Variational Iterative Methods for Nonsymmetric Systems of Linear Equations

Stanley C. Eisenstat
Howard C. Elman
Martin H. Schultz

Research Report #209

August, 1981

This work was supported in part by ONR Grant N00014-76-C-0277, FSU-ONR Grant F1.N00014-80-C-0076, and DOE Grant DE-AC02-77ET53053.
Abstract

We consider a class of iterative algorithms for solving systems of linear equations where the coefficient matrix is nonsymmetric with positive-definite symmetric part. The algorithms are modelled after the conjugate gradient method, and are well-suited for large sparse systems. They do not make use of any associated symmetric problems. Convergence results and error bounds are presented.
1. Introduction

The conjugate gradient method (CG), first described by Hestenes and Stiefel [8], is widely used for approximating the solutions of large sparse systems of linear equations

\[ A x = f \]

where \( A \) is an \( N \times N \), real, symmetric, positive-definite matrix [1, 3, 5, 12]. CG can be viewed as a direct method that, in the absence of round-off error, gives the exact solution in at most \( N \) steps; or as an iterative procedure that gives a good approximation to the solution in far fewer steps (see [13]). A feature of the method that makes it particularly suitable for large sparse systems is that all references to \( A \) are in the form of a matrix-vector product \( Av \), so that the storage requirements are usually lower than for direct methods. Another attractive feature is that, unlike most iterative methods, CG does not require any estimation of parameters. In this paper, we discuss a class of conjugate-gradient-like descent methods that can be used to solve nonsymmetric systems of linear equations. Numerical experiments with these methods are described in [6, 7].

A common technique [8] for solving nonsymmetric problems is to apply the conjugate gradient method to the normal equations

\[ A^T A x = A^T f, \]

in which the coefficient matrix is symmetric and positive-definite. On the \( i \)'th iteration, CG computes an approximate solution that is in some sense optimal in a Krylov subspace of the form \( \{v, A^T Av, \ldots, (A^T A)^{i-1} v\} \). This dependence on \( A^T A \) tends to make the convergence of CG slow (see [1], [3]).
Recently, Concus and Golub [4] and Widlund [18] devised a generalized conjugate gradient algorithm (GCG) for nonsymmetric systems in which the coefficient matrix has positive-definite symmetric part. Like the conjugate gradient method, GCG gives the exact solution in at most \( N \) iterations. However, on each iteration it requires the solution of an auxiliary system of equations in which the coefficient matrix is the symmetric part of \( A \). Also, if the nonsymmetric part is relatively large, then convergence may be slow.

The methods we present depend on a Krylov sequence based on \( A \) rather than \( A^T A \), and they do not require the solution of any auxiliary systems. They do require that the symmetric part of \( A \) be positive-definite. In Section 2, we present four variants that differ in their work and storage requirements. In Sections 3 and 4, we present convergence results and error bounds for each of the four variants. In Section 5, we discuss several alternative formulations.

**Notation**

The symmetric part of the coefficient matrix \( A \) is given by \( M := \frac{A + A^T}{2} \), and the skew-symmetric part by \( R := -\frac{A - A^T}{2} \). Thus \( A = M - R \). The Jordan canonical form of \( A \) is denoted by \( J := T^{-1} A T \).

For any square matrix \( X \), let \( \lambda_{\text{min}}(X) \) denote the eigenvalue of \( X \) of smallest absolute value, and let \( \lambda_{\text{max}}(X) \) denote the eigenvalue of largest absolute value. The spectral radius \( |\lambda_{\text{max}}(X)| \) of \( X \) is denoted by \( \rho(X) \). The set of eigenvalues of \( X \), also called the spectrum of \( X \), is denoted by \( \sigma(X) \). If \( X \) is nonsingular, then the condition number of \( X \), \( \kappa(X) \), is defined as \( \|X\|_2 \|X^{-1}\|_2 \).
Finally, given a set of vectors \( \{p_0, \ldots, p_k\} \), let \( \langle p_0, \ldots, p_k \rangle \) denote the space spanned by \( \{p_0, \ldots, p_k\} \).

2. Descent Methods for Nonsymmetric Systems

In this section, we present a class of descent methods for solving the system of linear equations

\[
A \mathbf{x} = \mathbf{f}
\]

(2.1)

where \( A \) is a nonsymmetric matrix of order \( N \) with positive-definite symmetric part. We consider four variants, all of which have the following general form:

(2.2a) Choose \( x_0 \).

(2.2b) Compute \( r_0 = f - Ax_0 \).

(2.2c) Set \( p_0 = r_0 \).

FOR \( i = 0 \) STEP 1 UNTIL Convergence DO

(2.2d) \[
a_i = \frac{(r_i, A p_i)}{(A p_i, A p_i)}
\]

(2.2e) \[
x_{i+1} = x_i + a_i p_i
\]

(2.2f) \[
r_{i+1} = r_i - a_i A p_i
\]

(2.2g) Compute \( p_{i+1} \).

The choice of \( a_i \) in (2.2d) minimizes \( ||r_{i+1}||_2 = ||f - A(x_i + a p_i)||_2 \) as a function of \( a \), so that the Euclidean norm of the residual decreases at each step. The variants differ in the technique used to compute the new direction vector \( p_{i+1} \).
A good choice for $p_{i+1}$ is one that results in a significant decrease in the norm of the residual $\|r_{i+1}\|_2$ but does not require a large amount of work to compute. When $A$ is symmetric and positive-definite, such a vector can be computed by the simple recurrence relation

$$(2.3a) \quad p_{i+1} = r_{i+1} + b_i p_i,$$

where

$$(2.3b) \quad b_i = -\frac{(A r_{i+1}, A p_i)}{(A p_i, A p_i)}.$$

The method defined by (2.2) and (2.3) is equivalent to a variant of CG known as the conjugate residual method (CR) [16]. The direction vectors produced are $A^T A$-orthogonal, that is

$$(2.4) \quad (A p_i, A p_j) = 0 \,, \quad \text{for } i \neq j \,,
$$

and $x_{i+1}$ minimizes the functional

$$E(w) := \|f - Aw\|_2$$

over the affine space $x_0 + \langle p_0, \ldots, p_i \rangle$.

If $A$ is nonsymmetric and the algorithm defined by (2.2) and (2.3) is applied to (2.1), then the orthogonality relation (2.4) does not hold in general. However, a set of $A^T A$-orthogonal directions can be generated by using all the previous vectors $\{p_j\}_{j=0}^i$ to compute $p_{i+1}$:

$$(2.5a) \quad p_{i+1} = r_{i+1} + \sum_{j=0}^{i} b_j^{(i)} p_j,$$

where

$$(2.5b) \quad b_j^{(i)} = -\frac{(A r_{i+1}, A p_j)}{(A p_j, A p_j)} \,, \quad j \leq i \,.$$

The iterate $x_{i+1}$ generated by (2.2) and (2.5) minimizes $E(w)$ over $x_0 + \langle p_0, \ldots, p_i \rangle$ (see Theorem 3.1). We refer to this algorithm as the generalized conjugate residual method (GCR). In the absence of roundoff
error, GCR gives the exact solution to (2.1) in at most N iterations (see Corollary 3.2).

The work and storage requirements per iteration of GCR may be prohibitively high when N is large. Vinsome [17] has proposed a method called Orthomin that can be viewed as a modification of GCR that is significantly less expensive per iteration. Instead of making \( p_{i+1} A^T A \)–orthogonal to all the preceding direction vectors \( \{ p_j \}_{j=0}^i \), one can make \( p_{i+1} \) orthogonal to only the last \( k \geq 0 \) vectors \( \{ p_j \}_{j=i-k+1}^i \):

\[
(2.6) \quad p_{i+1} = r_{i+1} + \sum_{j=i-k+1}^{i} b(j) p_j,
\]

with \( \{ b(j) \}_{j=i-k+1}^i \) defined as in (2.5b).\(^1\) Only \( k \) direction vectors need be saved. We refer to this method as Orthomin(k) (see [19]). Both GCR and Orthomin(k) for \( k \geq 1 \) are equivalent to the conjugate residual method when \( A \) is symmetric and positive–definite.

Another alternative is to restart GCR periodically: every \( k+1 \) iterations, the current iterate \( x_j^{(k+1)} \) is taken as the new starting guess.\(^2\) At most \( k \) direction vectors have to be saved, so that the storage costs are the same as for Orthomin(k). However, the cost per iteration is lower, since in general fewer than \( k \) direction vectors are used to compute \( p_{i+1} \). We refer to this restarted method as GCR(k).

\(^1\) The first \( k \) directions \( \{ p_j \}_{j=0}^{k-1} \) are computed by (2.5a), as in GCR.

\(^2\) Here \( j \) is a counter for the number of restarts. The \( j^{th} \) cycle of GCR(k) produces the sequence of approximate solutions \( \{ x_i \}_{i=(j-1)(k+1)+1}^{(j-1)(k+1)+k+1} \).
For the special case \( k = 0 \), Orthomin(k) and GCR(k) are identical, with
\[
P_{i+1} = r_{i+1}.
\]
This method, which we refer to as the minimum residual method (MR), has very modest work and storage requirements, and in the symmetric case resembles the method of steepest descent (see [10]). Because of its simplicity, we consider it separately from Orthomin(k) and GCR(k).

In Table I, we summarize the work and storage costs (excluding storage for \( A \)) of computing \( x_i \) for each of the methods. The entries in this table are determined as follows. For GCR, the storage cost includes space for the vectors \( x_i, r_i, A_r_i, p_0, \ldots, p_{i-1}, \) and \( A_p_0, \ldots, A_p_{i-1} \). The work/iteration includes two inner-products for \( A_{i-1} \), two scalar-vector products for \( x_i \) and \( r_i \), \( i \) inner-products for \( (b^{(i-1)}_j)_{j=0}^{i-1} \), \( i \) scalar-vector products for \( p_i \), \( i \) scalar-vector products to compute \( A_p_i \) by
\[
A_p_i = A_r_i + \sum_{j=0}^{i-1} b^{(i-1)}_j A_p_j,
\]
and one matrix-vector product \( A_r_i \). The total is thus \( (3i+4)N + 1 \) mv. The entries for Orthomin(k) correspond to the requirements after the \( k \)th iteration, and are the same as those for the \( k \)th iteration of GCR. The work given for GCR(k) is the average over \( k+1 \) iterations. The cost of MR is the same as the cost of Orthomin(0) or GCR(0). \(^3\)

\(^3\) Several other implementations are possible. In Orthomin(k) or GCR(k), it may be cheaper to compute \( A_p_i \) by a matrix-vector product for large \( k \). With a third matrix-vector product, \( b^{(i)}_j \) can be computed as
\[
-(A^T A_r_{i+1} p_j) / (A_p_j A_p_j),
\]
and the previous \( \{A_p_j\} \) need not be saved.
<table>
<thead>
<tr>
<th>GCR</th>
<th>Orthomin(k)</th>
<th>GCR(k)</th>
<th>MR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Work/</td>
<td>(3i+4)N</td>
<td>(3k+4)N</td>
<td>((3/2)k+4)N</td>
</tr>
<tr>
<td>Iteration</td>
<td>+ 1 mv</td>
<td>+ 1 mv</td>
<td>+ 1 mv</td>
</tr>
<tr>
<td>Storage</td>
<td>(2i+3)N</td>
<td>(2k+3)N</td>
<td>(2k+3)N</td>
</tr>
</tbody>
</table>

Table I  Work per iteration (mv denotes a matrix-vector product) and storage requirements.

3. Convergence of GCR and GCR(k)

In this section, we show that GCR gives the exact solution in at most N iterations and present error bounds for GCR and GCR(k). We first establish a set of relations among the vectors generated by GCR. (See [8] for an analogous result for the conjugate gradient method.)

Theorem 3.1. If \( \{x_i\} \), \( \{r_i\} \), and \( \{p_i\} \) are the iterates generated by GCR in solving the linear system (2.1), then the following relations hold:

(3.1a) \( (A_{p_i},A_{p_j}) = 0 \) , \( i \neq j \);

(3.1b) \( (r_i,A_{p_j}) = 0 \) , \( i > j \);

(3.1c) \( (r_i,A_{p_j}) = (r_i,A_{r_i}) \);

(3.1d) \( (r_i,A_{r_j}) = 0 \) , \( i > j \);

(3.1e) \( (r_i,A_{p_j}) = (r_0,A_{p_j}) \) , \( i < j \);

(3.1f) \( \langle p_0,\ldots,p_i \rangle = \langle p_0,A_{p_0},\ldots,A_{i}^{t}A_{0} \rangle = \langle r_0,\ldots,r_i \rangle \);

(3.1g) if \( r_i \neq 0 \), then \( p_i \neq 0 \);

(3.1h) \( x_{i+1} \) minimizes \( E(w) = \|f-Aw\|^2 \) over the affine space \( x_0 + \langle p_0,\ldots,p_i \rangle \).

Proof. The directions \( \{p_i\} \) are chosen so that (3.1a) holds.
Relation (3.1b) is proved by induction on $i$. It is vacuously true for $i = 0$. Assume that it holds for $i \leq t$. Then, using (2.2f) and taking the inner product with $A p_j$,

\[(r_{t+1}, A p_j) = (r_t, A p_j) - a_t \ (A p_t, A p_j) \quad .\]

If $j < t$, then the terms on the right-hand side are zero by the induction hypothesis and (3.1a). If $j = t$, then the right-hand side is zero by the definition of $a_t$. Hence (3.1b) holds for $i = t+1$.

For (3.1c), by premultiplying (2.5a) by $A$ and taking the inner product with $r_i$,

\[
(r_i, Ar_i) = (r_i, Ar_i) + \sum_{j=0}^{i} b^{(i-1)} (r_i, A p_j)
= (r_i, Ar_i) ,
\]

since all the terms in the sum are zero by (3.1b).

To prove (3.1d), we rewrite (2.5a) as

\[
r_j = p_j - \sum_{t=0}^{j-1} b^{(j-1)} (r_i, A p_t) .
\]

Premultiplying by $A$ and taking the inner product with $r_i$ ($i > j$),

\[
(r_i, Ar_j) = (r_i, Ar_j) - \sum_{t=0}^{j-1} b^{(j-1)} (r_i, A p_t)
= 0 ,
\]

by (3.1b).

Relation (3.1e) is proved by induction on $i$, for $i \leq j$. It is trivially true when $i = 0$. Assume that it holds for $i = t < j$. Using (3.2),

\[
(r_{t+1}, A p_j) = (r_t, A p_j) - a_t \ (A p_t, A p_j)
\]
by the induction hypothesis and (3.1a).

Relation (3.1f) is proved by induction on j. The three spaces are identical when j = 0. Assume that they are identical for j ≤ t. Then $\{p_j\}_{j=0}^t \subseteq \langle r_0, \ldots, r_{t+1} \rangle$. But by (2.5a),

$$p_{t+1} = r_{t+1} + \sum_{j=0}^t b(t) p_j,$$

so that $\langle p_0, \ldots, p_{t+1} \rangle$ is a subspace of $\langle r_0, \ldots, r_{t+1} \rangle$. By (3.1a), the vectors $\{p_j\}_{j=0}^{t+1}$ are linearly independent. Hence, the dimension of $\langle r_0, \ldots, r_{t+1} \rangle$ is greater than or equal to $t+1$, which implies that $\{r_j\}_{j=0}^{t+1}$ are linearly independent and $\langle p_0, \ldots, p_{t+1} \rangle = \langle r_0, \ldots, r_{t+1} \rangle$. Similarly, by (2.2f),

$$p_{t+1} = r_t - a_t A p_t + \sum_{j=0}^t b(t) p_j.$$

By the induction hypothesis, $r_t$, $A p_t$, and $\{p_j\}_{j=0}^t \subseteq \langle p_0, A p_0, \ldots, A^{t+1} p_0 \rangle$, so that $\langle p_0, \ldots, p_{t+1} \rangle$ is a subspace of $\langle p_0, A p_0, \ldots, A^{t+1} p_0 \rangle$. Again, the two spaces are equal because the $\{p_j\}$ are linearly independent.

The proof of (3.1g) depends on the fact that the symmetric part M of A is positive-definite. If $r_i \neq 0$, then by (3.1c),

$$(r_i, A p_i) = (r_i, A r_i) = (r_i, M r_i) > 0,$$

so that $(r_i, A p_i) \neq 0$, whence $p_i \neq 0$.

For the proof of (3.1h), note that

$$x_{i+1} = x_0 + \sum_{j=0}^i a_j p_j.$$

Thus, $E(x_{i+1})^2$ is a quadratic functional in $a = (a_0, \ldots, a_i)^T$. Indeed, using (3.1a) to simplify the quadratic term,
\[ E(x_{i+1})^2 = \|r_0 - \sum_{j=0}^{i} a_j A p_j \|^2 \]
\[ = (r_0, r_0) - 2 \sum_{j=0}^{i} a_j (r_0, A p_j) + \sum_{j=0}^{i} a_j^2 (A p_j, A p_j) \]

Thus, \( E(w) \) is minimized over \( x_0 + \langle p_0, \ldots, p_i \rangle \) when
by (3.1e).

Q.E.D.

**Corollary 3.2.** GCR gives the exact solution to (2.1) in at most \( N \) iterations.

**Proof.** If \( r_i = 0 \) for some \( i \leq N-1 \), then \( A x_i = f \) and the assertion is proved. If \( r_i \neq 0 \) for all \( i \leq N-1 \), then \( p_i \neq 0 \) for all \( i \leq N-1 \) by (3.1g). By (3.1a), \( \{p_i\}_{i=0}^{N-1} \) are linearly independent, so that \( \langle p_0, \ldots, p_{N-1} \rangle = \mathbb{R}^N \).
Hence, by (3.1h), \( x_N \) minimizes the functional \( E \) over \( \mathbb{R}^N \), i.e., \( x_N \) is the solution to the system.

Q.E.D.

This result does not give any insight into how close \( x_i \) is to the solution of (2.1) for \( i < N \). We now derive an error bound for GCR that proves that GCR converges as an iterative method. Let \( P_i \) denote the set of real polynomials \( q_i \) of degree less than or equal to \( i \) such that \( q_i(0) = 1 \).

**Theorem 3.3.** If \( \{r_i\} \) is the sequence of residuals generated by GCR, then
\[ \|r_i\|_2 \leq \min_{q_i \in P_i} \|q_i(A)\|_2 \|r_0\|_2 \leq \left[ 1 - \frac{\lambda_{\min}(M)^2}{\lambda_{\max}(AA^T)} \right]^{i/2} \|r_0\|_2 \]
Hence, GCR converges. If \( A \) has a complete set of eigenvectors, then

\[
\|r_i\|_2 \leq K(T) \ M_i \ |r_0|_2 ,
\]

where

\[
M_i := \min_{\lambda \in \sigma(A)} \ \max_{q_i \in P_i} |q_i(\lambda)| .
\]

Moreover, if \( A \) is normal, then

\[
\|r_i\|_2 \leq M_i \ |r_0|_2 .
\]

**Proof.** By (3.1f), the residuals \( \{r_i\} \) generated by GCR are of the form

\[
r_i = q_i(A)r_0 \text{ for some } q_i \in P_i.
\]

By (3.1h),

\[
\|r_i\|_2 = \min_{q_i \in P_i} |q_i(A)r_0|_2 .
\]

The first inequality of (3.3) is an immediate consequence of (3.6). To prove the second inequality of (3.3), note that for \( q_1(z) = 1 + az \in P_i \),

\[
\min_{q_i \in P_i} |q_i(\lambda)|_2 \leq |q_i(A)|_2 \leq |q_1(\lambda)|_2 .
\]

But

\[
|q_1(\lambda)|_2^2 = \max_{x \neq 0} \frac{((I+\alpha A)x, (I+\alpha A)x)}{(x,x)}
\]

\[
= \max_{x \neq 0} \left[ 1 + 2\alpha \frac{(x,Ax)}{(x,x)} + \alpha^2 \frac{(Ax,Ax)}{(x,x)} \right] .
\]

Moreover,

\[
\frac{(Ax,Ax)}{(x,x)} = \frac{(x,A^T Ax)}{(x,x)} \leq \lambda_{\max} (A^T A) ,
\]

and, using the positive-definiteness of \( M \),

\[
\frac{(x,Ax)}{(x,x)} = \frac{(x,Mx)}{(x,x)} \geq \lambda_{\min} (M) > 0 .
\]

Hence, if \( \alpha < 0 \),
\[ ||q_1(A)||^2_2 \leq 1 + 2 \lambda_{\min}(M) \alpha + \lambda_{\max}(A^T A) \alpha^2. \]

This expression is minimized by \( \alpha = -\frac{\lambda_{\min}(M)}{\lambda_{\max}(A^T A)} \), and with this choice of \( \alpha \),

\[ ||q_1(A)||^2 \leq \left[ 1 - \frac{\lambda_{\min}(M)^2}{\lambda_{\max}(A^T A)} \right]^{1/2}, \]

which concludes the proof of (3.3).

Recall that the Jordan canonical form of \( A \) is given by \( J = T^{-1} A T \). To prove (3.4), we rewrite (3.6) as

\[ ||r_i||_2 = \min_{q_i \in P_i} ||T q_i(J) T^{-1} r_0||_2 \]

\[ \leq ||T||_2 ||T^{-1}||_2 \min_{q_i \in P_i} ||q_i(J)||_2 ||r_0||_2. \]

Since \( A \) has a complete set of eigenvectors, \( J \) is diagonal, so that

\[ \min_{q_i \in P_i} ||q_i(j)||_2 = \min_{q_i \in P_i} \max_{\lambda \in \sigma(A)} |q_i(\lambda)|, \]

whence (3.4) follows.

If \( A \) is normal, then \( T \) can be chosen to be an orthonormal matrix, which proves (3.5).

Q.E.D.

Since the symmetric part of \( A \) is positive-definite, the spectrum of \( A \) lies in the open right half of the complex plane (see [9]). Thus, the analysis of Manteuffel [11] shows that \( \min_{q_i \in P_i} ||q_i(A)||_2 \) and \( M_i \) approach zero as \( i \) goes to infinity, which also implies that GCR converges.
Theorem 3.3 can also be used to establish an error bound for GCR(k).

Corollary 3.4. If \( \{ r_i \} \) is the sequence of residuals generated by GCR(k), then

\[
\| r_{j(k+1)} \|_2 \leq \left[ \min_{q_{k+1} \in P_{k+1}} \| q_{k+1}(A) \|_2 \right]^j \| r_0 \|_2 ,
\]

so that

\[
\| r_i \|_2 \leq \left[ 1 - \frac{\lambda_{\min}(M)^2}{\lambda_{\max}(A^T A)} \right]^{i/2} \| r_0 \|_2 .
\]

Hence, GCR(k) converges. Moreover, if \( A \) has a complete set of eigenvectors, then

\[
\| r_{j(k+1)} \|_2 \leq (K(T) M_{k+1})^j \| r_0 \|_2 ,
\]

and if \( A \) is normal, then

\[
\| r_{j(k+1)} \|_2 \leq (M_{k+1})^j \| r_0 \|_2 .
\]

Proof. Assertions (3.7), (3.9), and (3.10) follow from Theorem 3.3. To prove (3.8), let \( i = jk + t \) where \( 0 \leq t < k \). Then

\[
\| r_{jk+t} \|_2 \leq \left[ 1 - \frac{\lambda_{\min}(M)^2}{\lambda_{\max}(A^T A)} \right]^{t/2} \| r_{jk} \|_2 ,
\]

by (3.3), and

\[
\| r_{jk} \|_2 \leq \left[ 1 - \frac{\lambda_{\min}(M)^2}{\lambda_{\max}(A^T A)} \right]^{jk/2} \| r_0 \|_2 ,
\]

by (3.7) and the second inequality of (3.3).

Q.E.D.
4. Convergence of Orthomin

In this section, we present convergence results for Orthomin(k) and an alternative error bound for GCR and GCR(k). We also present an analysis of Orthomin in the special case when the symmetric part of $A$ is the identity.

The vectors generated by Orthomin(k) satisfy a set of relations analogous to (3.1):

**Theorem 4.1.** The iterates $\{x_i\}$, $\{r_i\}$, and $\{p_i\}$ generated by Orthomin(k) satisfy the relations:

\begin{align}
(4.1a) \quad (A_{p_i}A_{j}) &= 0, \quad j = i-k, \ldots, i-1, \quad i \geq k; \\
(4.1b) \quad (r_i,A_{j}) &= 0, \quad j = i-k-1, \ldots, i-1, \quad i \geq k+1; \\
(4.1c) \quad (r_i,A_{p_i}) &= (r_i,A_{r_i}); \\
(4.1d) \quad (r_i,A_{r_{i-1}}) &= 0; \\
(4.1e) \quad (r_i,A_{p_j}) &= (r_{j-k},A_{p_j}), \quad i = j-k, \ldots, j-1, \quad j \geq k; \\
(4.1f) \quad \text{if } r_i \neq 0, \text{ then } p_i \neq 0; \\
(4.1g) \quad \text{for } i \geq k, \text{ } x_{i+1} \text{ minimizes } E(w) \text{ over the affine space } x_{i-k} + \langle p_{i-k}, \ldots, p_i \rangle.
\end{align}

Corollary 3.4 with $k = 0$ implies that Orthomin(0) (MR) converges. We now prove that Orthomin(k) converges for $k > 0$. Since the analysis applies as well to GCR, GCR(k), and MR, we state the results in terms of all four methods. Recalling that $R$ is the skew-symmetric part of $A$, we first prove two preliminary results:

**Lemma 4.2.** The direction vectors $\{p_i\}$ and the residuals $\{r_i\}$ generated by GCR, Orthomin(k), GCR(k), and MR satisfy

\begin{align}
(4.2) \quad (A_{p_i}A_{p_i}) \preceq (A_{r_i}A_{r_i}).
\end{align}
Proof. The direction vectors are given by

\[ p_i = r_i + \sum b_j^{(i-1)} p_j, \]

where the limits of the sum are defined as in (2.5) for GCR and GCR(k), and (2.6) for Orthomin(k). Therefore, by the \( A^T A \)-orthogonality of the \( \{p_i\} \) and the definition of \( b_j^{(i-1)} \),

\[
(A_{p_i}A_{p_i}) = (A_{r_i}A_{r_i}) + 2 \sum b_j^{(i-1)} (A_{r_i}A_{p_j}) + \sum (b_j^{(i-1)})^2 (A_{p_j}A_{p_j})
\]

\[ = (A_{r_i}A_{r_i}) - \sum \frac{(A_{r_i}A_{p_i})^2}{(A_{p_j}A_{p_j})} \]

\[ \leq (A_{r_i}A_{r_i}). \]

Q.E.D.

Lemma 4.3. For any real \( x \neq 0 \),

\[ \frac{(x, Ax)}{(Ax, Ax)} \geq \frac{\lambda_{\min}(M)}{\lambda_{\min}(M)\lambda_{\max}(M) + \rho(R)^2}. \]

Proof. Letting \( y = Ax \),

\[ \frac{(x, Ax)}{(Ax, Ax)} = \frac{(y, A^{-1}y)}{(y, y)} = \frac{(y, \frac{A^{-1} + A^{-T}}{2}y)}{(y, y)} \geq \lambda_{\min} \left( \frac{A^{-1} + A^{-T}}{2} \right). \]

Thus, it suffices to bound \( \lambda_{\min}( \frac{A^{-1} + A^{-T}}{2} ) \). Consider the identity

\[ X^{-1} + Y^{-1} = [Y(X+Y)^{-1}X]^{-1}, \]

which holds for any nonsingular matrices \( X \) and \( Y \), provided that \( X+Y \) is nonsingular. With \( X = 2A \) and \( Y = 2A^T \), (4.4) leads to

\[ \frac{A^{-1} + A^{-T}}{2} = [(2A)^T (4M)^{-1} (2A)]^{-1} = [(M - R^T M^{-1} (M - R)]^{-1} \]
\[ = (M + R^T M^{-1} R)^{-1} \]

For any \( x \neq 0 \),

\[
(x, (M + R^T M^{-1} R)x) = (x, Mx) + (Rx, M^{-1} Rx) > 0
\]

so that \( M + R^T M^{-1} R \) is positive-definite. Therefore \( \frac{A^{-1} + A^{-T}}{2} \) is positive-definite and

\[
\lambda_{\min}(\frac{A^{-1} + A^{-T}}{2}) = \frac{1}{\lambda_{\max}(M + R^T M^{-1} R)}.
\]

But

\[
\lambda_{\max}(M + R^T M^{-1} R) = \max_{x \neq 0} \left[ \frac{(x, Mx)}{(x, x)} + \frac{(x, R^T M^{-1} Rx)}{(x, x)} \right]
\]

\[
\leq \lambda_{\max}(M) + \max_{x \neq 0, Rx \neq 0} \frac{(Rx, M^{-1} Rx)}{(Rx, Rx)} \frac{(Rx, Rx)}{(x, x)}
\]

\[
\leq \lambda_{\max}(M) + \lambda_{\max}(M^{-1}) ||R^T R||_2
\]

\[
= \lambda_{\max}(M) + \rho(R)^2/\lambda_{\min}(M).
\]

Hence

\[
\lambda_{\min}(\frac{A^{-1} + A^{-T}}{2}) \geq \frac{1}{\lambda_{\max}(M) + \rho(R)^2/\lambda_{\min}(M)}. 
\]

Q.E.D.

The following result proves that Orthomin\((k)\) converges and provides another error bound for GCR, GCR\((k)\), and MR.

**Theorem 4.4.** If \( \{r_i\} \) is the sequence of residuals generated by GCR, Orthomin\((k)\), GCR\((k)\), or MR, then
(4.5a) \[ \| r_i \|_2 \leq \left[ 1 - \frac{\lambda_{\text{min}}(M)^2}{\lambda_{\text{max}}(A^T A)} \right]^{i/2} \| r_0 \|_2, \]

and

(4.5b) \[ \| r_i \|_2 \leq \left[ 1 - \frac{\lambda_{\text{min}}(M)^2}{\lambda_{\text{min}}(M) \lambda_{\text{max}}(M) + \rho(R)^2} \right]^{i/2} \| r_0 \|_2. \]

**Proof.** By (2.2f),

\[
\| r_{i+1} \|_2^2 = (r_i, r_i) - 2a_i (r_i, Ap_i) + a_i^2 (Ap_i, Ap_i) \\
= \| r_i \|_2^2 - 2 \frac{(r_i, Ap_i)^2}{(Ap_i, Ap_i)} + \frac{(r_i, Ap_i)^2}{(Ap_i, Ap_i)} \\
= \| r_i \|_2^2 - \frac{(r_i, Ap_i)^2}{(Ap_i, Ap_i)}.
\]

Therefore,

\[
\frac{\| r_{i+1} \|_2^2}{\| r_i \|_2^2} = 1 - \frac{(r_i, Ap_i)}{(r_i, r_i)} \frac{(r_i, Ap_i)}{(Ap_i, Ap_i)} \\
\leq 1 - \frac{(r_i, Ar_i)}{(r_i, r_i)} \frac{(r_i, Ar_i)}{(Ar_i, Ar_i)},
\]

by (3.1c)/(4.1c) and (4.2). But

\[
\frac{(r_i, Ar_i)}{(r_i, r_i)} \leq \lambda_{\text{min}}(M),
\]

and

\[
\frac{(r_i, Ar_i)}{(Ar_i, Ar_i)} = \frac{(r_i, r_i)}{(r_i, A^T A r_i)} \frac{(r_i, Ar_i)}{(r_i, r_i)} \leq \frac{\lambda_{\text{min}}(M)}{\lambda_{\text{max}}(A^T A)}.
\]

so that

\[
\| r_{i+1} \|_2 \leq \left[ 1 - \frac{\lambda_{\text{min}}(M)^2}{\lambda_{\text{max}}(A^T A)} \right]^{(i+1)/2} \| r_i \|_2.
\]
which proves (4.5a). By (4.3),

$$\frac{(r_i, Ar_i)}{(Ar_i, Ar_i)} \preceq \frac{\lambda_{\text{min}}(M)}{\lambda_{\text{min}}(M) \lambda_{\text{max}}(M) + \rho(R)^2} ,$$

so that

$$\|r_{i+1}\|_2 \leq \left[1 - \frac{\lambda_{\text{min}}(M)^2}{\lambda_{\text{min}}(M) \lambda_{\text{max}}(M) + \rho(R)^2}\right]^{i/2} \|r_i\|_2 ,$$

which proves (4.5b).

Q.E.D.

In general, the two error bounds given in Theorem 4.4 are not comparable. They are equal when $M = I$ and (4.5b) is stronger when $R = 0$. When $R = 0$, the constant $\left[\frac{\lambda_{\text{max}}(A) - \lambda_{\text{min}}(A)}{2} \right]^{1/2}$ in (4.5b) resembles the constant $\left[\frac{\lambda_{\text{max}}(A) - \lambda_{\text{min}}(A)}{\lambda_{\text{max}}(A) + \lambda_{\text{min}}(A)} \right]^{1/2}$ in the error bound for the steepest descent method (see [10]). Thus, we believe that the bounds in Theorem 4.4 are not strict for $k \geq 1$.

If $A = I - R$ with $R$ skew-symmetric, then Orthomin(1) is equivalent to GCR, and we can improve the error bounds of Theorem 3.3 and Theorem 4.4.

**Theorem 4.5.** If $A = I - R$ with $R$ skew-symmetric, then Orthomin(1) is equivalent to GCR. The residuals $\{r_i\}$ generated by Orthomin(1) satisfy

$$\|r_t\|_2 \leq \frac{2 \rho(R)^t (1 + \sqrt{1 + \rho(R)^t})^t}{(1 + \sqrt{1 + \rho(R)^t})^{2t} + \rho(R)^{2t}} \|r_0\|_2 ,$$

for even $t$. 

Proof. To prove that Orthomin(1) is equivalent to GCR, it suffices to show that $b_j^{(i)} = 0$ in (2.5b) for $j \leq i-1$. But the numerator is

$$(Ar_{i+1}, Ap_j) = (r_{i+1}, Ap_j) - (Rr_{i+1}, Ap_j)$$

By (3.1b),

$$(r_{i+1}, Ap_j) = - (r_{i+1}, Ap_j) = 0$$

Hence, by the skew-symmetry of $R$,

$$(Ar_{i+1}, Ap_j) = - (r_{i+1}, Ap_j) + (r_{i+1}, RAp_j) = - (r_{i+1}, A^2 p_j)$$

But by (2.2f),

$$(r_{i+1}, A^2 p_j) = \frac{1}{a_j} (r_{i+1}, A(r_j - r_{j+1})) = 0$$

for $j \leq i-1$, by (3.1d).

For (4.6), observe that $A = I - R$ is a normal matrix, so that (3.5) holds. We prove (4.6) by bounding $M_t$. Widlund [18] has shown that

$$(4.7) \quad M_t \leq \left[ \cosh(t \log(\frac{1}{p(R)} (1 + \sqrt{1+p(R)^2}))) \right]^{-1}$$

for even $t$. Let $\eta = \frac{1}{p(R)} (1 + \sqrt{1+p(R)^2})$. Using

$$\cosh(z) = \frac{1}{2} (e^z + e^{-z})$$

(4.7) reduces to

$$M_t \leq \frac{2}{\eta^t + \eta^{-t}} = 2 \frac{\eta^t}{\eta^{2t} + 1}$$

from which (4.6) follows.

Q.E.D.
5. Other Approaches

In this section, we discuss several methods that are mathematically equivalent to GCR.

We derived GCR from CR by replacing the short recurrence for direction vectors (2.3) with (2.5), which produces a set of $A^T A$-orthogonal vectors when $A$ is nonsymmetric. Young and Jea [19] present an alternative, Lanczos-like method for computing $A^T A$-orthogonal direction vectors:

\[(5.1a)\quad p_{i+1}^* = A p_i^* + \sum_{j=0}^{i} b_j^{(i)} p_j^* ,\]

where

\[(5.1b)\quad b_j^{(i)} = - \frac{(A^2 p_i^*, A p_j^*)}{(A p_j^*, A p_j^*)} , \quad j \leq i .\]

If $\{p_i\}$ is the set of direction vectors generated by GCR and $p_0^* = p_0^*$, then $p_i^* = c_i p_i$ for some scalar $c_i$ (see [19]). Hence, this procedure can be used to compute directions in place of (2.5). The resulting algorithm is equivalent to GCR, but does not require the symmetric part of $A$ to be positive-definite.

Axelsson [2] takes a somewhat different approach. Let $x_0$, $r_0$, and $p_0$ be as in (2.2). Then one iteration of Axelsson's method is given by:

\[(5.2a)\quad x_{i+1} = x_i + \sum_{j=0}^{i} a_j^{(i)} p_j ,\]

\[(5.2b)\quad r_{i+1} = f - A x_{i+1} ,\]

\[(5.2c)\quad b_i = - \frac{(A r_{i+1}, A p_i)}{(A p_i, A p_i)} ,\]

\[(5.2d)\quad p_{i+1} = r_{i+1} + b_i p_i .\]
where the scalars \(a_j^{(i)}\) are computed so that \(\|r_{i+1}\|_2\) is minimized. This requires the solution of a symmetric system of equations of order \(i+1\)

\[ B_{st}^{(i)} = g, \]

where \(B_{st} = (Ap_s, Ap_t)\) and \(g_s = (r_i, Ap_s)\). Thus, the solution update is more complicated than in GCR, but the computation of a set of linearly independent direction vectors is simpler. Although the direction vectors are not all \(A^TA\)-orthogonal, (5.2d) and the choice of \(a_j^{(i-1)}\) force

\[ \|r_i\|_2 = \min_{q_i \in P_i} \|q_i(A)r_0\|_2. \]

to be satisfied, so that this method is equivalent to GCR.

If these methods are restarted every \(k+1\) steps, then the resulting methods are equivalent to GCR\((k)\). Both methods can also be modified to produce methods analogous to Orthomin\((k)\): only the \(k\) previous vectors \(\{p_j^i\}_{j=i-k+1}^{i}\) are used in (5.1a), and only the \(k\) vectors \(\{p_j^i\}_{j=i-k+1}^{i}\) are used in (5.2a), with \(\{a_j^{(i)}\}_{j=i-k+1}^{i}\) computed to minimize \(\|r_{i+1}\|_2\). However, both these truncated methods may fail to converge in some cases. (We have encountered situations in which such failure occurs for the truncated version of (5.1); see [2] for a discussion of the truncated version of (5.2).) For this reason, we favor the formulation of GCR given in Section 2.

In discussing the methods of this paper, we have emphasized their variational property, i.e., that \(x_i\) is such that \(\|r_i\|_2\) is minimized over some subspace. Saad [14, 15] has developed a class of CG-like methods for nonsymmetric problems by restricting his attention to the properties of projection and orthogonality. Let \(\{v_j\}_{j=0}^i\) and \(\{w_j\}_{j=0}^i\) be two sets of linearly independent vectors, and let \(K_i := \langle v_0, \ldots, v_i \rangle\) and
$L_i := \langle w_0, \ldots, w_i \rangle$. Saad defines an oblique projection method as one that computes an approximate solution $x_{i+1} \in x_0 + K_i$ whose residual $r_{i+1}$ is orthogonal to $L_i$. For example, GCR is such a method with $K_i = \langle p_0, \ldots, p_i \rangle$ and $L_i = \langle Ap_0, \ldots, Ap_i \rangle$.

Saad presents several oblique projection methods in [14, 15]. One of these is in some sense an alternative formulation of GCR. Let $v_0 = \frac{r_0}{||r_0||_2}$, and let $\{v_t\}_{t=1}^i$ be defined by

$$h_{t+1,t} v_{t+1} = Av_t - \sum_{j=0}^t h_{jt} v_j,$$

(5.3)

where $\{h_{jt}\}_{j=0}^t$ are chosen so that

$$(v_{t+1}, Av_j) = 0, \quad j \leq t,$$

and $h_{t+1,t}$ is chosen so that $||v_{t+1}||_2 = 1$. Let $A^{(i)}$ be the solution of the system of equations

$$H_i A^{(i)} = ||r_0||_2 (1,0,\ldots,0)^T,$$

(5.4)

where $H_i$ is the upper-Hessenberg matrix whose nonzero elements are the $h_{jt}$ defined above, and let

$$x_{i+1} = x_0 + \sum_{j=0}^i a^{(i)}_j v_j.$$

(5.5)

By construction, $x_{i+1} \in x_0 + K_i$, where $K_i := \langle v_0, \ldots, v_i \rangle = \langle v_0, Av_0, \ldots, A^i v_0 \rangle$. It can be shown that $v_{i+1}$ is proportional to $r_{i+1}$, so that $r_{i+1}$ is orthogonal to $L_i := \langle Av_0, \ldots, Av_i \rangle$. It can also be shown that $x_{i+1}$ minimizes $||r_{i+1}||_2$ over $x_0 + \langle v_0, Av_0, \ldots, A^i v_0 \rangle$, so that $x_{i+1}$ is equal to the $(i+1)$'st iterate generated by GCR.
Note that the approximate solution $x_{i+1}$ is computed only after $\{v_t\}_{t=0}^i$ have been computed, so that this method lends itself naturally to restarting. Several other heuristics can be used to cut expenses (see [14, 15]). In particular, the computation of the $\{v_t\}$ can be truncated, so that at most $k$ vectors are used to compute $v_{t+1}$:

$$h_{t+1, t} v_{t+1} = A v_t - \sum_{j=\max(0, t-k+1)}^t h_{j,t} v_j.$$  

(5.6)

This procedure can then be integrated into an algorithm with restarts every $i+1$ steps, for $i > k$. After $\{v_t\}_{t=0}^i$ have been computed by (5.6), $x_{i+1}$ is computed as in (5.4) and (5.5), and the algorithm is restarted. The effect of truncating the computation of the $\{v_t\}$ is to make $H_i$ a banded upper-Hessenberg matrix with bandwidth $k$. We do not know when this method converges.
REFERENCES


