basis vector with 1 in the ith position and 0 elsewhere. We also obtain the simple inequalities

$$\text{cond}(G_n \circ F_n) \leq \text{cond}(G_n)\text{cond}(F_n), \tag{10a}$$

$$\text{cond}(G_n) \leq \text{cond}(G_n \circ F_n)\text{cond}(F_n), \tag{10b}$$

$$\text{cond}(F_n) \leq \text{cond}(G_n)\text{cond}(G_n \circ F_n). \tag{10c}$$

This setting leads to the possibility of evaluating the behavior of the condition of $A\tilde{Q}_n$ based on the maps F_n and G_n.

3.2   The Condition of F_n

Standard results exist from approximation theory which deal with the condition of the map F_n, for various choices of the polynomial basis. A norm is commonly placed on the space $\mathscr{P}_{n-1}$ which corresponds to a weighted integral involving the polynomial over a given region $\Omega \subseteq \mathbb{C}$.

We are concerned here with polynomials which are orthogonal over the interior of an ellipse in $\mathbb{C}$. We define $\mathscr{E}(c, d, a)$ to be the (boundary of the) ellipse centered at c with foci $c \pm d$ whose major axis is the line segment defined by the endpoints $c \pm a$ in the complex plane (see [Davis 1963], [Saad 1986]). We note that $|d| \leq |a|$. We then let $\Omega = \text{hull}(\mathscr{E}(c, d, a))$, where $\text{hull}(\cdot)$ denotes the convex hull.

We define the $\mathscr{L}^2$ inner product $(f, g)_{2,\Omega} = \int_\Omega \bar{f}g$ and the weighted $\mathscr{L}^2$ inner product $(f, g)_{2,\Omega,w} = \int_\Omega w\bar{f}g$ on the space $\mathscr{P}_{n-1}$, where $w(z) = |(c + d - z)(c - d - z)|^{-1/2}$ is a real-valued weight function. We note that the given integral is over a two-dimensional domain in $\mathbb{C}$ if and only if $|a| > |d|$. We also define the $\mathscr{L}^p$ norm $\|f\|_{p,\Omega} = (\int_\Omega |f|^p)^{1/p}$ and the weighted $\mathscr{L}^p$ norm $\|f\|_{p,\Omega,w} = (\int_\Omega w|f|^p)^{1/p}$, and we let $\{\mathscr{P}_{n-1}, \|\cdot\|_{p,\Omega}\}$ and $\{\mathscr{P}_{n-1}, \|\cdot\|_{p,\Omega,w}\}$ denote the space $\mathscr{P}_{n-1}$ endowed with the respective norms.

For $d \neq 0$ we define the polynomials $\tilde{P}_i(z) = T_i((z - c)/d)$, where $T_i(z) = \cosh(i \cosh^{-1}(z))$ is the Chebyshev polynomial of degree i [Rivlin 1976]. We then let $q_i(z) = \tilde{P}_i(z)/\|\tilde{P}_i(z)\|_{2,\Omega,w}$. Note that $\{q_i\}_{i=0}^{n-1}$ is an orthonormal basis of $\{\mathscr{P}_{n-1}, \|\cdot\|_{2,\Omega,w}\}$ [Davis 1963, p. 240]. On the other hand, in the degenerate case of $d = 0$ we let $\tilde{P}_i(z) = (z - c)^i$ and $q_i(z) = \tilde{P}_i(z)/\|\tilde{P}_i(z)\|_{2,\Omega}$. In this case, $\{q_i\}_{i=0}^{n-1}$ is an orthonormal basis of $\{\mathscr{P}_{n-1}, \|\cdot\|_{2,\Omega}\}$.

It is necessary to estimate the condition of the map $F_n: l^2(\mathbb{C}^n) \to \{\mathscr{P}_{n-1}, \|\cdot\|_{2,\Omega}\}$ defined by $[\alpha_0, \ldots, \alpha_{n-1}]^T \mapsto \sum \alpha_i q_i$. For the case of $d = 0$, as noted in [Davis 1963, p. 240], the map F_n is perfectly conditioned, due to the orthonormality of these polynomials in $\{\mathscr{P}_{n-1}, \|\cdot\|_{2,\Omega}\}$ over the disk of radius a. When $d \neq 0$, the situation is more complicated. We consider F_n to be the composition of maps $l^2(\mathbb{C}^n) \to \{\mathscr{P}_{n-1}, \|\cdot\|_{2,\Omega,w}\} \to \{\mathscr{P}_{n-1}, \|\cdot\|_{2,\Omega}\}$, where the second map of the sequence is the identity map. The map from $l^2(\mathbb{C}^n)$ to $\{\mathscr{P}_{n-1}, \|\cdot\|_{2,\Omega,w}\}$ which maps $[\alpha_0, \ldots, \alpha_{n-1}]^T$ to $\sum \alpha_i q_i$ has a condition of 1. Thus the condition of F_n equals exactly the condition of the identity map from $\{\mathscr{P}_{n-1}, \|\cdot\|_{2,\Omega,w}\}$ to $\{\mathscr{P}_{n-1}, \|\cdot\|_{2,\Omega}\}$. In what follows we show that this condition has a small growth rate with respect to n.

In order to show $\text{cond } F_n$ has a small growth rate for the $d \neq 0$ case, we first prove a basic lemma. Let $\mu(\Omega)$ denote the measure of the set Ω. Then we have

LEMMA   Let $\{\varphi_i\}$ be an orthogonal basis for $\{\mathscr{P}_{n-1}, \|\cdot\|_{2,\Omega}\}$. Then for $p \in \mathscr{P}_{n-1}$,

$$\|p\|_{\infty,\Omega} \leq \|p\|_{2,\Omega} \cdot \sup_\Omega \sum_i \frac{|\varphi_i(z)|}{\|\varphi_i\|_{2,\Omega}}, \tag{11a}$$

$$\|p\|_{2,\Omega} \leq \sqrt{\mu(\Omega)}\|p\|_{\infty,\Omega}. \tag{11b}$$

Thus the condition of the identity map from $\{\mathscr{P}_{n-1}, \|\cdot\|_{2,\Omega}\}$ *to* $\{\mathscr{P}_{n-1}, \|\cdot\|_{\infty,\Omega}\}$ *is bounded by* $\sqrt{\mu(\Omega)}\, \sup_\Omega \sum_i (|\varphi_i(z)|/\|\varphi_i\|_{2,\Omega})$.

Proof We have by Cauchy-Schwartz

$$|p(z)| = \left| \sum_i \frac{(\varphi_i, p)_{2,\Omega}}{(\varphi_i, \varphi_i)_{2,\Omega}} \varphi_i(z) \right| \le \sum_i \frac{|(\varphi_i, p)_{2,\Omega}|}{(\varphi_i, \varphi_i)_{2,\Omega}} |\varphi_i(z)| \le \sum_i \frac{\|\varphi_i\|_{2,\Omega} \cdot \|p\|_{2,\Omega}}{(\varphi_i, \varphi_i)_{2,\Omega}} |\varphi_i(z)|$$

By taking suprema, (11a) follows. Further,

$$\|p\|_{2,\Omega}^2 = \int_\Omega |p|^2 \le \int_\Omega \|p\|_{\infty,\Omega}^2 = \mu(\Omega)\|p\|_{\infty,\Omega}^2$$

to yield (11b).   ∎

Let us next derive a bound for $\|F_n\|$. We have

$$\|v\|_{2,\Omega}^2 = \left\| \frac{w^{1/2}v}{w^{1/2}} \right\|_{2,\Omega}^2 \le \|w^{1/2}v\|_{2,\Omega}^2 \sup_\Omega w^{-1} = \|v\|_{2,\Omega,w}^2 \sup_\Omega w^{-1}.$$

But

$$\sup_\Omega w^{-1} = \sup_\Omega |(c+d-z)(c-d-z)|^{1/2} = \sup_{\text{hull}(\mathscr{E}(0,d,a))} |d^2 - z^2|^{1/2} \le (|d|^2 + |a|^2)^{1/2}.$$

It follows that

$$\|v\|_{2,\Omega} \le (|d|^2 + |a|^2)^{1/4} \|v\|_{2,\Omega,w} \tag{12}$$

so that $\|F_n\| \le (|d|^2 + |a|^2)^{1/4}$.

We now consider $\|F_n^{-1}\|$. Let B be a set which contains the singularities $c \pm d$ of w. Specifically, let $B = B_1 \cup B_2$, $B_1 \cap B_2 = \{\}$, and let B_1 contain $c + d$ and B_2 contain $c - d$. We also assume $B \subseteq \Omega$. Then

$$\|v\|_{2,\Omega,w}^2 = \|w^{1/2}v\|_{2,\Omega}^2 = \int_{\Omega \setminus B} w|v|^2 + \int_{B_1} w|v|^2 + \int_{B_2} w|v|^2$$

$$\le \|v\|_{2,\Omega}^2 \sup_{\Omega \setminus B} w + \left[\sup_{B_1} |v| \right]^2 \int_{B_1} w + \left[\sup_{B_2} |v| \right]^2 \int_{B_2} w. \tag{13}$$

We seek to bound $\sup_{B_m} |v|$, $m = 1, 2$, We let $\varphi_{i,m}$ be a set of orthonormal polynomials over B_m in the sense that $\int_{B_m} \bar{\varphi}_{i,m} \varphi_{j,m} = \delta_{i,j}$, the Kronecker delta. Then by the Lemma and since $B_m \subseteq \Omega$,

$$\sup_{B_m} |v| \le \left(\sup_{B_m} \sum_i |\varphi_{i,m}| \right) \|v\|_{2,B_m} \le \left(\sup_{B_m} \sum_i |\varphi_{i,m}| \right) \|v\|_{2,\Omega}. \tag{14}$$

Combining (13) and (14), we have

$$\| F_n^{-1} \| = \sup_v \frac{\| v \|_{2,\Omega,w}}{\| v \|_{2,\Omega}}$$

$$\leq \left[\sup_{\Omega \backslash B} w + \left(\int_{B_1} w \right) \left(\sup_{B_1} \sum_i | \varphi_{i,1} | \right)^2 + \left(\int_{B_2} w \right) \left(\sup_{B_2} \sum_i | \varphi_{i,2} | \right)^2 \right]^{1/2}$$

$$\leq \left[\sup_{\Omega \backslash B} w \right]^{1/2} + \left(\int_{B_1} w \right)^{1/2} \sup_{B_1} \sum_i | \varphi_{i,1} | + \left(\int_{B_2} w \right)^{1/2} \sup_{B_2} \sum_i | \varphi_{i,2} |. \qquad (15)$$

We now consider two specific cases. First, suppose $|a| > |d|$. We than let B_1 (alt. B_2) be the disk of radius r centered at $c + d$ (alt. $c - d$). For sufficiently small r, these disks are contained within the ellipse $\mathscr{E}(c, d, a)$. Then $\sup_{\Omega \backslash B} w$ and $\int_{B_m} w$ are finite quantities bounded independent of the polynomial degree $n - 1$. On the other hand, the quantity $\sup_{B_m} \sum_i | \varphi_{i,m} |$ may be bounded by choosing

$$\varphi_{i,m}(z) = \sqrt{\frac{i + 1}{\pi}} \frac{(z - (c \pm d))^i}{r^{i+1}}.$$

Then

$$\sup_{B_m} \sum_i | \varphi_{i,m} | \leq \sum_i \frac{1}{r} \sqrt{\frac{i + 1}{\pi}} \leq \frac{n}{r} \sqrt{\frac{n}{\pi}}.$$

These polynomials are orthonormal over the disk B_m [Davis 1963, p. 240]. We conclude that for $|a| > |d|$,

$$\text{cond } F_n \leq (|d|^2 + |a|^2)^{1/4} \left[\left[\sup_{\Omega \backslash B} w \right]^{1/2} + \left(\int_{B_1} w \right)^{1/2} \frac{2n}{r} \sqrt{\frac{n}{\pi}} \right]. \qquad (16)$$

Thus, the growth rate is no worse than order $n^{3/2}$.

On the other hand, for the case $|a| = |d|$ we have the case of a degenerate ellipse which is the line segment $[c - d, c + d]$ in the complex plane. We let B_1 (alt. B_2) be a line segment of length $2r$ contained in $[c - d, c + d]$ and containing $c + d$ (alt. $c - d$). We choose $\varphi_{i,j}$ to be scaled and translated Legendre polynomials for the appropriate intervals. Using the properties of the Legendre polynomials (see also [Gautschi 1972, p. 923]) we conclude that for the $|a| = |d|$ case,

$$\sup_{B_m} \sum_i | \varphi_{i,m} | \leq \sum_i \frac{1}{r} \sqrt{\frac{2i + 1}{2}} \leq \frac{n}{r} \sqrt{\frac{2n - 1}{2}}$$

and thus

$$\text{cond } F_n \leq (|d|^2 + |a|^2)^{1/4} \left[\left[\sup_{\Omega \backslash B} w \right]^{1/2} + \left(\int_{B_1} w \right)^{1/2} \frac{2n}{r} \sqrt{\frac{2n - 1}{2}} \right]. \qquad (17)$$

Again, the growth rate is no worse than order $n^{3/2}$.

This growth of $n^{3/2}$ is generally more favorable than an exponential growth rate of the form ξ^n for some $\xi > 1$, which is the case for the power basis in some important cases [Gautschi 1984]. Unfortunately, each of the above bounds has a constant multiplier based on the quantity r, and this multiplier may become large as the ellipse degenerates into a line segment or a circle. We conjecture however that these bounds are pessimistic. The dependence of the condition of F_n on the shape of the ellipse is a topic of further research.

3.3 The Condition of G_n

The condition of the map G_n may be arbitrarily large, as $r^{(0)}$ is made to be a vector with sufficient eigenvector deficiency, i.e. $A\tilde{Q}_n$ becomes rank deficient. However, cond $G_n = \infty$ if and only if $n < d(r^{(0)}, A)$, in which case $u^{(n-1)} = u$ and the minimal residual method has already converged. By a similar argument, if cond G_n is extremely large, then one might suspect that $r^{(0)}$ is "nearly" annihilated by an $(n-1)$-degree polynomial in A, in which case $r^{(n-1)}$ will be very small, and the algorithm is likely to have converged already to the desired stopping tolerance.

We now give a more meaningful result. To insure that cond G_n is small, it is desirable that the quantity $\| P(A)Ar^{(0)} \|$ is a reasonable measure of the size of the polynomial P over the given region $\Omega \supseteq \sigma(A)$, in a sense compatible with the norm placed on $\mathscr{P}_{n-1}$. We now make this precise. Assuming for the moment that A is diagonalizable and $V^{-1}AV = \Lambda = \mathrm{diag}\{\lambda_i\}$, where $\{\lambda_i\} = \sigma(A)$, and letting $r = V^{-1}r^{(0)} = [r_1, \ldots, r_N]^T$, we have

$$(\text{cond } G_n)^2 = \max_{P \in \mathscr{P}_{n-1}} \frac{\| P(A)Ar^{(0)} \|^2}{\int_\Omega |P|^2} \min_{P \in \mathscr{P}_{n-1}} \frac{\| P(A)Ar^{(0)} \|^2}{\int_\Omega |P|^2}$$

$$\leq (\text{cond } V)^2 \cdot \max_{P \in \mathscr{P}_{n-1}} \frac{\sum |P(\lambda_i)|^2 |\lambda_i r_i|^2}{\int_\Omega |P|^2} \min_{P \in \mathscr{P}_{n-1}} \frac{\sum |P(\lambda_i)|^2 |\lambda_i r_i|^2}{\int_\Omega |P|^2}. \quad (18)$$

In this case the condition of G_n is measured by the relationship between $\sum |P(\lambda_i)|^2 |\lambda_i r_i|^2$ and $\int_\Omega |P|^2$. If A has at least n distinct eigenvalues, then there exist values of r_i such that these two quantities are precisely equal for all $P \in \mathscr{P}_{n-1}$. These values may be obtained using the principles of numerical quadrature. However, in general it is not possible to predict the values of $\{r_i\}$, which are the eigencomponents of $r^{(0)}$. In lieu of precise knowledge about $\{r_i\}$, two things are desirable: first, that the components $\{r_i\}$ be evenly distributed (this could be enhanced by choosing $u^{(0)}$ to be a vector of random components), and secondly that the shape of Ω would match $\sigma(A)$ as closely as possible.

A failure to accurately estimate the spectrum may cause a significant increase of the condition of the basis. This need not always occur, though. For example, if $\Omega = \text{hull}(\mathscr{E}(c, d, a))$ but the spectrum $\sigma(A)$ is well-approximated by an alternate ellipse of the form $\Omega' \equiv \text{hull}(\mathscr{E}(c, d, a'))$ (having the same foci but different size), then the appropriate Chebyshev polynomials for Ω after an appropriate scaling will also be orthonormal over Ω' with respect to the appropriate weight function. Since the basis matrix AQ_n is typically scaled so that each column has norm 1, the use of an incorrect

ellipse is not a problem as long as the foci are correct. This indicates the relative importance of estimating the position and the orientation of the ellipse rather than its exact size.

Cases do exist for which scaling does not help, however. For example, suppose that Ω is estimated to be the disk centered at 2 and of radius 1, so that the shifted power basis $(z - 2)^i$ is generated, but $\sigma(A)$ is actually composed of a large number of values spread densely through the interval $[1, 3]$. According to [Gautschi 1984], the coefficient map $\alpha \in l^\infty(\mathbb{C}^n) \mapsto \sum (e_i^* \alpha)(z - 2)^{i-1} \in \{\mathscr{P}_{n-1}, \|\cdot\|_{\infty,[1,3]}\}$ has a condition that grows exponentially, and at least as fast as a constant multiplied by $(1 + \sqrt{2})^n$. Now, we desire to look at the same map but using the l^2 and $\mathscr{L}^2$ norms instead of the l^∞ and $\mathscr{L}^\infty$ norms. The condition of the identity map from $l^2(\mathbb{C}^n)$ to $l^\infty(\mathbb{C}^n)$ is bounded by $\sqrt{n}$ [Golub/Van Loan 1989]. On the other hand, using the above Lemma and the properties of the Legendre polynomials (see also [Gautschi 1972]), we see that the condition of the identity map from $\{\mathscr{P}_{n-1}, \|\cdot\|_{2,[1,3]}\}$ to $\{\mathscr{P}_{n-1}, \|\cdot\|_{\infty,[1,3]}\}$ has a growth rate no worse than $n\sqrt{2n-1}$. Since the exponential growth of $(1 + \sqrt{2})^n$ dominates, we conclude that the condition of the map $\alpha \in l^2(\mathbb{C}^n) \mapsto \sum (e_i^* \alpha) z^{i-1} \in \{\mathscr{P}_{n-1}, \|\cdot\|_{2,[1,3]}\}$ also grows exponentially, at least as fast as a constant times $(1 + \sqrt{2})^n$.

Letting $\Omega = [1, 3]$ in this example and assuming that cond V is small and that $\{r_i\}$ are of the same order of magnitude, then typically cond G_n will be small, as noted above in (18). Since $\text{cond}(G_n \circ F_n) \geq \text{cond}(F_n)/\text{cond}(G_n)$ from (10c), and since we just showed cond F_n grows exponentially, the quantity $\text{cond}(G_n \circ F_n)$ will typically grow exponentially. The basis $\{(z - 2)^i\}$ may be replaced by the scaled basis $\{[\sqrt{(2i + 1)/2}](z - 2)^i\}$ which is orthonormal with respect to $(\cdot, \cdot)_{2,[1,3]}$. However, the condition of this scaling operation is $\sqrt{(2n - 1)/2}$ which is still dominated by $(1 + \sqrt{2})^n$. This shows that the power basis for the Krylov space is inferior to the Chebyshev basis in some important cases.

The bound given above for cond G_n becomes worse as cond(V) becomes large, i.e. as A becomes highly non-normal. In this case, the behavior of the condition is less clear, and in general the properties of non-normal matrices are less well-understood. On the other hand, it appears likely that the bound based on cond(V) is pessimistic, as are similar bounds for the non-normal case (see e.g. [Joubert 1990b, Chapter 3]).

This analysis does not give a final answer to the question of the condition of Chebyshev bases, but it does illustrate some of the issues involved and gives a motivation and rationale for using Chebyshev bases. Specifically, this analysis shows that Chebyshev polynomials are a more natural choice than the power basis, because the region over which these polynomials are orthogonal has a better match to the spectrum of A than that of the power basis.

We now turn to the practical matter of particular algorithms for generating the minimal residual iterates.

4   RESTARTED MINIMAL RESIDUAL ALGORITHMS

We now consider particular algorithms. Unless otherwise noted, the computational scheme for a cycle of the given algorithm is arranged as indicated below. For

simplicity of notation, we refer here to the first k-step cycle of the given restarted algorithm.

1) Form the basis matrix Q_{k+1}. Avoid explicit scaling of the vectors when possible by saving scale factors in an auxiliary table and applying the scalings explicitly only to avoid overflow or underflow.

2) Form the inner products $Q_{k+1}^* Q_{k+1}$. If necessary, update the current estimate of $\|r^{(0)}\|$ by using the explicitly computed value of $(q^{(0)}, q^{(0)})$.

3) Form the matrices $[Q_{k+1}^* Q_{k+1}] H_{k+1}$ and $H_{k+1}^* [[Q_{k+1}^* Q_{k+1}] H_{k+1}]$. Apply diagonal scaling to the latter matrix.

4) Solve for y_k by (8c).

5) Form $u^{(k)}$ by (8a).

6) Form $H_{k+1} y_k$ and then $r^{(k)}$ by (8b).

7) Calculate $\|r^{(k)}\|$ by (9) or by an explicit computation.

This scheme may be directly implemented on serial computers; implementation details for parallel computers are discussed in Part II. Most of the algorithms to be discussed in this section closely follow the scheme given here; the Orthomin and GMRES algorithms depart slightly, for reasons to be described below.

The formation of y_k in step 4 by the solution of the small system may be accomplished by using Gaussian elimination. However, if the restart frequency k is extremely large, then for some algorithms the small matrix may be numerically singular, i.e. the condition may be greater than $1/\epsilon_M$, where ϵ_M is the relative machine precision, the smallest positive number representable on the given computer such that $1 + \varepsilon_M > 1$. As indicated in [Golub/Van Loan 1989, p. 124], the direction of the error from a Gaussian elimination solve with a nearly-singular matrix is largely in the direction of the null space. This effect in part mitigates the numerical error in $u^{(n)}$ and $r^{(n)}$, but this is not an adequate final solution. One way to avoid the problem is to shut off the Choleski factorization of the small matrix when a small or negative pivot is encountered and use only the part of the Krylov space accumulated up to that point. However, this approach is not guaranteed to make greatest possible use of all the Krylov vectors that have been generated.

Thus we use an alternative approach in order to execute step 4 above. Specifically we apply the pseudo-inverse of the small matrix by calculating a spectral decomposition using the approach of [Golub/Van Loan 1989, p. 418f]. This involves a Householder tridiagonalization followed by a diagonalization by Givens rotations. Eigencomponents of size less than ϵ_M are discarded. When properly implemented, the cost of this process is only about twice the cost of Choleski factorization.

The effect of this scheme is to eliminate the part of the Krylov basis which cannot be used due to ill-conditionedness. We thus obtain an iterative method which is guaranteed not to break down. However, if k is too large and the basis is ill-conditioned, then part of the work in generating the vectors will be wasted. In order to minimize the impact of this, it is possible to develop an adaptive scheme which switches to a safer restarted algorithm such as GMRES for future cycles in the case in which the current algorithm is generating highly ill-conditioned bases. This would

yield an algorithm as safe as GMRES in all cases, wasting at most the work of a single restart cycle.

On a scalar or vector machine, the work required for the small system solve is usually negligible compared to the vector work. However, on highly parallel machines, it becomes more significant. Furthermore, for the GCR and GMRES algorithms, the scalar work per cycle is order k^2, whereas for the other algorithms, which require a factorization of a $k \times k$ matrix, the scalar work per cycle is order k^3. This is the cost accrued from eliminating most of the vector SAXPY operations (operations of the form $y \leftarrow y + ax$ where x and y are vectors and a is a scalar) in the new algorithms. The parallelization of the linear solve with this small matrix is a topic of further study.

We make one further point about the above restarted scheme. Unfortunately, the formula (9) for calculating $\|r^{(k)}\|$ used in its unmodified form is often subject to numerical error and may not be reliable. Thus for these algorithms we perform an explicit computation of the residual norm, except for the case of GMRES, for which the residual norm formula is generally reliable. Means of dealing with this numerical problem are discussed in [Joubert 1990b].

We now discuss particular choices for the basis of the Krylov space as represented by Q_{k+1}.

4.1 The GCR and Orthomin Algorithms

The Orthomin algorithm ([Vinsome 1975], [Young/Jea 1980]) is defined as follows:

$$q^{(n)} = r^{(n)} + \sum_{i=\phi(n)}^{n-1} \alpha_{n,i} q^{(i)}, \quad \alpha_{n,i} = -\frac{(Aq^{(i)}, Ar^{(n)})}{(Aq^{(i)}, Aq^{(i)})}.$$

$$u^{(n+1)} = u^{(n)} + \lambda_n q^{(n)} \quad \lambda_n = \frac{(Aq^{(n)}, r^{(n)})}{(Aq^{(n)}, Aq^{(n)})} = \frac{(Ar^{(n)}, r^{(n)})}{(Aq^{(n)}, Aq^{(n)})}.$$
$$r^{(n+1)} = r^{(n)} - \lambda_n Aq^{(n)}$$

Setting $\phi(n) = 0$ gives the full Orthomin algorithm, whereas $\phi(n) = n - \text{mod}(n, k)$ defines the restarted algorithm, GCR($k - 1$). For the full algorithm, the vectors $\{Aq^{(i)}\}$ are mutually l^2-orthogonal. The inversion of $(AQ_k)^*(AQ_k)$ is performed implicitly through the formation of the coefficients λ_n.

4.2 The GMRES Algorithm

Since the GMRES algorithm generates the columns of Q_{k+1} to be orthonormal, the computation given in (8) may be rearranged to be more stable. Steps 3 and 4 of the scheme given at the beginning of this section are modified in what follows, by making use of a QR factorization of H_{k+1}.

As above we have

$$u^{(k)} = u^{(0)} + Q_k y_k, \tag{19a}$$

$$r^{(k)} = r^{(0)} - Q_{k+1} H_{k+1} y_k. \tag{19b}$$

In order that (3b) is satisfied, y_k must satisfy

$$H^*_{k+1} H_{k+1} y_k = H^*_{k+1} h_{1,0} e_1. \tag{20}$$

We solve this least-squares problem by calculating a QR factorization

$$H_{k+1} = P_{k+1} \begin{bmatrix} U_k \\ 0 \end{bmatrix} = P_{k+1} \begin{bmatrix} I_k \\ 0 \end{bmatrix} U_k \tag{21}$$

where I_k is the identity, U_k is upper triangular and $P_{k+1} = \prod_{i=1}^{k} R_{k+1,i} = R_{k+1,1} R_{k+1,2} \cdots R_{k+1,k}$ is the left-to-right product of Givens rotations

$$R_{k+1,i} = \begin{bmatrix} I_{i-1} & & \\ & \begin{bmatrix} c_i & -s_i^* \\ s_i & c_i^* \end{bmatrix} & \\ & & I_{k-i} \end{bmatrix},$$

where c_i and s_i are chosen to annihilate the lower subdiagonal elements of H_{k+1}. Then y_k satisfies

$$U_k y_k = \begin{bmatrix} I_k \\ 0 \end{bmatrix}^* P^*_{k+1} h_{1,0} e_1 \tag{22}$$

and thus

$$u^{(k)} = u^{(0)} + Q_k U_k^{-1} \begin{bmatrix} I_k \\ 0 \end{bmatrix}^* P^*_{k+1} h_{1,0} e_1 = u^{(0)} + Q_k U_k^{-1} \check{z}_k,$$

$$r^{(k)} = r^{(0)} - Q_{k+1} H_{k+1} U_k^{-1} \check{z}_k$$

where

$$\check{z}_k = U_k y_k. \tag{23}$$

We find $\check{z}_k$ as follows. We let

$$\check{z}_k = \begin{bmatrix} I_k \\ 0 \end{bmatrix}^* z_{k+1} \tag{24}$$

where $z_{k+1} = P^*_{k+1} h_{1,0} e_1$ and develop a recurrence for z_{k+1}. We have

$$z_{k+1} = R^*_{k+1,k} \begin{bmatrix} z_k \\ 0 \end{bmatrix} \tag{25a}$$

or

$$z_{k+1} \equiv \begin{bmatrix} z_{k-1} \\ \hat{\zeta}_k \\ \tilde{\zeta}_k \end{bmatrix} = R^*_{k+1,k} \begin{bmatrix} z_{k-2} \\ \hat{\zeta}_{k-1} \\ \tilde{\zeta}_{k-1} \\ 0 \end{bmatrix} = \begin{bmatrix} z_{k-2} \\ \hat{\zeta}_{k-1} \\ c^*_k \tilde{\zeta}_{k-1} \\ -s_k \tilde{\zeta}_{k-1} \end{bmatrix} \qquad (25b)$$

yielding the recurrences $\hat{\zeta}_k = c^*_k \tilde{\zeta}_{k-1}$, $\tilde{\zeta}_k = -s_k \tilde{\zeta}_{k-1}$ and $\tilde{\zeta}_0 = h_{1,0}$.
    We note also that

$$H_{k+1} y_k = P_{k+1} \begin{bmatrix} U_k \\ 0 \end{bmatrix} U_k^{-1} \check{z}_k = P_{k+1} \begin{bmatrix} \check{z}_k \\ 0 \end{bmatrix}. \qquad (26)$$

The residual norm may be calculated by the formula

$$\| r^{(k)} \|^2 = \| r^{(k-1)} \|^2 - |\hat{\zeta}_k|^2 = |\tilde{\zeta}_k|^2. \qquad (27)$$

To sum up, we have the following scheme for this algorithm:

1) Calculate Q_{k+1}, H_{k+1}.
2) Calculate H_{k+1} by (21).
3) Calculate z_{k+1}, $\check{z}_k$ using (25a, b), (24).
4) Calculate y_k by backward solution with U_k (23).
5) Form $u^{(k)}$ by (19a).
6) Calculate $H_{k+1} y_k$ by (26), form $r^{(k)}$ by (19b).
7) Form $\| r^{(k)} \|$ by (27).

4.3 The Power Basis Algorithm

We next define a simple restarted algorithm which uses the power basis: set $q^{(0)} = r^{(0)}$ and then $\gamma_i q^{(i)} = A q^{(i-1)}$ for $i \geq 1$. In this formulation we choose $\gamma_i \equiv \gamma$ to be a fixed scaling factor on the order of the size of $\| A \|$, to avoid possible overflow or underflow on the given computer.

4.4 The Chebyshev Basis Algorithm

We now define an algorithm based on Chebyshev polynomials. We assume that the spectrum of A is contained in the ellipse with axes parallel to the real and imaginary axes and which is circumscribed by the rectangle $\{x + iy : c - a \leq x \leq c + a, -b \leq y \leq b\}$, where $a \geq 0$, $b \geq 0$ and c are real. This assumption is natural when A is real, since the non-real eigenvalues occur in complex conjugate pairs. Let $d = \sqrt{a^2 - b^2}$, so that the foci are at $c \pm d$, and choose $\gamma \neq 0$ on the order of $\| A - cI \|$, e.g. $\gamma = \max\{a, b\}$. Then define the scaled and translated Chebyshev polynomials

[Manteuffel 1977]:

$$P_{c,0}(z) = 1,$$

$$P_{c,1}(z) = \frac{1}{2\gamma}\,[z - c],$$

$$P_{c,n}(z) = \frac{1}{\gamma}\left[(z - c)P_{c,n-1}(z) - \frac{d^2}{4\gamma}P_{c,n-2}(z)\right].$$

We then let $q^{(n)} = P_{c,n}(A)r^{(0)}$. Note that the resulting vectors satisfy a three-term recurrence, so that H_{k+1} is tridiagonal.

Based on the discussion in Section 3 above, the resulting matrix $Q_k^* A^* A Q_k$ might be expected in many cases to be well-conditioned. The exact estimation of the extremal eigenvalues of A is not as critical in this context as it is for adaptive Chebyshev acceleration ([Manteuffel 1978], [Hageman/Young 1981]), since a failure to estimate the spectrum precisely results only in a somewhat less well-conditioned basis, rather than possible divergence of the method, as may occur with Chebyshev acceleration. It is also true that the condition of the basis is shift and scaling invariant, in the sense that if a multiple of the identity is added to A or A is multiplied by a constant then the corresponding polynomials may be shifted and scaled to give the same basis condition as for the original matrix A. Thus, indefinite matrices are not a problem for this technique as they are for Chebyshev acceleration.

It is desirable, nonetheless, that the ellipse match the spectrum of A as well as possible, to minimize the basis condition. For spectral distributions that do not match ellipses well, other types of orthogonal polynomials may be used, but they may not possess three-term recurrences. The GMRES/repeated basis algorithm defined below uses polynomial recurrences which in some sense are tailor-made to the first restart cycle of the run and thus are likely to give well-conditioned bases for subsequent runs, for arbitrary spectral distributions.

We point out that other well-conditioned bases with short recurrences may be developed which, like the Chebyshev basis, are better than the power basis; see [Bai/Hu/Reichel 1991] for an alternative approach based on a Leja ordering of approximate eigenvalues of A.

4.5   The GMRES/Chebyshev Basis Algorithm

The Chebyshev basis algorithm defined above depends on estimates of the spectrum of A in order to be implemented effectively. Typically, the spectrum of a given matrix is not known beforehand. We thus define the GMRES/Chebyshev basis algorithm to be one cycle of the restarted GMRES algorithm for the first restart cycle, to provide eigenvalue estimates, followed by subsequent restart cycles using the Chebyshev basis algorithm, based on these estimates.

We obtain the spectral estimates as follows. The matrix $\tilde{H}_k$, which is defined to be the upper k rows of the matrix H_{k+1} from a cycle of GMRES, is an orthogonal section of A, whose eigenvalues approximate those of A. We seek a rectangular region

in the complex plane which contains the spectrum of $\tilde{H}_k$. To estimate the extremal real parts of the spectrum of $\tilde{H}_{k+1}$, we may use estimates for the extremal eigenvalues of the Hermitian part of $\tilde{H}_k$, namely $\tilde{H}_{k,H} \equiv (\tilde{H}_k + \tilde{H}_k^*)/2$. Likewise, to estimate the extremal imaginary parts of the spectrum of $\tilde{H}_k$, we may estimate the extremal eigenvalues of the skew-Hermitian part $\tilde{H}_{k,N} \equiv (\tilde{H}_k - \tilde{H}_k^*)/2$. An alternative approach is to use directly the real and imaginary parts of the eigenvalues of $\tilde{H}_k$. Numerical comparisons of these approaches are given in the next section.

To estimate the spectra of the matrices $\tilde{H}_{k,H}$ and $\tilde{H}_{k,N}$, several techniques may be applied. First, we make use of a Gershgorin-type theorem due to Ostrowski based on Cassini ovals [Householder 1964, p. 67] to find the smallest interval containing all eigenvalues for each of these two matrices. A more accurate but more costly scheme is to calculate exactly the eigenvalues of $\tilde{H}_k$, or alternatively $\tilde{H}_{k,H}$ and $\tilde{H}_{k,N}$. The cost of this calculation by standard techniques [Golub/Van Loan 1989, p. 378] is only several times the cost of a linear solve with this small system.

A rectangle of the form $\{x + iy: c - \alpha < x < c + \alpha, -\beta < y < \beta\}$ is obtained by this scheme. From this, an inscribed ellipse of proportional shape may be obtained by letting $a = \alpha$ and $b = \beta$. An inscribing rather than circumscribing ellipse is used so that in the limiting cases of $\alpha = 0$ or $\beta = 0$, the ellipse is matched exactly. Even these approximate values of a, b and c can give substantial improvement over the power basis as will be demonstrated later.

In brief, the algorithm is summarized as follows:

- Perform one cycle of GMRES.

- Obtain eigenvalue estimates from the associated Hessenberg matrix.

- Form the smallest bounding rectangle with horizontal and vertical sides.

- Form an inscribing ellipse.

- Use the associated Chebyshev polynomials for all subsequent cycles. Use the pseudo-inverse to solve the small system.

It is possible to modify this algorithm so that near the end of the iteration process a switch back to GMRES is performed, so that reliable residual norm estimates could be obtained which would allow termination of the iteration in the middle of a restart cycle if the stopping test were satisfied; see [Joubert 1990b, Section 3.2.2]. We will not pursue these ideas here.

This algorithm is generally less expensive than GMRES, and in many cases requires about half the cost of restarted GMRES. Furthermore, by using the pseudo-inverse for the small system solve, breakdown of the algorithm is avoided. As mentioned earlier, the algorithm can be easily modified to be as "safe" as GMRES, as follows. If in the calculation of the pseudo-inverses, one of the small matrices is found to be numerically rank deficient, then the decision can be made to revert to GMRES for all future restart cycles. Thus, in the worst case, only the work for a single restart cycle is lost.

A further modification of the algorithm is also possible which prevents the loss of even one restart cycle but at the same time requires the computation of inner products at each step, thus diminishing the useful parallel properties of the algorithm. To do this, at each step not only $q^{(n)}$ but also all associated inner products of $Q_n^* A^* A Q_n$ are

computed. The condition of the diagonally scaled $Q_n^* A^* A Q_n$ is then computed, and if this quantity is too large, then $q^{(n)}$ can then be safely orthogonalized against the previous basis vectors. Subsequent basis vectors of the cycle may likewise be orthogonalized. This procedure guarantees that $Q_n^* A^* A Q_n$ is well-conditioned. We omit the details of this strategy.

It should be pointed out that as the restart frequency k increases, the likelihood of an ill-conditioned basis for the GMRES/Chebyshev basis algorithm increases. On the other hand, this is compensated for in part by the fact that as k grows, the spectral estimates for A become better, so that the operative ellipse can better match the spectrum of A.

In situations in which k is extremely large and a reliable residual norm estimate is desired before the cycle is over in order to test for termination, then it is possible to perform the inner product calculations more often, say, every s steps, for some $s \geq 1$.

4.6   The GMRES/Repeated Basis Algorithm

The basis condition bound derived in Section 3 depended on cond V, where $AV = VJ$ is a Jordan decomposition. If cond V is large, it is not clear that a well-conditioned basis can in general be formed by means of a three-term recurrence. Furthermore, the condition of the basis may also be large if the spectrum of A is not well-matched by *any* ellipse, i.e. no ellipse exists for which the eigenvalues of A are uniformly distributed within the entire ellipse. We therefore define the following algorithm which uses a long recurrence. For the first restart cycle we apply GMRES. For future cycles, the basis vectors $q^{(i)}$ are generated by means of the identical recurrence coefficients used for the first GMRES cycle, embodied in the original matrix H_{k+1}. This algorithm resembles the hybrid GMRES algorithm proposed in [Nachtigal/Reichel/Trefethen 1990], which builds a polynomial for Richardson's method based on an initial GMRES cycle; however, their approach does not make use of an inner product matrix of the form $(AQ_k)^*(AQ_k)$, so there is no need that the Krylov basis be well-conditioned. This basis approach may also be combined with the periodic orthogonalization ideas of [Chronopoulos/Kim 1990].

The algorithm defined here requires nearly the same work as restarted GMRES, with only a single extra residual norm calculation at the end of each cycle. However, as noted previously, the algorithm has benefits for various parallel architectures or when cache or out-of core memory is used. Furthermore it may be expected that the bases thus generated are typically well-conditioned. It remains an open question as to what is the "best" (in some average or minimax sense) set of long-recurrence or three-term recurrence polynomials which would typically give well-conditioned bases.

5.   PERFORMANCE OF THE ALGORITHMS

We now give comparison of the CPU requirements and numerical properties of the algorithms. Actual timings for these algorithms on various computers are given in Part II.

5.1 Work Requirements

The work requirements for the algorithms described in Section 4 are compared in Tables 1 and 2 below. An explicit residual norm computation at each restart for the power basis and Chebyshev basis algorithms is included. Otherwise, work for the stopping test is not included. In all cases, k denotes the restart frequency. As can be seen, for large k, the cost of the vector operations relative to the matrix-vector product becomes larger for k sufficiently large the Chebyshev basis algorithm requires close to half the work of GMRES.

In Table 2 we assume equal relative costs for the dot product and the SAXPY operations.

Table 1 Average work per iteration (excluding matrix-vector products)

Method	Dot products	SAXPY's	Total
GCR($k-1$)	$\dfrac{(k+3)}{2}$	$k+1$	$\dfrac{3}{2}k + \dfrac{5}{2}$
GMRES(k)	$\dfrac{(k+1)(k+2)}{(2k)}$	$\dfrac{(k+1)(k+2)}{(2k)} + \dfrac{1}{k}$	$k + 3 + \dfrac{3}{k}$
Power(k)	$\dfrac{(k+1)(k+2)}{(2k)} + \dfrac{1}{k}$	2	$\dfrac{1}{2}k + \dfrac{7}{2} + \dfrac{2}{k}$
Chebyshev(k)	$\dfrac{(k+1)(k+2)}{(2k)} + \dfrac{1}{k}$	$4 - \dfrac{1}{k}$	$\dfrac{1}{2}k + \dfrac{11}{2} + \dfrac{1}{k}$

Table 2 Average vector operations per iteration for some values of k (excluding matrix-vector products)

k\Method	GCR($k-1$)	GMRES(k)	Power(k)	Chebyshev(k)
1	4.0	7.000	6.000	7.000
2	5.5	6.500	5.500	7.000
3	7.0	7.000	5.667	7.333
4	8.5	7.750	6.000	7.750
5	10.0	8.600	6.400	8.200
6	11.5	9.500	6.833	8.667
7	13.0	10.429	7.286	9.143
8	14.5	11.375	7.750	9.625
9	16.0	12.333	8.222	10.111
10	17.5	13.300	8.700	10.600
11	19.0	14.273	9.182	11.091
12	20.5	15.250	9.667	11.583
13	22.0	16.231	10.154	12.077
14	23.5	17.214	10.643	12.571
15	25.0	18.200	11.133	13.067

5.2   The Condition of Bases

We now give some numerical results to demonstrate the behavior of the basis condition associated with these algorithms. We use the following test problem. We consider a finite difference discretization of the boundary value problem

$$- u_{xx}(x, y) - u_{yy}(x, y) + Cu_x(x, y) = g(x, y) \quad \text{on} \quad \Omega = [0, 1]^2,$$

$$u(x, y) = 1 + xy \quad \text{on} \quad \partial\Omega.$$

Central differencing, with uniform mesh spacing h_x and h_y in the x and y directions, yields a sparse matrix of size $N = (n_x - 1)(n_y - 1)$ (where $h_x = 1/n_x$, $h_y = 1/n_y$), after boundary points have been eliminated. The right hand function $g(x, y)$ is defined so that the true solution is $u(x, y) = 1 + xy$ on Ω. By varying the constant C, the amount of nonsymmetry of the matrix may be varied, and Krylov bases with different properties may thus be examined. Specifically, for given h_x and h_y there exist a symmetric matrix A_S and a skew-symmetric matrix A_N fixed independent of C such that $A = A_S + C \cdot A_N$. Problems of this type arise frequently in transport processes and are of significant practical importance.

The initial guess vector is $u^{(0)} = 0$ and no preconditioning is used for these experiments. Uniprocessor calculations are made using double precision on the University of Texas System Cray Y-MP 8/864 vector computer.

For the test problem we let $n_x = n_y = 51$ and use several values of C corresponding to a symmetric, a mildly nonsymmetric and a highly nonsymmetric problem. In all runs, the restart frequency k is 50, the first restart cycle is performed by GMRES, and the condition numbers of the small matrices $Q_n^* A^* A Q_n$, $1 \leq n \leq 50$ (after diagonal scaling is applied), are given for the second restart cycle. Though the choice of $u^{(0)} = 0$ (and thus $r^{(0)} = b$) is used, other choices of $r^{(0)}$ also gave similar results to those shown below.

We give comparative results in Figures 1, 2 and 3 with $Ch_x = 0$, 4 and 128 respectively, for the power basis, the Chebyshev basis provided with the exact spectrum rectangle for the problem, the GMRES/Chebyshev algorithm with several eigenvalue estimation schemes (the Cassini ovals technique for $\tilde{H}_{n,H}$ and $\tilde{H}_{n,N}$ described in Section 4, a full eigenvalue decomposition of $\tilde{H}_n$, and full eigenvalue decompositions of $\tilde{H}_{n,H}$ and $\tilde{H}_{n,N}$), and finally the GMRES/repeated basis algorithm. We note that for the second and third test problems, A has some non-real eigenvalues.

The following observations on the small matrix condition may be made. First, the power basis gives very strong exponential growth. The Chebyshev basis with ellipse based on the spectral estimates from the Cassini ovals technique is better but still grows exponentially; apparently this is due to a mismatch between the ellipse and the spectrum of A. On the other hand, the Chebyshev basis algorithm performed extremely well, even showing a sub-exponential behavior for the symmetric case. Furthermore, the GMRES/Chebyshev tests based on the full eigenvalue estimation techniques on the small matrix $\tilde{H}_k$ or on $\tilde{H}_{k,H}$ and $\tilde{H}_{k,N}$ from the first GMRES cycle performed nearly as well. The GMRES/repeated basis algorithm also performed well,
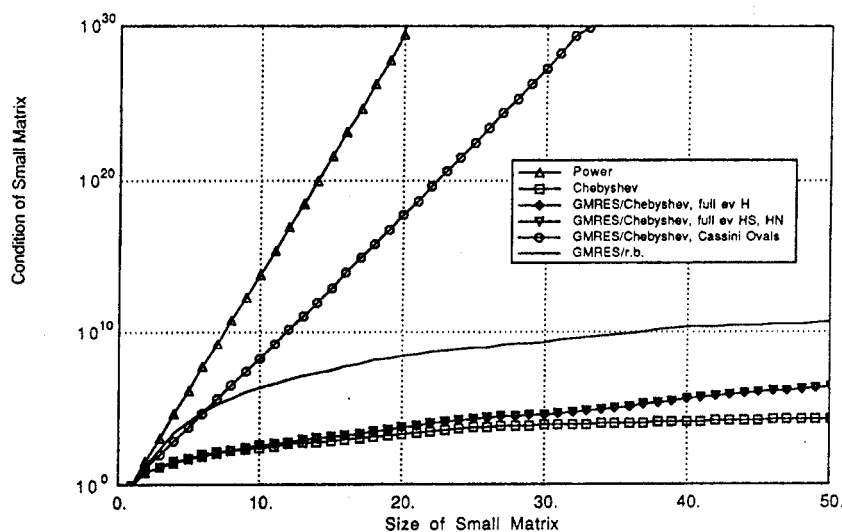
Figure 1   Model problem, $Ch_x = 0$.

particularly in the nonsymmetric case for which cond V is large $(Ch_x = 4)$ (see Section 3).

The condition size is also significant in comparison to the number of digits representable on the given computer. These runs suggest that a Chebyshev basis algorithm may be practical for many problems when using 64-bit word length machines (having about 15 digits accuracy), for most of the common choices of k, e.g. $k \leq 50$. The potential for savings in these cases is substantial.
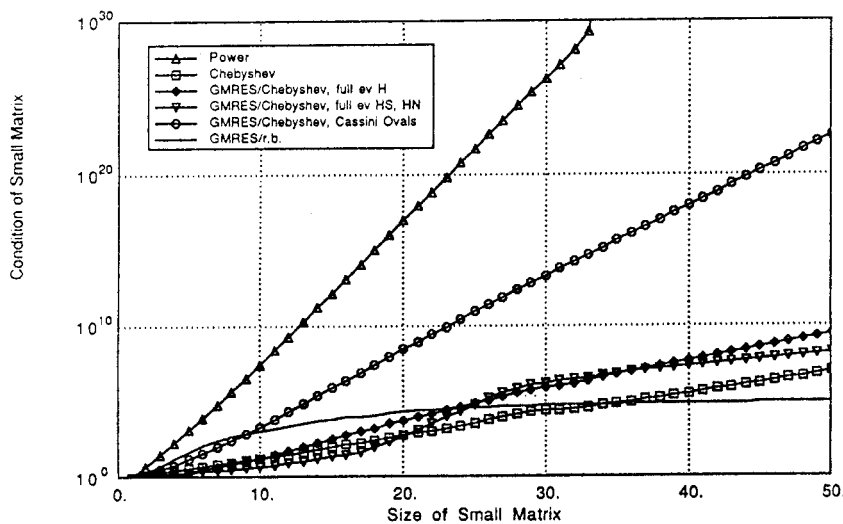
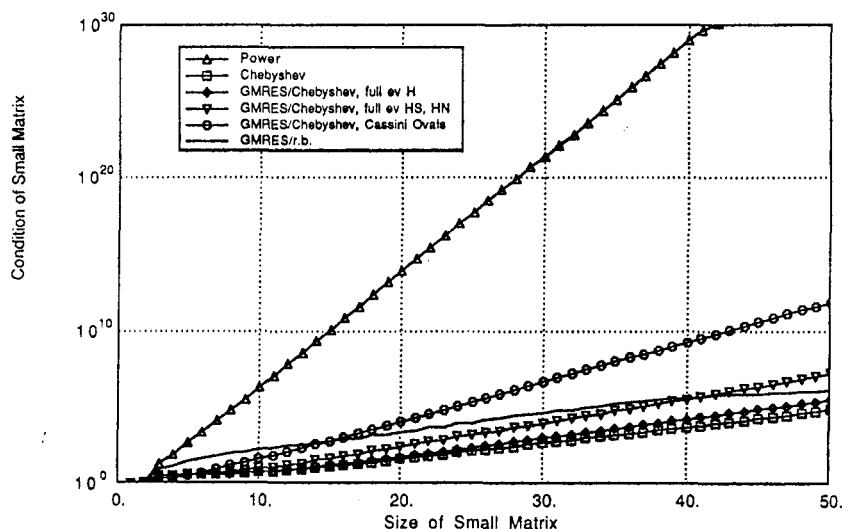**Figure 2**   Model problem, $Ch_x = 4$.

Figure 3   Model problem, $Ch_x = 128$.

6   CONCLUSIONS

In this paper we have presented several new algorithms which give the same iterates as the restarted GMRES algorithm but offer decreased computational expense.

The algorithms exhibit good numerical properties. When 64-bit precision arithmetic is used, the results in this paper suggest that these algorithms should be sufficiently numerically stable to solve a wide range of common problems. Furthermore, the new algorithms are "safe", in the sense that if an ill-conditioned basis is generated for one cycle, then the iterate is not corrupted by this, and if it is desired a switch can be made to the safer GMRES algorithm for the remaining steps.

Details of the parallel implementation of these algorithms and further numerical experiments are given in Part II.

References

Ashby, S. F., T. A. Manteuffel and P. E. Saylor, A taxonomy of conjugate gradient methods, *SIAM J. Numer. Anal.*, **27** (6) (1990), 1542–1568.

Bai, Z., D. Hu and L. Reichel, *A Newton Basis GMRES Implementation*, University of Kentucky, Department of Mathematics, Lexington, KY, Research Report 91-03, April 1991.

Barragy, E. and G. F. Carey, A parallel element-by-element solution scheme, *International Journal for Numerical Methods in Engineering*, **26** (1989), 2367–2382.

Businger, P. A., *Matrix Scaling with Respect to the Maximum-Norm, the Sum-Norm and the Euclidean Norm*, Ph.D. thesis, the University of Texas at Austin, August 1967, Report TNN-71, the University of Texas Computation Center, August 1967.

Carey, G. F. and E. Barragy, Basis function selection and preconditioning high degree finite element and spectral methods, *BIT*, **29** (1989), 794–804.

Chronopoulos, A., *A Class of Parallel Iterative Methods Implemented on Multiprocessors*, Report UIUCDCS-R-86-1267, University of Illinois at Urbana-Champaign, Ph.D. Thesis, November 1986.

Chronopoulos, A., T. and S. K. Kim, *s-Step Orthomin and GMRES Implemented on Parallel Computers*,

Technical Report TR 90-15, Department of Computer Science, University of Minnesota, Minneapolis, Minesota, February 1990.

Davis, P. J., *Interpolation and Approximation*. New York: Blaisdell, 1963.

Eisenstat, S. C., H. C. Elman and M. H. Schultz, Variational iterative methods of nonsymmetric systems of linear equations, *SIAM J. Numer. Anal.*, **20** (2) (1983), 345–357.

Faber, V. and T. Manteuffel, Necessary and sufficient conditions for the existence of a conjugate gradient method, *SIAM J. Numer. Anal.*, **21**(2) (1984), 352–362.

Faber, V. and T. Manteuffel, Orthogonal error methods, *SIAM J. Numer. Anal.*, **24**(1) (1987), 170–187.

Gautschi, W., The condition of orthogonal polynomials, *Mathematics of Computation*, **26**(120) (1972), 923–924.

Gautschi, W., The condition of polynomials in power form, *Mathematics of Computation*, **33**(145) (1979), 343–352.

Gautschi, W., *Questions of Numerical Condition Related to Polynomials*, in Studies in Numerical Analysis, Gene Golub, ed., Mathematical Association of America, 1984, pp. 140–177.

Golub, G. H. and C. F. Van Loan, *Matrix Computations*, second edition. Baltimore: John Hopkins University Press, 1989.

Grcar, J. F., Operator coefficient methods for linear equations, Report SAND89-8691, Sandia National Laboratory, November 1989.

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

Hestenes, M. R. and E. L. Stiefel, Methods of conjugate gradients for solving linear systems, *J. Res. Nat. Bur. Standards*, **49** (1952), 409–436.

Householder, A. S., *The Theory of Matrices in Numerical Analysis*. New York: Dover, 1964.

Joubert, W. D., *Generalized Conjugate Gradient and Lanczos Methods for the Solution of Nonsymmetric Systems of Linear Equations*, the University of Texas at Austin, Ph.D. thesis, Report CNA-238, Center for Numerical Analysis, Austin, Texas, January 1990.

Joubert, W., *Iterative Methods for the Solution of Nonsymmetric Systems of Linear Equations*, the University of Texas at Austin, Center for Numerical Analysis, Report CNA-242, February 1990.

Joubert, W. D. and T. A. Manteuffel, *Iterative Methods for Nonsymmetric Linear Systems*, in *Iterative Methods for Large Linear Systems*, David R. Kincaid and Linda J. Hayes eds. Boston: Academic Press, 1990, pp. 149–171.

Joubert, W. D. and D. M. Young, Necessary and sufficient conditions for the simplification of generalized conjugate gradient algorithms, *Linear Algebra and its Applications*, **88/89** (1987), 449–485.

Manteuffel, T. A., The Tchebychev iteration for nonsymmetric linear systems, *Numerische Mathematik*, **28** (1977), 307–327.

Manteuffel, T. A., Adaptive procedure for estimation of parameters for the nonsymmetric Tchebychev iteration, *Numerische Mathematik*, **31** (1978), 187–208.

Nachtigal, N. M., L. Reichel and L. N. Trefethen, a hybrid GMRES algorithm for nonsymmetric linear systems, *SIAM J. Matrix Anal. Appl.*, to appear..

Rivlin, T. J., *The Chebyshev Polynomials*. New York: Wiley, 1976.

Saad, Y., On the condition number of some gram matrices arising from least squares approximation in the complex plane, *Numerische Mathematik*, **48** (1986), 337–347.

Saad, Y. and M. H. Schultz, Conjugate gradient-like algorithms for solving nonsymmetric linear systems, *Mathematics of Computation*, **44**(170) (1985), 417–424.

Saad, Y. and M. H. Schultz, GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems, *SIAM J. Sci. Stat. Comp.*, **7**(3) (1986), 856–869.

Saylor, P. E., *Leapfrog Variants of Iterative Methods for Linear Algebraic Equations*, Report UIUCDCS-R-87-1373, Department of Computer Science, University of Illinois at Urbana-Champaign, December 1987.

Vinsome, P. K. W., *ORTHOMIN, an Iterative Method for Solving Sparse Sets of Simultaneous Linear Equations*, in *4th Symposium of Numerical Simulation of Resevoir Performance of the Society of Petroleum Engineers of the AIME*, Los Angeles, Calif., 1976, Paper SPE 5739.

Walker, H. F., Implementation of the GMRES method using Householder transformations, *SIAM J. Sci. Stat. Comput.*, **9** (1988), 152–163.

Walker, H. F., Implementations of the GMRES method, *Computer Physics Communications*, **53** (1989), 311–20.

Young, D. M. and K. C. Jea, Generalized conjugate-gradient acceleration of nonsymmetrizable iterative methods, *Linear Algebra and its Applications*, **34** (1980), 159–194.