

Conjugate Residual Methods for
Almost Symmetric Linear Systems¹

by

Juan Camilo Meza

Technical Report 86-9, April 1986

¹A Thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, Rice University.

RICE UNIVERSITY

CONJUGATE RESIDUAL METHODS
FOR ALMOST SYMMETRIC LINEAR SYSTEMS

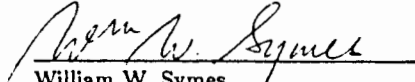
by

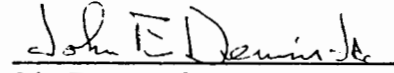
JUAN CAMILO MEZA

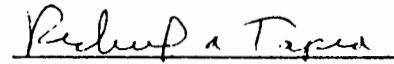
A Thesis Submitted
In Partial Fulfillment Of The
Requirements For The Degree

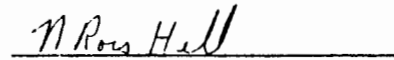
DOCTOR OF PHILOSOPHY

Approved, Thesis Committee:


William W. Symes,
Professor of Mathematical Sciences
Chairman


John E. Dennis Jr.,
Professor of Mathematical Sciences


Richard A. Tapia,
Professor of Mathematical Sciences


N. Ross Hill,
Associate Professor of Geophysics

Houston, Texas

May, 1986

CONJUGATE RESIDUAL METHODS
FOR
ALMOST SYMMETRIC LINEAR SYSTEMS

JUAN CAMILO MEZA

ABSTRACT

This study concerns the use of conjugate residual methods for the solution of nonsymmetric linear systems arising from seismic inverse problems. We focus on an application which has two distinguishing features. The first feature is that the linear system is not readily available. The second feature is that the linear system is almost symmetric. We state and prove a new convergence theorem for a class of Generalized Conjugate Residual methods which shows that in some cases the perturbed symmetric problem can be solved with an error bound similar to the one for the symmetric case.

ACKNOWLEDGEMENTS

I would like to thank my committee members John Dennis, Ross Hill, Bill Symes, and Richard Tapia for their help throughout my graduate career. I would especially like to thank my advisor Bill Symes, whose ideas and patience were invaluable. My thanks also go to Ingram Olkin who carefully read several drafts of this thesis, and whose timely advice motivated me to finish. I would also like to thank Anna Apanel, Dan Woods, Maria Rosa Celis, and Mike Pearlman who each in their own way provided help, and inspiration when the going got rough. To my parents, Carmen and Camilo, and my entire family, who were always there when I needed them, I dedicate this work to you. Finally, I would like to thank my wife Julia who put up with me through so much even in the midst of her own thesis work - we made it !

This work was supported in part by NSF grant DMS-8403148, and in part by NSF grant DCR81-16770.

TABLE OF CONTENTS

Chapter 1 Introduction	1
1.1 Statement of the Problem	1
1.2 Motivation	2
1.3 Goals	9
Chapter 2 Preliminaries	10
2.1 Notation	10
2.2 Basic Linear Algebra Theory	11
2.3 Iterative Methods	12
Chapter 3 Krylov Space Methods	16
3.1 Generalized Conjugate Residual Methods	16
3.2 Restarted and Truncated Methods	19
3.3 Convergence Results	20
Chapter 4 The Nonsymmetric Problem	24
4.1 Previous Work	24
4.2 Convergence Analysis for the Symmetric Problem	28
4.3 Perturbational Analysis for GCR	32
4.4 Special Distributions of Eigenvalues	46
Chapter 5 Numerical Results	51
5.1 Numerical Examples for Small Perturbations	51
Chapter 6 Conclusions	65
Bibliography	67

LIST OF TABLES

Chapter 3 Krylov Space Methods	
3.1. GCR work and storage requirements.	20
Chapter 4 The Nonsymmetric Problem	
4.1. Chebyshev Coefficients.	40
4.2. Error term τ_k for $\delta = 10^{-8}$	40
4.3. Error term τ_k for $\delta = 10^{-3}$	41
4.4. Error term τ_k for $\delta = 10^{-1}$	41
Chapter 5 Numerical Results	
5.1. Perturbation to Identity, N=5.	52
5.2. Perturbation to Identity, N=50.	53
5.3. Two Clusters at $\lambda_n = .1, \lambda_1 = 1.0$; N=50.	54
5.4. Two Clusters at $\lambda_n = .01, \lambda_1 = 1.0$; N=50.	55
5.5. Two Clusters at $\lambda_n = .001, \lambda_1 = 1.0$; N=50.	55
5.6. Predicted Number of Iterations for the 2 Cluster Cases.	56
5.7. Three Clusters at a=0.1, b=0.5, c=1.0; N=9.	56
5.8. Three Clusters at a=0.1, b=0.5, c=1.0; N=30.	57
5.9. Uniformly Spaced Eigenvalues $\in [1.0, 10.0]$; N=10.	58
5.10. Uniformly Spaced Eigenvalues $\in [1.0, 10.0]$; N=50.	58
5.11. Uniformly Spaced Eigenvalues $\in [1.0, 100.0]$; N=50.	59
5.12. One Isolated Large Eigenvalue at $\lambda = 100$; N=10.	61
5.13. Jordan Blocks, Small α	62
5.14. Jordan Blocks, $\alpha=1$	63
5.15. Spectral Properties of Jordan Blocks.	63

CHAPTER 1

Introduction

1.1. Statement of the Problem

This study concerns the use of conjugate residual methods for the solution of almost symmetric linear systems such as those arising from seismic inverse problems. The conjugate residual method was originally developed for symmetric positive definite systems, and is usually both efficient and effective over a wide range of problems. Many important physical problems, however, give rise to nonsymmetric linear systems (see Concus and Golub (1976), Vinsome (1976), Symes (1982)). In this study, we focus on an application arising from a seismic inverse problem which has two distinguishing features. The first feature is that the linear system is not readily available. This means that for most practical problems we must resort to an iterative procedure. The second feature is that the linear system arising from the seismic inverse problem is nearly symmetric.

Many authors have attempted to generalize the conjugate gradient methods to nonsymmetric systems. One such example is the class of Generalized Conjugate Residual (GCR) methods suggested by Eisenstat, Elman, and Schultz (1983). They prove convergence, along with a rate of convergence, for these

methods. The convergence rate derived for the GCR methods is similar to the convergence rate for steepest descent, which can be considerably slower than the rate for the conjugate gradient methods. Since the nonsymmetries in our application are small, it seems plausible that the convergence rate for the nonsymmetric conjugate gradient methods might be similar to the convergence rate for the symmetric problem. In this study, we state and prove a new convergence theorem for a class of GCR methods which shows that in some cases the perturbed symmetric problem can be solved with an error bound similar to the one for the symmetric case.

1.2. Motivation

The velocity inversion problem is a member of a class of problems known as seismic inverse problems. The idea behind the seismic inverse problem is to determine a set of parameters describing a medium, such as the earth, from another set of data usually given on the boundary of the medium. A typical example is the exploration for oil whereby small charges of explosives are set off near the ground and the resulting echoes are recorded at receivers placed near the surface at certain distances away from the explosion. The object of the seismic experiment is to determine a set of parameters that describe the structure of the earth from the data taken at the receivers.

By the velocity inversion problem we mean the problem of determining the sound speed structure of a medium from its response to an energy source.

Consider the one-dimensional velocity model:

$$\begin{aligned} \left(\frac{1}{c^2(z)} \partial_t^2 - \partial_z^2\right) u &= 0, & z > 0, \\ c(0)\partial_t u &= -f(t), & z = 0, \\ u &\equiv 0, & z > 0, t < 0. \end{aligned} \tag{1.2.1}$$

Here $c(z)$ is the wave speed, $f(t)$ is a source wavelet, and $u(z,t)$ is the wavefield. In our example $f(t)$ is the energy source, that is, the explosion. The wavefield $u(z,t)$ may be thought of as displacement or pressure. In this study we assume $f(t)$ is given and that $c(0)$ is known from measurements taken near the surface.

Define a seismogram by

$$s(t) = \left. \frac{\partial u(z,t)}{\partial t} \right|_{z=0}$$

The seismogram may be thought of as the pressure or displacement measured at the receivers after the explosive charge is set off. Notice that every quantity in the boundary value problem (1.2.1) is fixed, except for $c(z)$, so that if the wave speed is varied then the wavefield $u(z,t)$ changes. Since the seismogram depends on $u(z,t)$, it may be regarded as a function of c , that is,

$$s = F(c). \tag{1.2.2}$$

The relation (1.2.2) is known as the forward problem. By the *inverse problem* we mean the problem of determining c given a seismogram s .

As in most physical experiments, the data is known to have some noise. Under these conditions it is unlikely that we can fit the data exactly. Instead

we consider the least squares problem.

$$\min \|s - F(c)\|^2. \quad (1.2.3)$$

This is a nonlinear least squares problem. A natural choice to consider for solving this problem is some type of Newton method. For example consider the Gauss-Newton method

$$J^* J \cdot \delta c = -J^*(F(c) - s), \quad (1.2.4)$$

where $J = DF(c)$, J^* is the adjoint of J :

$$\langle J^* \cdot x, y \rangle = \langle x, J \cdot y \rangle, \quad (1.2.5)$$

and $\langle x, y \rangle$ denotes the L^2 -inner product. In order to calculate a Gauss-Newton step it is necessary to compute the actions of J and J^* on vectors. Symes (1985) shows that the action of J on a vector is given by

$$J(c) \cdot \delta c = \langle DF(c) \cdot \delta c, F(c) - s \rangle.$$

The gradient $DF(c) \cdot \delta c$ may be computed from the solution of the perturbational problem

$$\begin{aligned} \left(\frac{1}{c^2(z)} \partial_t^2 - \partial_z^2 \right) \delta u &= \frac{2\delta c}{c^3} \frac{\partial^2 u}{\partial t^2} \\ \partial_z \delta u &\equiv 0, \quad z = 0, \\ \delta u &\equiv 0, \quad t < 0, \end{aligned} \quad (1.2.6)$$

where $u(z, t)$ solves the boundary value problem (1.2.1). The gradient is then computed by

$$DF(c) \cdot \delta c = \left. \frac{\partial \delta u}{\partial t} \right|_{z=0}.$$

The adjoint is calculated by a similar process.

Two remarks are in order. The first remark is that $J \cdot \delta c$ is defined by the solution of a boundary value problem. The second remark is that each evaluation of $J \cdot \delta c$ is subject to a certain amount of discretization error. Both of these remarks also apply to the computation of the action of J^* on a vector.

Let us consider the consequences of the second remark. Assume that the boundary value problem (1.2.1) is discretized on a rectangular grid and solved by a finite-difference method. Let the matrix A denote a discretization of J , and let the matrix \tilde{A} denote a discretization of the adjoint J^* . Then we can write the discretized version of equation (1.2.4) as

$$\tilde{A} A x = \tilde{A} b. \quad (1.2.7)$$

where A is an $m \times n$ matrix, b is an m -dimensional vector, and x is an n -dimensional vector. Depending on the discretization used, both m and n can be very large. Typical values are $m=10,000$, and $n=5,000$.

Notice that equation (1.2.4) yields a symmetric positive definite system. However, neither the matrix A nor the matrix $\tilde{A} A$ is readily available, since the action of A on a vector must be computed from the solution of a boundary value problem. Moreover, given the size of a typical problem, computing either matrix by using a set of basis vectors is entirely out of the question. Therefore direct methods for the solution of the discretized version of (1.2.4) can be ruled

out.

Among the iterative methods available, the conjugate gradient algorithm proposed by Hestenes and Stiefel (1952) is a popular method for symmetric positive definite systems. This approach also has the advantage that we do not have to access the elements of the matrix A directly.

Unfortunately, the discretized equation (1.2.7) is not symmetric. When we discretize both J and J^* , we cannot hope to satisfy the adjoint relation (1.2.5) exactly for the operators A and \tilde{A} , since the discretization errors generated by the computation of A and \tilde{A} are independent. If (x, y) denotes the standard l_2 inner product, then

$$(\tilde{A}x, y) \neq (x, Ay), \quad (1.2.8)$$

that is, $\tilde{A} \neq A^T$. We can model this discretization error by the system

$$\tilde{N}x = \tilde{b}, \quad (1.2.9)$$

where

$$\begin{aligned} \tilde{N} &= (A^T + E^T)A, \\ \tilde{b} &= (A^T + E^T)b, \end{aligned} \quad (1.2.10)$$

and the matrix E can be thought of as noise generated by the calculation of $A^T x$. The matrix E is unrelated to the matrix A so that \tilde{N} is nonsymmetric. Notice that for simplicity we have chosen to model the perturbed system as if the discretization error arose from the computation of $A^T x$.

At first glance, it appears that if the discretization errors are small then the behavior of the conjugate gradient method for this problem might be similar to that for the symmetric problem. Unfortunately this is not the case. Symes (1982) has shown that even for small discretization errors, the standard conjugate gradient method applied to equation (1.2.7) may diverge. An explanation of this behavior was provided by Dennis (1984). It is well-known that the conjugate gradient method may be viewed as a minimization algorithm applied to a certain quadratic functional. The conjugate gradient method minimizes this functional by computing a search direction and taking a step along this direction. In this application the search direction depends on the vector $A^T x$. Since the calculation of $A^T x$ is contaminated by noise generated in the discretization process the search direction computed by the conjugate gradient method may not be a descent direction. Moreover, using the standard formulas for the conjugate gradient method (Hestenes and Stiefel (1952)) the steplength will be positive, so that the new iterate must increase the function value. Therefore, the sequence of iterates generated by the conjugate gradient method on this problem is not guaranteed to converge to the minimizer, and worse the iterates may diverge. This suggests that we use a nonsymmetric version of the conjugate gradient method.

Many authors have worked on the problem of generalizing the conjugate gradient method for nonsymmetric systems. However, much of this work has been in the field of elliptic equations, especially those problems arising in

reservoir engineering. Our application is different. The discretization errors can be adjusted depending on how accurately we solve the various boundary value problems. Therefore, even though the problem is nonsymmetric, it is best thought of as a small perturbation of a symmetric operator.

In Chapter 2, we define the notation used and review the basic linear algebra theory necessary in this study. Chapter 3 introduces Krylov space methods for the solution of linear systems. An example of such a method is the class of Generalized Conjugate Residual (GCR) methods, proposed by Eisenstat, Elman and Schultz (1983). Among these methods, the truncated and restarted versions of GCR are discussed. In Section 3.4 we present some of the convergence theorems for these methods proved by Eisenstat, Elman and Schultz. The nonsymmetric problem is discussed in Chapter 4. We briefly review this field and present the main result of this study in Section 4.3. We show that the GCR method converges with a bound which deviates from the error bound for the symmetric case by a term which depends on the size of the nonsymmetry. An application of the main result for the restarted version of the GCR method is also presented. Several other applications for special distributions of eigenvalues are presented in Section 4.4. In Chapter 5 we present some numerical results for test problems dealing with small perturbations to a symmetric operator. Chapter 6 contains some concluding remarks.

1.3. Goals

In this study, we investigate the behavior of conjugate residual methods for the solution of almost symmetric linear systems such as those arising from the velocity inversion problem, with particular emphasis on the following factors:

- 1) Robust modifications to conjugate residual methods in the presence of small errors.
- 2) Generalizations of the Chebyshev analysis.
- 3) A better understanding of the behavior of nonsymmetric conjugate residual methods for nearly symmetric problems.

CHAPTER 2

Notation and Preliminaries

This chapter deals with notation and preliminaries used in this study. Section 2.1 introduces the notation. In Section 2.2 we briefly review the basic linear algebra theory necessary in this study. Section 2.3 discusses iterative methods for the solution of linear systems. Matrix polynomials, which are used extensively in later chapters, are also introduced in this section.

2.1. Notation

Let x and y be real n vectors, and let A be an $n \times n$ real matrix. By (x, y) we mean the standard l_2 inner product. The l_2 norm is defined by

$$\|x\|_2 = (x, x)^{1/2}.$$

The set of eigenvalues, $\lambda(A) = \{\lambda_1(A), \dots, \lambda_n(A)\}$, of a matrix A are the n roots of the characteristic equation, $|A - \lambda I| = 0$, of A . Eigenvalues are ordered

$$|\lambda_1| \geq |\lambda_2| \geq \dots \geq |\lambda_n|.$$

The *spectral radius*, $\rho(A)$ of an $n \times n$ matrix A is defined by

$$\rho(A) = \max_{\lambda \in \lambda(A)} |\lambda_i|.$$

2.2. Basic Linear Algebra Theory

Much of the theory in this study revolves around symmetric positive definite matrices. A symmetric matrix satisfies the equation $A = A^T$. The matrix A is said to be positive definite if

$$(x, Ax) > 0 \quad \text{for all } x \neq 0.$$

For nonsymmetric matrices we define the splitting

$$A = M - R,$$

where

$$M = \frac{1}{2}(A + A^T),$$

$$R = -\frac{1}{2}(A - A^T).$$

The matrix M is called the *symmetric* part of A ; the matrix R is called the *skew-symmetric* part of A . Many of our proofs require that the symmetric part of A be positive definite.

The condition of a matrix turns out to be an important concept. By an *ill-conditioned* matrix we mean a matrix where small changes in x may cause large changes in the product Ax . For any vector norm $\|\cdot\|$, define a corresponding matrix norm by

$$\|A\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|}.$$

In particular, it can be shown (see Noble and Daniel (1977) p. 442) that for the Euclidean vector norm $\|\cdot\|_2$, the corresponding matrix norm is

$$\|A\|_2 = \sqrt{\max \lambda(A^T A)}.$$

Unless otherwise stated we will just write $\|A\|$ to denote the Euclidean matrix norm. The condition number, $\kappa(A)$, of a matrix A can now be defined by

$$\kappa(A) = \|A\| \cdot \|A^{-1}\|.$$

If the matrix is symmetric then it is straightforward to show that

$$\kappa(A) = \frac{|\lambda_1(A)|}{|\lambda_n(A)|}.$$

2.3. Iterative Methods

Consider the system of linear equations

$$Ax = b. \quad (2.2.1)$$

Techniques for solving this system of linear equations are usually classified as either direct or iterative methods. A direct method is one which guarantees a solution to equation (2.2.1) in a finite number of operations. The number of operations depends on the size of the system. If the matrix A is large then direct methods tend to take considerable time and storage. This may be reduced when the matrix A has a special structure, in which case, special direct methods may take advantage of the particular structure (see Duff (1977)). Regardless of the size or structure however, direct methods always assume that the coefficients

of the matrix A are available. This is not the case in our application.

In our application, the entries of the matrix A are not readily available. However, we can compute the action of the matrix A on a vector by solving a boundary value problem. This leads us into the area of iterative methods. By an iterative method we mean any method which generates a sequence of approximations to the solution of equation (2.2.1). Iterative methods have the advantage that they do not require that the matrix A be stored. The disadvantage is that they may converge slowly or may even diverge for some applications. In particular, we are interested in *polynomial-based* iterative methods. These methods generate a sequence of iterates of the form

$$x_k = x_0 + P_k(A)(x - x_0), \quad (2.2.2)$$

where $P_k(A)$ is a polynomial in the matrix A of degree at most k . If we denote the residual r_k by

$$r_k = b - Ax_k, \quad (2.2.3)$$

then equation (2.2.2) is equivalent to

$$r_k = Q_k(A)r_0. \quad (2.2.4)$$

Here $Q_k(A)$ is a polynomial in the matrix A of degree at most k , such that $Q_k(0) = 1$.

One important and useful fact about matrix polynomials is their behavior under orthogonal transformations. For any matrix polynomial $P_k(A)$, and any orthogonal matrix Q , if $A = Q^T T Q$, then

$$P_k(A) = Q^T P_k(T) Q. \quad (2.2.5)$$

If the matrix A is symmetric then it may be diagonalized by an orthogonal matrix so that $T = \text{diag}(\lambda_1, \dots, \lambda_n)$. As a consequence equation (2.2.5) simplifies to

$$P_k(A) = Q^T \text{diag}(P_k(\lambda_1), \dots, P_k(\lambda_n)) Q.$$

In other words the matrix polynomial in A is reduced to a polynomial in the real variables λ .

Iterative methods require a stopping rule. Usually a measure is defined in terms of how close the approximation is to the solution; the method terminates when this measure is small. We discuss two measures commonly used in the literature.

For a symmetric and positive definite matrix A , define the error functional

$$E_1(x_k) = (x - x_k, A(x - x_k))^q \equiv \|x - x_k\|_A^q,$$

where x is the solution to the linear system (2.2.1). Although this appears to be a reasonable measure of the error, it suffers from two deficiencies. The first is that, in general, we do not know what the solution x is. The second is that the A -norm is only valid when the matrix A is positive definite. However, E_1 will be used in some of our convergence analyses.

A second measure is based on the error functional

$$E_2(x_k) = (A(x - x_k), A(x - x_k))^q = \|b - Ax_k\|_2^q.$$

This error functional is more practical for many iterative procedures since the residual is already computed. This measure is also used in our convergence analysis.

CHAPTER 3

Krylov Space Methods

This chapter introduces Krylov space methods for the solution of linear systems. Section 3.1 defines a Krylov space method. In Section 3.2, we discuss the Generalized Conjugate Residual (GCR) algorithm and present some of its basic properties. Section 3.3 discusses two modifications to the GCR algorithm called *truncated* and *restarted* methods. In the last section of this chapter, we review some of the convergence results for the GCR methods.

3.1. Generalized Conjugate Residual Methods

Consider the system of linear equations

$$Ax = b, \quad (3.1.1)$$

where x and b are n dimensional vectors and A is an $n \times n$ real matrix. If the matrix A is large and sparse then this system is often solved by iterative procedures. In this chapter we present several methods proposed by Eisenstat, Elman, and Schultz (1983), which are in the class of Krylov space methods.

By a Krylov space we mean the vector space defined by

$$\kappa(v, A, k) = \text{span}\{v, Av, \dots, A^{k-1}v\}.$$

A Krylov space method is an iterative method that approximates the solution to equation (3.1.1) by generating iterates of the form

$$x_k \in x_0 + \kappa(r_0, A, k),$$

where x_0 is an initial point, and r_0 is its corresponding residual.

There are many examples of Krylov space methods in the literature (for a survey see Saad (1985)). We concentrate on a particular class of methods, namely the Generalized Conjugate Residual (GCR) methods. In the following discussion, we now assume that the symmetric part of A is positive definite.

Eisenstat, Elman, and Schultz (1983) suggest the following class of descent algorithms for the solution of equation (3.1.1).

ALGORITHM 3.1. Generalized Conjugate Residual Method

Choose x_0

Compute $r_0 = b - Ax_0$

Set $p_0 = r_0$

For $i=0, 1, \dots$

$$a_i = \frac{(r_i, Ap_i)}{(Ap_i, Ap_i)}$$

$$x_{i+1} = x_i + a_i p_i$$

$$r_{i+1} = r_i - a_i Ap_i$$

Compute p_{i+1}

The particular choice of a_i is one that minimizes $\|r_{i+1}\|_2$ as a function of a so

that $\|r_i\|_2$ decreases at each iteration.

There are different versions of this algorithm; these vary in how the new direction, p_{i+1} , is computed. If we impose the condition that

$$(Ap_i, Ap_j) = 0 \quad \text{for } i \neq j, \quad (3.1.2)$$

then at each iteration x_{i+1} minimizes the residual over the affine space $x_0 + \langle p_0, \dots, p_i \rangle$. Any set of vectors which satisfy condition (3.1.2) are said to be *A^TA-conjugate*. Condition (3.1.2) leads to the following formulas:

$$\begin{aligned} p_{i+1} &= r_{i+1} + \sum_{j=0}^i b_j^{(i)} p_j, \\ b_j^{(i)} &= \frac{-(Ar_{i+1}, Ap_j)}{(Ap_j, Ap_j)}, \quad j = 0, 1, \dots, i. \end{aligned} \quad (3.1.3)$$

The algorithm requires storage for the solution vector x , the residual r , the vector Ar , and $2(i+1)$ additional vectors for p and Ap , where i is the iteration number. The vectors Ar_{i+1} and Ap_{i+1} can share storage thereby reducing the total storage to $2(i+1) + 2$ vectors of length n . The work requirements are $[3(i+1) + 4]n$ multiplications plus 1 matrix vector multiplication per iteration. It is thus apparent that as i increases the method requires a large amount of storage and computations.

3.2. Restarted and Truncated Methods

As noted in Section 3.1, the GCR method becomes expensive as the iteration proceeds. At each iteration we must orthogonalize the new direction against every previous direction. To overcome this difficulty, we could orthogonalize the new direction against some small number of previous directions. This can be accomplished using a variety of different methods.

One alternative is to orthogonalize the current direction against the last k directions. We refer to any such method as a *truncated method*.

The formulas for the direction vectors are given by:

$$\begin{aligned} p_{i+1} &= r_{i+1} + \sum_{j=i-k+1}^i b_j^{(i)} p_j, \quad i = 0, 1, \dots \\ b_j^{(i)} &= \frac{-(Ar_{i+1}, Ap_j)}{(Ap_j, Ap_j)}, \quad j = i-k+1, \dots, i. \end{aligned} \quad (3.2.1)$$

Another method for saving storage and computing time is to restart the algorithm every $k+1$ iterations, using the current estimate for the solution as the new starting guess. Any such method is referred to as a *restarted method*.

Both of these approaches are discussed by Eisenstat, Elman, and Schultz (1983). Their version of the truncated method is also known as Orthomin(k) (see Vinsome (1976)). The restarted method is known as GCR(k). The special case for $k=0$ is known as the Minimum Residual (MR) method. Work and storage requirements for these methods are presented in Table 3.1.

Table 1. Work and Storage Requirements for GCR methods.

	GCR	Orthomin(k)	GCR(k)	MR
Work/Iter	$(3(i+1)+4)n$ + 1 Mv	$(3k+4)n$ + 1 Mv	$((3/2)k+4)n$ + 1 Mv	$4n$ + 1 Mv
Storage	$(2(i+2) + 2)n$	$(2k+3)n$	$(2k+3)n$	$3n$

Mv = Matrix-vector multiply.

3.3. Convergence Results

The basic properties of the GCR method are given by Eisenstat, Elman, and Schultz (1983). Since the direction vectors are chosen to be $A^T A$ -conjugate, a direct argument shows that x_{i+1} minimizes $\|r_{i+1}\|_2$ over Krylov spaces of increasing dimension. Eventually x_{i+1} minimizes the norm of the residual over the whole space. This can be summarized by the following theorem proved by Eisenstat, Elman, and Schultz (1983, Corollary 3.2).

THEOREM 3.1. Let A be an $n \times n$ real matrix such that $M = (A + A^T)/2$ is positive definite. Then the GCR method gives the exact solution to the system $Ax = b$ in at most n iterations.

Although Theorem 3.1 tells us that the GCR method converges in at most n iterations it does not provide information as to the rate of convergence of the method. The convergence rate is given by the following theorem also proved by Eisenstat, Elman, and Schultz (1983, Theorem 3.3).

THEOREM 3.2. If A is an $n \times n$ real matrix such that $M = (A + A^T)/2$ is positive definite, and if $\{r_i\}$ is the sequence of residuals generated by GCR, then

$$\|r_i\|_2 \leq \min_{q_i \in P_i} \|q_i(A)\|_2 \cdot \|r_0\|_2,$$

where P_i is the class of i -th degree polynomials. Moreover, if A has a complete set of eigenvectors, and if $J = T^{-1}AT$ is the Jordan canonical form of A , then

$$\|r_i\|_2 \leq \kappa(T) \min_{q_i \in P_i} \max_{\lambda \in \lambda(A)} |q_i(\lambda)| \cdot \|r_0\|_2.$$

Theorem 3.2 states that the GCR method is optimal among all polynomial-based iterative methods. Without any information about the structure of the eigenvalues, we cannot pick the best polynomial a priori. However, it can be shown that all of the GCR methods converge using simple properties derived from the iteration process. This convergence proof was provided by Eisenstat, Elman, and Schultz (1983, Theorem 4.4).

THEOREM 3.3. If A is an arbitrary real matrix such that $M = (A + A^T)/2$ is positive definite, and $R = (A^T - A)/2$, and if $\{r_i\}$ is the sequence of residuals generated by GCR, Orthomin(k), GCR(k), or MR then

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

and

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

Elman (1982 p. 141) points out that these bounds are probably not sharp, and his numerical experiments seem to indicate this. We note that if A is symmetric and positive definite so that $R = 0$, then the second bound resembles the steepest descent bound (see Luenberger (1973)). This is not too surprising, since the proof for these error bounds is the same for both the GCR method and the MR method. If we don't save any previous directions, that is $k=0$, then all the methods reduce to the MR algorithm which resembles the steepest descent algorithm.

We also note that Theorem 3.3 does not tell us how to choose k . Current folklore is that a value of $k=1$ or $k=2$ provides a good tradeoff between the work and storage requirements and an improved rate of convergence. However, this type of analysis is inadequate for determining the effect of the number of

saved directions on the rate of convergence. We show in Chapter 4 how to derive a sharper error bound which can be used, in some cases, to determine the optimal number of directions to save.

For the special case of a symmetric operator, the algorithms take on a particularly simple form. An argument parallel to the one used by Eisenstat, Elman, and Schultz (1983 Theorem 4.5) shows the following.

THEOREM 3.4. Let A be an $n \times n$ symmetric positive definite matrix.

Then Orthomin(1) generates the same iterates as the GCR method.

In essence, Theorem 3.4 states that when the GCR method is applied to a symmetric positive definite matrix the algorithm reduces to the well-known Conjugate Residual Method. This will become important in Chapter 4 when we study the effects of small perturbations to a symmetric operator on the convergence behavior of the GCR algorithm.

CHAPTER 4

The Nonsymmetric Problem

This chapter discusses the solution of large, sparse nonsymmetric linear systems. Krylov methods, introduced in Chapter 3, are discussed in relation to the nonsymmetric problem. First, we review some previous work for nonsymmetric problems. In Section 4.2, we present the standard Chebyshev convergence analysis for the Conjugate Residual method. The main result is presented in Section 4.3. We show that the GCR(k) method converges with a bound which deviates from the error bound for the symmetric case by a term which depends on the size of the nonsymmetry, provided that the method is restarted sufficiently often. Section 4.4 treats two applications of our main result for special distributions of eigenvalues.

4.1. Previous Work

The Conjugate Gradient method is a popular method for the solution of symmetric positive definite linear systems. However, many important problems give rise to nonsymmetric linear systems, which are usually large and sparse. Therefore, it seems natural to generalize the methods used for the symmetric case to the nonsymmetric case. There are various ways to extend the Conjugate

Gradient method to nonsymmetric systems. Most of these modifications are generalizations of the Conjugate Gradient (CG) method introduced by Hestenes and Stiefel (1952), or the Conjugate Residual (CR) method developed by Stiefel (1955). These methods impose conditions on the iteration method which force certain properties of the Conjugate Gradient method to be satisfied.

Historically, the first suggestion for using the Conjugate Gradient method for general linear systems is due to Hestenes and Stiefel (1952). They suggested using the CG method on the normal equations. If the matrix A has full rank, then the normal equations will be symmetric and positive definite. Fortunately, it is not necessary to form the product $A^T A$ since this could lead to a significant loss of precision. Moreover, use of the normal equations has the disadvantage that the convergence rate for conjugate gradients depends on $\kappa(A^T A)$ instead of $\kappa(A)$. If the problem is already moderately ill-conditioned then the resulting iteration scheme could converge slowly.

The Generalized Conjugate Gradient (GCG) method developed by Concus and Golub (1976), and by Widlund (1978) was an attempt to modify the CG method to nonsymmetric systems. The GCG method uses a three term recurrence formula for the solution update where certain scalars are chosen to make the residuals of the iteration mutually orthogonal. Consider the iteration

$$x_{i+1} = x_{i-1} + \omega_{i+1}(r_i + x_i - x_{i-1}), \quad i = 0, 1, \dots \quad (4.1.1)$$

If we force the residuals of this iteration to be mutually orthogonal, then we can solve for the scalars ω_i . The formulas are given by

$$\begin{aligned} \omega_1 &= 1 \\ \omega_{i+1} &= \left[1 + \frac{(\eta_i / \eta_{i-1})}{\omega_i} \right]^{-1}, \quad i = 1, 2, \dots, \end{aligned} \quad (4.1.2)$$

where $\eta_i = (r_i, r_i)$.

Axelsson (1979) developed a generalization of the conjugate residual method that differs in the formulas for the solution update. Axelsson computes the steplengths, α_i , by solving the least squares problem:

$$\min \| B^{(i)} \alpha^{(i)} - r_i \|_2, \quad (4.1.3)$$

where

$$B^{(i)} \equiv [Ap_0, \dots, Ap_i].$$

The solution to the least squares problem (4.1.3) is equivalent to minimizing the residual at each iteration.

Young and Jea (1980) proposed a modification, Orthodir, to the CR method. The formula for the direction vectors is replaced with a more expensive calculation to try to improve convergence. In particular, they choose

$$p_{i+1} = Ap_{i+1} + \sum_{j=0}^i b_j^{(i)} p_j, \quad (4.1.4)$$

$$b_j^{(i)} = \frac{-(A^2 p_i, Ap_j)}{(Ap_j, Ap_j)}.$$

Both Axelsson's method and Orthodir, together with another method proposed by Saad (1983) called GMRES, are mathematically equivalent to GCR.

They all share the property that at each iteration the residual is minimized over a certain Krylov subspace.

Saad (1981) used the relationship between the conjugate gradient method and Lanczos (1950) method to develop a class of oblique projection methods. Arnoldi's (1951) method, which is a generalization of the Lanczos method for nonsymmetric systems, is the basis for these projection methods.

Other authors have produced methods not based on the CG method for large nonsymmetric systems. Manteuffel (1977) developed a nonsymmetric version of the Chebyshev method with an adaptive procedure for estimating eigenvalues. The main disadvantage of this method is the need for good estimates of the eigenvalues of the linear operator. These estimates are usually difficult to obtain even for the simplest problems.

Gay (1979) analyzed Broyden's (1965) method for linear systems. Although Broyden's method was originally developed for nonlinear systems, Gay showed that for a nonsingular linear system, Broyden's method converges in at most $2n$ iterations for a system of order n , and proved that there exist systems for which $2n$ iterations are required.

We concentrate on the GCR methods developed by Elman (1982), who showed that several versions of the GCR method converge under the assumption that the symmetric part of the matrix A is positive definite. There are two distinguishing features in our application. The first is that we do not have access to the coefficients of the matrix A , and so we cannot form A^T (see

Section 1.2). Unfortunately, in our application we require the vector $A^T x$ in all of the above algorithms. The second feature is that the size of the nonsymmetries is small. Elman's analysis predicts a rate of convergence which is too pessimistic in many cases. One would hope that the convergence behavior for our type of problem is similar to the standard CR methods applied to the symmetric problem. We show that for small perturbations to a symmetric operator that the error bound for GCR is not too different from that given by the error bound for CR on the symmetric system. Unfortunately this bound deteriorates as the number of iterations increases so that we may have to restart the algorithm to obtain an acceptable convergence rate. First we review the standard convergence analysis for the symmetric problem.

4.2. Convergence Analysis for the Symmetric Problem

The standard Chebyshev analysis for Conjugate Gradient methods is well known (see for example Chandra (1978), Cline (1976), Axelsson (1984)) and yields optimal error bounds for the algorithm. The analysis for both the CG and CR methods is the same, but since we are mainly interested in the GCR methods we only present the error bounds for the Conjugate Residual method.

Consider the system of linear equations

$$Ax = b, \quad (4.2.1)$$

where the matrix A is symmetric and positive definite. Let P_k^1 denote the class of polynomials p_k of degree k such that $p_k(0) = 1$. The following result is due

to Chandra (1974, Theorem 3.5).

THEOREM 4.1. Let A be a symmetric positive definite matrix. Then for any $k \geq 0$, the iterates of the Conjugate Residual method satisfy

$$\|r_k\|_2 \leq \min_{p_k \in P_k^1} \max_{\lambda \in \lambda(A)} |p_k(\lambda)| \|r_0\|_2.$$

Theorem 4.1 states that the conjugate residual method generates the optimal polynomial with respect to the l_2 norm of the residual. The particular error bounds for CR found in the literature are all derived by considering specific polynomials. For the general case, Engeli, Ginsburg, Rutishauser, and Stiefel (1959) suggest as a candidate polynomial, p_k , the one that minimizes the maximum value in an interval containing the spectrum, $\lambda(A)$. The solution using this criterion is given by the normalized Chebyshev polynomial

$$p_k(\lambda) = \frac{T_k\left(\frac{2\lambda - (\lambda_1 + \lambda_n)}{\lambda_1 - \lambda_n}\right)}{T_k\left(\frac{\lambda_1 + \lambda_n}{\lambda_n - \lambda_1}\right)}, \quad (4.2.2)$$

where $T_k(z) = \cos(k \arccos z)$, $-1 \leq z \leq 1$. Using the polynomial, $p_k(\lambda)$, the following well-known bound can be derived. Although the proof can be found in several places (see for example Cline (1976)) we include it for completeness.

THEOREM 4.2. Let A be a symmetric positive definite matrix. Then for any $k \geq 0$, the iterates of the Conjugate Residual method satisfy the error bound

$$\|r_k\|_2 \leq 2 \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^k \|r_0\|_2, \quad (4.2.3)$$

where $\kappa \equiv \kappa(A) = \lambda_1/\lambda_n$.

Proof. Consider the normalized Chebyshev polynomial defined by (4.2.2). Clearly $p_k(\lambda) \in P_k^1$, so that an application of Theorem 4.1 yields

$$\|r_k\|_2 \leq \max_{\lambda \in \lambda(A)} \frac{\left| T_k \left(\frac{2\lambda - (\lambda_1 + \lambda_n)}{\lambda_1 - \lambda_n} \right) \right|}{\left| T_k \left(\frac{\lambda_1 + \lambda_n}{\lambda_n - \lambda_1} \right) \right|} \|r_0\|_2. \quad (4.2.4)$$

Using a property of Chebyshev polynomials that $|T_k(z)| \leq 1$, $-1 \leq z \leq 1$, and noting that $\left| \frac{2\lambda - (\lambda_1 + \lambda_n)}{\lambda_1 - \lambda_n} \right| \leq 1$ (4.2.4) reduces to

$$\|r_k\|_2 \leq \frac{1}{\left| T_k \left(\frac{\lambda_1 + \lambda_n}{\lambda_n - \lambda_1} \right) \right|} \|r_0\|_2. \quad (4.2.5)$$

To bound the term, $T_k \left(\frac{\lambda_1 + \lambda_n}{\lambda_n - \lambda_1} \right)$, consider the k -th degree Chebyshev polynomial,

$$T_k(z) = \cos(k \arccos z). \quad (4.2.6)$$

From the definition

$$\cos(\alpha) = \frac{1}{2}(e^{i\alpha} + e^{-i\alpha}),$$

and the relation

$$e^{ik\theta} = (\cos \theta + i \sin \theta)^k,$$

equation (4.2.6) can be rewritten as

$$T_k(z) = \frac{1}{2} \left[(z + \sqrt{z^2 - 1})^k + (z - \sqrt{z^2 - 1})^k \right]. \quad (4.2.7)$$

For the value $z = (\lambda_1 + \lambda_n) / (\lambda_n - \lambda_1) = (1 + \kappa) / (1 - \kappa)$, (4.2.7) becomes

$$T_k \left(\frac{\lambda_1 + \lambda_n}{\lambda_n - \lambda_1} \right) = \frac{1}{2} \left[\left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^k + \left(\frac{\sqrt{\kappa} + 1}{\sqrt{\kappa} - 1} \right)^k \right], \quad (4.2.8)$$

where κ is the condition number of the matrix A . Combining (4.2.8) and (4.2.5) results in

$$\|r_k\|_2 \leq 2 \left[\left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^k + \left(\frac{\sqrt{\kappa} + 1}{\sqrt{\kappa} - 1} \right)^k \right]^{-1} \|r_0\|_2. \quad (4.2.9)$$

Notice that, $\kappa(A) \geq 1$, so that

$$0 \leq \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} < 1 < \frac{\sqrt{\kappa} + 1}{\sqrt{\kappa} - 1},$$

and hence

$$\left[\left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^k + \left(\frac{\sqrt{\kappa} + 1}{\sqrt{\kappa} - 1} \right)^k \right]^{-1} \leq \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^k. \quad (4.2.10)$$

The proof is completed by substituting (4.2.10) into (4.2.9). \square

The error bound (4.2.3) depends on two facts, both of which are properties of symmetric matrices: (i) the eigenvalues of the matrix A are known to be real, (ii) the matrix A is guaranteed to have a complete set of orthogonal eigenvectors and hence is unitarily diagonalizable. Neither one of these two properties is true in general for a nonsymmetric matrix and makes the analysis of Krylov space methods for nonsymmetric matrices more difficult.

In practice, the error bound (4.2.3) can be quite pessimistic for certain problems. Whereas this is the best error bound for the general case, the bound can be improved for special distributions of the eigenvalues of A . Axelsson (1984) derived an improved error bound by assuming that the eigenvalues were distributed over two well separated intervals of equal length on the positive real axis. Jennings (1977) and Stewart (1975) also obtained results for special distributions of eigenvalues. Jennings considered the effect of one isolated eigenvalue on the convergence rate. Stewart also considered the case of one isolated eigenvalue. However, he concentrated on the convergence rate of the eigenvector associated with the isolated eigenvalue.

4.3. Perturbational Analysis

The standard convergence rate analysis for Conjugate Gradient methods is based on the assumption that the system is symmetric and positive definite. We consider the system

$$A(\epsilon)x = b, \quad (4.3.1)$$

where

$$A(\epsilon) = A + \epsilon E, \quad \epsilon > 0,$$

A is an $n \times n$ symmetric positive definite matrix, and E is a general nonsymmetric matrix such that $\|E\|_2 = 1$.

Elman (1982) has shown that the GCR method generates iterates whose residuals are bounded by

$$\|r_k\|_2 \leq \max \|q_k(A(\epsilon))\| \cdot \|r_0\|_2 \quad (4.3.2)$$

for all polynomials q_k of degree k such that $q_k(0) = 1$. However without additional information on the structure of the eigenvalues of the matrix A we cannot deduce a general result from this bound. We point out that there exist matrices for which the GCR method converges in no less than n iterations, that is, the residual will not decrease substantially until the last iteration. This point may be clarified by an example. Consider the matrix J defined by

$$J = \begin{bmatrix} \lambda & \alpha & 0 & \dots & 0 \\ 0 & \lambda & \alpha & \dots & 0 \\ 0 & 0 & \lambda & \dots & \alpha \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \lambda \end{bmatrix}$$

In this case $\|q_k(J)\|$ is on the order of 1. In fact, the minimum polynomial for J is the characteristic polynomial (see Wilkinson (1965) p.41-42). Therefore the GCR method will not produce any substantial decrease until $k=n$.

In many important examples the perturbed matrix is only slightly nonsymmetric. For such systems we would like to consider the effect of small perturbations on the convergence rate of the GCR method. We show that the convergence rate is similar to the standard convergence rate for the CR method, at least for the first few iterations.

We first prove a lemma which gives us a bound on the perturbation of a matrix polynomial.

LEMMA 4.3. Let $\phi_k(D)$ be a matrix polynomial of degree k , where D is a diagonal matrix, and let E be a general nonsymmetric matrix such that $\|E\| = 1$ for some consistent matrix norm. Then

$$\|\phi_k(D+\epsilon E) - \phi_k(D)\| \leq \epsilon \sum_{j=1}^k j \cdot |c_{kj}| \cdot (\|D\| + \epsilon)^{j-1},$$

where c_{kj} are the coefficients of the polynomial ϕ_k .

Proof. By the fundamental theorem of calculus

$$\phi_k(D+\epsilon E) - \phi_k(D) = \int_0^\epsilon d\delta \frac{d\phi_k(D+\delta E)}{d\delta}. \quad (4.3.3)$$

Taking norms on both sides of (4.3.3) and using Hölder's inequality we obtain

$$\|\phi_k(D+\epsilon E) - \phi_k(D)\| \leq \sup_{0 \leq \delta \leq \epsilon} \left\| \frac{d\phi_k(D+\delta E)}{d\delta} \right\| \cdot \int_0^\epsilon d\delta. \quad (4.3.4)$$

To bound the right-hand side of (4.3.4) consider the derivative term

$$\frac{d\phi_k(D+\delta E)}{d\delta}$$

since ϕ_k is a polynomial of degree k we can write

$$\frac{d\phi_k(D+\delta E)}{d\delta} = \frac{d}{d\delta} \sum_{j=0}^k c_{kj} (D+\delta E)^j = \sum_{j=0}^k c_{kj} \frac{d(D+\delta E)^j}{d\delta}. \quad (4.3.5)$$

Using Leibniz's rule, (4.3.5) can be rewritten as

$$\frac{d\phi_k(D+\delta E)}{d\delta} = \sum_{j=0}^k c_{kj} \sum_{i=1}^j (D+\delta E)^{i-1} E (D+\delta E)^{j-i}. \quad (4.3.6)$$

Taking norms in (4.3.6) and using the triangle inequality yields

$$\begin{aligned} \left\| \frac{d\phi_k(D+\delta E)}{d\delta} \right\| &\leq \left\| \sum_{j=0}^k c_{kj} \sum_{i=1}^j (D+\delta E)^{i-1} E (D+\delta E)^{j-i} \right\| \\ &\leq \sum_{j=1}^k j \cdot |c_{kj}| \cdot \|E\| \cdot \|D+\delta E\|^{j-1} \end{aligned} \quad (4.3.7)$$

Substituting (4.3.7) into (4.3.4) results in the inequality

$$\|\phi_k(D+\epsilon E) - \phi_k(D)\| \leq \sup_{0 \leq \delta \leq \epsilon} \left[\sum_{j=1}^k j \cdot |c_{kj}| \cdot \|E\| \cdot \|D+\delta E\|^{j-1} \right] \epsilon. \quad (4.3.8)$$

Notice that

$$\begin{aligned} \sup_{0 \leq \delta \leq \epsilon} \|D+\delta E\| &\leq \sup_{0 \leq \delta \leq \epsilon} (\|D\| + |\delta| \|E\|) \\ &\leq \|D\| + \epsilon \|E\|, \end{aligned}$$

which we may substitute into inequality (4.3.8) to obtain

$$\|\phi_k(D+\epsilon E) - \phi_k(D)\| \leq \epsilon \|E\| \sum_{j=1}^k j \cdot |c_{kj}| (\|D\| + \epsilon \|E\|)^{j-1}. \quad (4.3.9)$$

The proof is completed by using the assumption $\|E\| = 1$. \square

The main result of this study is a bound for the GCR method which applies to the perturbed problem described in equation (4.3.1). We show that when the GCR method is applied to a symmetric operator which has been perturbed, then for the first few iterations the GCR method generates iterates whose residuals satisfy an error bound that is close to the well-known error bound for the symmetric case.

THEOREM 4.4. Let $A(\epsilon) = A + \epsilon E$, where A is an $n \times n$ symmetric positive definite matrix, and E is an arbitrary matrix such that $\|E\|_2 = 1$. Then the GCR method applied to the perturbed system

$$A(\epsilon)x = b$$

yields a sequence of residuals that satisfy the inequality

$$\frac{\|r_k\|}{\|r_0\|} \leq 2 \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^k + \tau_k,$$

where

$$\tau_k = \frac{2\kappa\delta}{(\sqrt{\kappa} + 1)^{k+1}} \sum_{j=1}^k j \cdot |c_{kj}| \left[\frac{\kappa(1+\delta) - 1}{\sqrt{\kappa} + 1} \right]^{j-1} (\sqrt{\kappa} - 1)^{k-j},$$

$\delta = 2\epsilon/\lambda_1$, and c_{kj} are the coefficients of the k -th degree Chebyshev polynomial.

Proof. Since the matrix A is symmetric, it is unitarily diagonalizable, so let $A \equiv D$ be diagonal. From (4.3.2)

$$\frac{\|r_k\|_2}{\|r_0\|_2} \leq \|q_k(D)\|_2 + \|q_k(D+\epsilon E) - q_k(D)\|_2. \quad (4.3.10)$$

The first term on the right hand side of (4.3.10) is exactly the standard convergence rate bound from the symmetric problem. The second term depends on the perturbation and the polynomial chosen. To bound this term we choose a particular polynomial and apply Lemma 4.3.

By analogy to the symmetric case consider the matrix polynomial

$$q_k(D+\epsilon E) = \frac{T_k \left\{ \frac{2(D+\epsilon E) - (\lambda_1 + \lambda_n)I}{\lambda_1 - \lambda_n} \right\}}{T_k \left\{ \frac{\lambda_1 + \lambda_n}{\lambda_n - \lambda_1} \right\}}. \quad (4.3.11)$$

Define

$$\begin{aligned} \hat{D} &= \frac{2D - (\lambda_1 + \lambda_n)I}{\lambda_1 - \lambda_n}, \\ \hat{\epsilon} &= \frac{2\epsilon}{\lambda_1 - \lambda_n}. \end{aligned} \quad (4.3.12)$$

Then

$$q_k(D+\epsilon E) = \frac{T_k(\hat{D} + \hat{\epsilon})}{T_k \left\{ \frac{\lambda_1 + \lambda_n}{\lambda_n - \lambda_1} \right\}}.$$

To bound the second term of (4.3.10) use Lemma 4.3 to yield

$$\begin{aligned} \|q_k(D+\epsilon E) - q_k(D)\| &\leq \frac{\|T_k(\hat{D} + \epsilon E) - T_k(\hat{D})\|}{\left|T_k\left(\frac{\lambda_1 + \lambda_n}{\lambda_n - \lambda_1}\right)\right|} \\ &\leq \frac{\epsilon \sum_{j=1}^k j \cdot |c_{kj}| \cdot (\|\hat{D}\| + \epsilon)^{j-1}}{\left|T_k\left(\frac{\lambda_1 + \lambda_n}{\lambda_n - \lambda_1}\right)\right|}. \end{aligned}$$

By (4.3.12), $\|\hat{D}\| = 1$, so that

$$\|q_k(D+\epsilon E) - q_k(D)\|_2 \leq \frac{\frac{2\epsilon}{\lambda_1 - \lambda_n} \sum_{j=1}^k j \cdot |c_{kj}| \left(1 + \frac{2\epsilon}{\lambda_1 - \lambda_n}\right)^{j-1}}{\left|T_k\left(\frac{\lambda_1 + \lambda_n}{\lambda_n - \lambda_1}\right)\right|}. \quad (4.3.13)$$

The term in the denominator is bounded (see Cline (1976)) by

$$\left|T_k\left(\frac{\lambda_1 + \lambda_n}{\lambda_n - \lambda_1}\right)\right|^{-1} \leq 2 \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}\right)^k, \quad (4.3.14)$$

Substitute (4.3.14) into (4.3.13) and let $\delta = 2\epsilon/\lambda_1$. Then (4.3.13) becomes

$$\|q_k(D+\epsilon E) - q_k(D)\|_2 \leq 2 \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}\right)^k \frac{\kappa\delta}{\kappa-1} \sum_{j=1}^k j \cdot |c_{kj}| \left(1 + \frac{\kappa\delta}{\kappa-1}\right)^{j-1}, \quad (4.3.15)$$

where the quantity δ may be thought of as a normalized error.

Define η_k by

$$\eta_k = \frac{\kappa\delta}{\kappa-1} \sum_{j=1}^k j \cdot |c_{kj}| \left(1 + \frac{\kappa\delta}{\kappa-1}\right)^{j-1}, \quad (4.3.16)$$

which is a measure of the perturbation in the Chebyshev matrix polynomial due to the normalized error δ . Then the right hand side of (4.3.15) may be

simplified to

$$\begin{aligned} 2 \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}\right)^k \eta_k &= \frac{2\kappa\delta}{(\sqrt{\kappa} + 1)^{k+1}} \sum_{j=1}^k j \cdot |c_{kj}| \left[\frac{\kappa(1+\delta) - 1}{\sqrt{\kappa} + 1}\right]^{j-1} (\sqrt{\kappa} - 1)^{k-j} \\ &= \tau_k, \end{aligned} \quad (4.3.17)$$

so that (4.3.15) becomes

$$\|q_k(D+\epsilon E) - q_k(D)\|_2 \leq \tau_k. \quad (4.3.18)$$

The proof is completed by substituting (4.3.18) into (4.3.10). \square

Remark 1. The values of the coefficients c_{kj} of the Chebyshev polynomials are easily computed (see for example Lanczos (1961) p. 455). The coefficients for the first 10 Chebyshev polynomials are provided in Table 4.1.

Remark 2. If τ_k is a slowly growing function of k then for the first few iterations we should get a convergence rate similar to the one for the CR method on the unperturbed symmetric problem. A few of the values of τ_k for various values of δ and $\kappa(A)$ are given in Tables 4.2-4.4. Here we have used the formula

$$\tau_k = 2 \left[\left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}\right)^k + \left(\frac{\sqrt{\kappa} + 1}{\sqrt{\kappa} - 1}\right)^k \right]^{-1} \eta_k, \quad (4.3.19)$$

instead of the expression (4.3.17). For $k \leq 10$ (4.3.19) provides a tighter bound than the asymptotic formula used in (4.3.17).

Table 4.1. Chebyshev Coefficients

Coefficients for $T_k(x)$												
k	j											
	0	1	2	3	4	5	6	7	8	9	10	
0	1											
1	0	1										
2	-1	0	2									
3	0	-3	0	4								
4	1	0	-8	0	8							
5	0	5	0	-20	0	16						
6	-1	0	18	0	-48	0	32					
7	0	-7	0	56	0	-112	0	64				
8	1	0	-32	0	160	0	-256	0	128			
9	0	9	0	-120	0	432	0	-576	0	256		
10	-1	0	50	0	-400	0	1120	0	-1280	0	512	

Table 4.2. Values of τ_k for Different Condition Numbers
Normalized Error = 10^{-6}

τ_k				
k	Condition Number			
	10	100	1000	10000
1	$1.11 \cdot 10^{-6}$	$1.01 \cdot 10^{-6}$	$1.00 \cdot 10^{-6}$	$1.00 \cdot 10^{-6}$
2	$4.44 \cdot 10^{-6}$	$4.04 \cdot 10^{-6}$	$4.00 \cdot 10^{-6}$	$4.00 \cdot 10^{-6}$
3	$1.67 \cdot 10^{-5}$	$1.52 \cdot 10^{-5}$	$1.50 \cdot 10^{-5}$	$1.50 \cdot 10^{-5}$
4	$5.33 \cdot 10^{-5}$	$4.85 \cdot 10^{-5}$	$4.80 \cdot 10^{-5}$	$4.80 \cdot 10^{-5}$
5	$1.61 \cdot 10^{-4}$	$1.46 \cdot 10^{-4}$	$1.45 \cdot 10^{-4}$	$1.45 \cdot 10^{-4}$
6	$4.67 \cdot 10^{-4}$	$4.24 \cdot 10^{-4}$	$4.20 \cdot 10^{-4}$	$4.20 \cdot 10^{-4}$
7	$1.31 \cdot 10^{-3}$	$1.19 \cdot 10^{-3}$	$1.18 \cdot 10^{-3}$	$1.18 \cdot 10^{-3}$
8	$3.63 \cdot 10^{-3}$	$3.33 \cdot 10^{-3}$	$3.27 \cdot 10^{-3}$	$3.26 \cdot 10^{-3}$
9	$9.85 \cdot 10^{-3}$	$8.95 \cdot 10^{-3}$	$8.87 \cdot 10^{-3}$	$8.87 \cdot 10^{-3}$
10	$2.64 \cdot 10^{-2}$	$2.40 \cdot 10^{-2}$	$2.38 \cdot 10^{-2}$	$2.38 \cdot 10^{-2}$

Table 4.3. Values of τ_k for Different Condition Numbers
Normalized Error = 10^{-3}

τ_k				
k	Condition Number			
	10	100	1000	10000
1	$1.11 \cdot 10^{-3}$	$1.01 \cdot 10^{-3}$	$1.00 \cdot 10^{-3}$	$1.00 \cdot 10^{-3}$
2	$4.45 \cdot 10^{-3}$	$4.04 \cdot 10^{-3}$	$4.01 \cdot 10^{-3}$	$4.00 \cdot 10^{-3}$
3	$1.67 \cdot 10^{-2}$	$1.52 \cdot 10^{-2}$	$1.50 \cdot 10^{-2}$	$1.50 \cdot 10^{-2}$
4	$5.35 \cdot 10^{-2}$	$4.86 \cdot 10^{-2}$	$4.82 \cdot 10^{-2}$	$4.81 \cdot 10^{-2}$
5	$1.62 \cdot 10^{-1}$	$1.47 \cdot 10^{-1}$	$1.46 \cdot 10^{-1}$	$1.45 \cdot 10^{-1}$
6	$4.69 \cdot 10^{-1}$	$4.26 \cdot 10^{-1}$	$4.22 \cdot 10^{-1}$	$4.22 \cdot 10^{-1}$
7	$1.32 \cdot 10^0$	$1.20 \cdot 10^0$	$1.19 \cdot 10^0$	$1.19 \cdot 10^0$
8	$3.65 \cdot 10^0$	$3.31 \cdot 10^0$	$3.28 \cdot 10^0$	$3.28 \cdot 10^0$
9	$9.91 \cdot 10^0$	$9.01 \cdot 10^0$	$8.93 \cdot 10^0$	$8.92 \cdot 10^0$
10	$2.66 \cdot 10^1$	$2.42 \cdot 10^1$	$2.40 \cdot 10^1$	$2.39 \cdot 10^1$

Table 4.4. Values of τ_k for Different Condition Numbers
Normalized Error = 10^{-1}

τ_k				
k	Condition Number			
	10	100	1000	10000
1	$1.11 \cdot 10^{-1}$	$1.01 \cdot 10^{-1}$	$1.00 \cdot 10^{-1}$	$1.00 \cdot 10^{-1}$
2	$4.94 \cdot 10^{-1}$	$4.45 \cdot 10^{-1}$	$4.40 \cdot 10^{-1}$	$4.40 \cdot 10^{-1}$
3	$1.98 \cdot 10^0$	$1.77 \cdot 10^0$	$1.75 \cdot 10^0$	$1.75 \cdot 10^0$
4	$6.85 \cdot 10^0$	$6.09 \cdot 10^0$	$6.03 \cdot 10^0$	$6.02 \cdot 10^0$
5	$2.23 \cdot 10^1$	$1.97 \cdot 10^1$	$1.95 \cdot 10^1$	$1.95 \cdot 10^1$
6	$6.98 \cdot 10^1$	$6.13 \cdot 10^1$	$6.05 \cdot 10^1$	$6.04 \cdot 10^1$
7	$2.12 \cdot 10^2$	$1.85 \cdot 10^2$	$1.83 \cdot 10^2$	$1.82 \cdot 10^2$
8	$6.32 \cdot 10^2$	$5.47 \cdot 10^2$	$5.40 \cdot 10^2$	$5.39 \cdot 10^2$
9	$1.85 \cdot 10^3$	$1.59 \cdot 10^3$	$1.57 \cdot 10^3$	$1.57 \cdot 10^3$
10	$5.37 \cdot 10^3$	$4.58 \cdot 10^3$	$4.52 \cdot 10^3$	$4.51 \cdot 10^3$

An interesting point evident from Tables 4.2-4.4 is that τ_k approaches a limit as $\kappa(A) \rightarrow \infty$. Using equation (4.3.10) and taking the limit as $\kappa(A) \rightarrow \infty$ yields

$$\lim_{\kappa \rightarrow \infty} \tau_k = \delta \sum_{j=1}^k j \cdot |c_{kj}| (1 + \delta)^{j-1}.$$

Unfortunately, the case where the condition number of A is large is not of interest in our application (nor in any practical problem since the CR method would probably converge too slowly).

An immediate consequence of Theorem 4.4 is the special case of a small perturbation to the identity matrix.

COROLLARY 4.5. The GCR method applied to the perturbed system

$$(I + \epsilon E)x = b, \quad \|E\| = 1,$$

yields a sequence of residuals that satisfy the inequalities

$$\frac{\|r_k\|}{\|r_0\|} \leq k\epsilon^k.$$

Proof. An application of Theorem 4.4 shows that

$$\frac{\|r_k\|}{\|r_0\|} \leq \tau_k. \quad (4.3.20)$$

Note that $\kappa(I) = 1$, so that (4.3.20) is reduced to

$$\frac{\|r_k\|}{\|r_0\|} \leq \frac{2\delta}{2^{k+1}} k |c_{kk}| \left(\frac{\delta}{2}\right)^{k-1}. \quad (4.3.21)$$

The coefficient c_{kk} in (4.3.21), which is the leading term of the k -th Chebyshev polynomial $T_k(x)$, is

$$c_{kk} = 2^{k-1}. \quad (4.3.22)$$

Substituting (4.3.22) into (4.3.21) and using the definition of $\delta = 2\epsilon/\lambda_1$ completes the proof. \square

As we already argued in Chapter 3, the GCR method is really not a practical algorithm for the types of problems we are interested in. Setting aside the issue of storage for the moment, the GCR method is not a practical algorithm because of the large amount of computation needed as the iteration proceeds. Most of this work is in computing the inner products necessary to compute the scalars $b_j^{(i)}$, which are used in the calculation of the new direction. In some applications, for example in elliptic partial differential equations, the matrix-vector multiply is not too expensive compared to an inner product, so that as the iteration proceeds it becomes expensive to calculate a new direction. In our application a matrix-vector multiply is defined by the solution of a boundary value problem, so that the inner products are cheap compared to the matrix-vector multiply. Therefore the question of practicality will depend on the specific application. For the most part, we have also a limited amount of storage so that we are forced to use one of the truncated or restarted methods.

The restarted version of the GCR method, GCR(k), is particularly easy to analyze with the aid of Theorem 4.4.

Recall that the GCR(k) method is the GCR method restarted every $k+1$ iterations. By a *cycle* we mean any set of residuals generated between any two restarts. For example the j -th cycle is

$$\{r_{j(1)}, r_{j(2)}, \dots, r_{j(k+1)}\}.$$

Denote the sequence of residuals generated by the GCR(k) method by

$$\{r_{\alpha(0)}, r_{\alpha(1)}, \dots, r_{\alpha(k+1)}, r_{1(1)}, \dots, r_{j(l)}, \dots\}.$$

Notice that

$$r_{j(0)} = r_{j-1(k+1)} \quad j = 1, 2, \dots \quad (4.3.23)$$

Let

$$B_i = 2 \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^i + \tau_i,$$

where τ_i is the error term from Theorem 4.4, and κ is the condition number of

A.

We are now in a position to prove an error bound for the GCR(k) method.

THEOREM 4.6. The GCR(k) method applied to the perturbed system (4.3.1) generates residuals that satisfy the inequalities

$$\frac{\|r_{j(l)}\|}{\|r_{\alpha(0)}\|} \leq B_l B_{k+1}^j, \quad l = 1, \dots, k+1.$$

Proof. By Theorem 4.4 it follows that within any cycle

$$\frac{\|r_{j(l)}\|}{\|r_{j(0)}\|} \leq B_l. \quad (4.3.24)$$

Now consider the total reduction in the residual

$$\frac{\|r_{j(l)}\|}{\|r_{\alpha(0)}\|} = \frac{\|r_{j(l)}\|}{\|r_{j(0)}\|} \cdot \frac{\|r_{j(0)}\|}{\|r_{\alpha(0)}\|}. \quad (4.3.25)$$

Using (4.3.23), equation (4.3.25) reduces to

$$\frac{\|r_{j(l)}\|}{\|r_{\alpha(0)}\|} = \frac{\|r_{j(l)}\|}{\|r_{j(0)}\|} \cdot \frac{\|r_{j-1(k+1)}\|}{\|r_{\alpha(0)}\|}.$$

Repeated application of this procedure yields

$$\frac{\|r_{j(l)}\|}{\|r_{\alpha(0)}\|} = \frac{\|r_{j(l)}\|}{\|r_{j(0)}\|} \cdot \frac{\|r_{j-1(k+1)}\|}{\|r_{j-1(0)}\|} \cdots \frac{\|r_{\alpha(k+1)}\|}{\|r_{\alpha(0)}\|}.$$

An application of Theorem 4.4 to each term

$$\frac{\|r_{j(l)}\|}{\|r_{\alpha(0)}\|} = B_l \cdot B_{k+1}^j \quad l=1, \dots, k+1$$

completes the proof. \square

4.4. Special Distributions of Eigenvalues

In this section we discuss two applications of Theorem 4.4 for matrices with special distributions of eigenvalues. As in the symmetric case, the error bounds derived for the GCR method depend on the particular polynomial chosen in Theorem 4.4. The first case we consider is a matrix with eigenvalues that lie in one of two clusters. The second case is that of a matrix with one isolated large eigenvalue. In both of these cases the theory for the symmetric problem predicts an error bound which is superior to the error bound predicted for the general case (see Axelsson (1984)). The idea in both cases is to choose a polynomial, $p_k(\lambda)$, with $p_k(0) = 1$, that takes into account the special structure of the spectrum. Using this polynomial, Lemma 4.3 is applied to derive a bound for the maximum of the matrix polynomial over the spectrum of A . This bound is then used in place of the standard error bound used for the general case in Theorem 4.4.

Consider the case where the eigenvalues of the matrix A are separated into two distinct clusters of equal width. Let

$$\lambda(A) \in [\lambda_n, b] \cup [c, \lambda_1],$$

where $b - \lambda_n = \lambda_1 - c$, and define the polynomial

$$P_2(\lambda) = 1 - \omega \lambda (\lambda_1 + \lambda_n - \lambda). \quad (4.4.1)$$

If we add the additional constraint that $P_2(c) = -P_2(\lambda_1)$, then we can solve for ω so that

$$\tilde{P}_2(\lambda) = 1 - 2 [\lambda_1(c + \lambda_n) - c(c - \lambda_n)]^{-1} \lambda (\lambda_1 + \lambda_n - \lambda). \quad (4.4.2)$$

By analogy to the symmetric case consider the polynomial defined by

$$P_2(\lambda) = \frac{T_1 \left\{ \frac{2(1 - \tilde{P}_2(\lambda)) - (\beta + \alpha)}{\beta - \alpha} \right\}}{T_1 \left\{ \frac{\beta + \alpha}{\alpha - \beta} \right\}}, \quad (4.4.3)$$

where $\alpha = 1 - \tilde{P}_2(\lambda_n)$, and $\beta = 1 - \tilde{P}_2(b)$. Notice that $P_2(0) = 1$.

The Chebyshev polynomial $T_1(z) = z$, so that (4.4.3) reduces to

$$P_2(\lambda) = 1 - \frac{2}{\beta + \alpha} (1 - \tilde{P}_2(\lambda)). \quad (4.4.4)$$

Substituting (4.4.1) into (4.4.4) results in

$$P_2(\lambda) = 1 - \frac{2\omega(\lambda_1 + \lambda_n)}{(\beta + \alpha)} \lambda + \frac{2\omega}{(\beta + \alpha)} \lambda^2. \quad (4.4.5)$$

Applying Lemma 4.3 to the matrix polynomial $P_2(D + \epsilon E)$, we obtain the inequality

$$\|P_2(D + \epsilon E) - P_2(D)\| \leq \epsilon [|c_{21}| + 2 |c_{22}| (\|D\| + \epsilon)], \quad (4.4.6)$$

where c_{kj} are the coefficients of the polynomial $P_2(\lambda)$. A straightforward calculation reduces inequality (4.4.6) to

$$\|P_2(D + \epsilon E) - P_2(D)\| \leq \frac{\delta}{\beta + \alpha} \left[\frac{(3 + \delta)\kappa(A) + 1}{1 + \left(\frac{1}{\lambda_1} + \frac{1}{\lambda_n}\right)c - \frac{1}{\lambda_1\lambda_n}c^2} \right], \quad (4.4.7)$$

where $\delta = 2\epsilon/\lambda_1$.

If we use the GCR(1) method then after every 2 iterations the method will generate a polynomial of degree 2. Therefore, applying Theorem 4.4 with the bound (4.4.7) yields

$$\frac{\|r_2\|}{\|r_0\|} \leq \frac{1}{\left|T_1\left(\frac{\beta+\alpha}{\alpha-\beta}\right)\right|} + \frac{\delta}{\beta+\alpha} \left[\frac{(3+\delta)\kappa(A)+1}{1 + \left(\frac{1}{\lambda_1} + \frac{1}{\lambda_n}\right)c - \frac{1}{\lambda_1\lambda_n}c^2} \right].$$

Using the definition of the Chebyshev polynomial for $k=1$ results in

$$\frac{\|r_2\|}{\|r_0\|} \leq \frac{\alpha-\beta}{\beta+\alpha} + \frac{\delta}{\beta+\alpha} \left[\frac{(3+\delta)\kappa(A)+1}{1 + \left(\frac{1}{\lambda_1} + \frac{1}{\lambda_n}\right)c - \frac{1}{\lambda_1\lambda_n}c^2} \right]. \quad (4.4.8)$$

Remark. In the special case when $b = \lambda_n, c = \lambda_1$, then $\alpha = \beta = 1$, and (4.4.8) reduces to

$$\frac{\|r_2\|}{\|r_0\|} \leq \frac{\delta}{4} [(3+\delta)\kappa(A)+1]. \quad (4.4.9)$$

The second case of interest is the case of an isolated large eigenvalue.

Consider the case where

$$\lambda \in [\lambda_n, b] \cup [\lambda_1].$$

The standard error bound involves the condition number of A defined by λ_1/λ_n . If $\lambda_1 \gg b$, then this error bound may be a severe overestimate. Cline (1976) has shown that the effect of one isolated large eigenvalue is that of adding one iteration to a problem with a condition number of $\kappa'(A) = b/\lambda_n$. We show that the perturbed problem behaves similarly.

Consider the polynomial

$$P_k(\lambda) = \left(1 - \frac{\lambda}{\lambda_1}\right) \frac{T_{k-1}\left\{\frac{2\lambda - (b + \lambda_n)}{b - \lambda_n}\right\}}{T_{k-1}\left\{\frac{b + \lambda_n}{\lambda_n - b}\right\}}. \quad (4.4.10)$$

We seek to bound the term $P_k(D + \epsilon E)$, so as in the proof of Theorem 4.4 first write

$$P_k(D + \epsilon E) = P_k(D) + [P_k(D + \epsilon E) - P_k(D)]. \quad (4.4.11)$$

The first term can be bounded by using the standard convergence analysis of Section 4.2. To obtain a bound for the second term of (4.4.11) substitute (4.4.10) and rearrange terms to yield

$$\begin{aligned} & P_k(D + \epsilon E) - P_k(D) \\ &= \left(1 - \frac{D}{\lambda_1}\right) \left[\frac{T_{k-1}(\hat{D} + \hat{\epsilon}E) - T_{k-1}(\hat{D})}{T_{k-1}\left\{\frac{b + \lambda_n}{\lambda_n - b}\right\}} \right] - \frac{\epsilon E}{\lambda_1} \frac{T_{k-1}(\hat{D} + \hat{\epsilon}E)}{T_{k-1}\left\{\frac{b + \lambda_n}{\lambda_n - b}\right\}}, \end{aligned} \quad (4.4.12)$$

where \hat{D} and $\hat{\epsilon}$ are defined in Section 4.3. Taking norms on both sides of (4.4.12) and noting that $\|1 - \frac{D}{\lambda_1}\| \leq 1$, we obtain from (4.4.12)

$$\begin{aligned} & \|P_k(D + \epsilon E) - P_k(D)\| \\ &\leq \frac{\|T_{k-1}(\hat{D} + \hat{\epsilon}E) - T_{k-1}(\hat{D})\|}{\left|T_{k-1}\left\{\frac{b + \lambda_n}{\lambda_n - b}\right\}\right|} + \frac{\delta}{2} \frac{\|T_{k-1}(\hat{D} + \hat{\epsilon}E)\|}{\left|T_{k-1}\left\{\frac{b + \lambda_n}{\lambda_n - b}\right\}\right|}. \end{aligned} \quad (4.4.13)$$

Here again define $\delta = 2\epsilon/\lambda_1$, and let

$$\eta_{k-1} = \|T_{k-1}(\hat{D} + iE) - T_{k-1}(\hat{D})\|.$$

Then using the triangle inequality and properties of the Chebyshev polynomials

$$\begin{aligned} \|T_{k-1}(\hat{D} + iE)\| &\leq \|T_{k-1}(\hat{D})\| + \eta_{k-1} \\ &\leq 1 + \eta_{k-1}. \end{aligned} \quad (4.4.14)$$

Substituting (4.4.14) into (4.4.13) yields the bound

$$\|P_k(D + iE) - P_k(D)\| \leq \frac{1}{\left|T_{k-1}\left(\frac{b + \lambda_n}{\lambda_n - b}\right)\right|} \left[\eta_{k-1} + \frac{\delta}{2}(1 + \eta_{k-1})\right]. \quad (4.4.15)$$

Taking norms on both sides of equation (4.4.11) and using the triangle inequality yields

$$\|P_k(D + iE)\| \leq \frac{1}{\left|T_{k-1}\left(\frac{b + \lambda_n}{\lambda_n - b}\right)\right|} \left[(1 + \eta_{k-1}) + \frac{\delta}{2}(1 + \eta_{k-1})\right], \quad (4.4.16)$$

which reduces to

$$\|P_k(D + iE)\| \leq 2 \left\{ \frac{\sqrt{\kappa'} - 1}{\sqrt{\kappa'} + 1} \right\}^{k-1} \left[(1 + \eta_{k-1}) + \frac{\delta}{2}(1 + \eta_{k-1})\right], \quad (4.4.17)$$

where $\kappa' = b/\lambda_n$.

As in the symmetric case, the error bound depends on the *effective* condition number κ' . We also note that the effect of the isolated large eigenvalue is that of losing one iteration, if the term η_{k-1} is not too large.

CHAPTER 5

Applications and Numerical Results

This chapter presents some numerical results for certain applications of interest. The purpose of these numerical examples is to illustrate some of the important aspects of Theorem 4.4.

5.1. Numerical Examples for Small Perturbations

Theorem 4.4 states that for small nonsymmetric perturbations to symmetric operators the convergence rate for the GCR(k) method is similar to the convergence rate for the CR method applied to the symmetric system. We present several small numerical examples that illustrate this point.

These experiments were run on a Pyramid computer, using double precision arithmetic. The method was said to converge whenever

$$\frac{\|r_k\|}{\|r_0\|} \leq 10^{-6}.$$

In these test cases, the *noise level* refers to the size of ϵ in the equation

$$A(\epsilon) = A + \epsilon E. \quad (5.1.1)$$

Since the matrix A is symmetric positive definite we assumed that it was already

diagonalized, so that $A = \text{diag}(d_1, d_2, \dots, d_n)$. The nonsymmetric perturbations were generated using the random number generator, URAND, from IMSL. The matrices, E , were computed by generating uniform random numbers between $[-0.5, +0.5]$, and normalizing so that $\|E\|_2 = 1$. The noise level was adjusted by varying ϵ .

The first test case is an application of Corollary 4.5. The matrices in this test case are all of the form $A = I + \epsilon E$. Corollary 4.5 predicts the error bound

$$\frac{\|r_k\|}{\|r_0\|} \leq k\epsilon^k.$$

Tables 5.1-5.2 display the number of iterations required for the GCR(k) method to converge for two matrices of order 10 and 50. Both of these matrices are small perturbations of the identity matrix.

Table 5.1
Perturbation to Identity, $N = 5$.

Number of Iterations			
k	Noise Level		
	10^{-6}	10^{-3}	10^{-1}
1	1	2	5
2	1	2	5
3	1	2	4
4	1	2	5
5	1	2	4

Table 5.2
Perturbation to Identity, $N = 50$.

Number of Iterations			
k	Noise Level		
	10^{-6}	10^{-3}	10^{-1}
1	1	2	5
2	1	2	5
3	1	2	5
4	1	2	5
5	1	2	5

These results show that the convergence rate is predicted quite well by the theory.

The second set of test cases was chosen to demonstrate the effect of clusters of eigenvalues on the convergence rate. For the symmetric case it is well known that the CR method will converge in at most m iterations, where m is the number of distinct eigenvalues. In fact, the clustering of eigenvalues tends to improve the convergence rate more than would be expected from the standard bound given in Section 3.3. The purpose of this set of tests is to determine if the nonsymmetries would destroy this clustering effect. A secondary goal is to determine if there is an optimal number of directions to save depending on the number of clusters. We ran several cases, with a various number of clusters of eigenvalues. Tables 5.3-5.5 demonstrate the effect of the noise level on the two cluster case. Table 5.6 displays the predicted number of iterations using the theoretical bounds derived in Section 4.4. In the two cluster cases, the matrix A

was formed so that it had two eigenvalues each with multiplicity $n/2$, that is the matrices are of the form

$$d_i = \lambda_1, \quad 1 \leq i \leq \frac{n}{2},$$

$$d_i = \lambda_n, \quad \left(\frac{n}{2} + 1\right) \leq i \leq n.$$

Tables 5.7-5.8 present the results for the 3 cluster cases. In the three cluster case, the matrix A has 3 eigenvalues each with multiplicity $n/3$. These matrices are of the form

$$d_i = \lambda_1, \quad 1 \leq i \leq \frac{n}{3},$$

$$d_i = \lambda_2, \quad \left(\frac{n}{3} + 1\right) \leq i \leq \frac{2n}{3},$$

$$d_i = \lambda_n, \quad \left(\frac{2n}{3} + 1\right) \leq i \leq n.$$

Table 5.3
Two Clusters at $\lambda_n = .1, \lambda_1 = 1.0$; $N = 50$.

Number of Iterations			
k	Noise Level		
	10^{-6}	10^{-3}	10^{-1}
0	6	7	69
1	3	6	22
2	3	5	26

Table 5.4
Two Clusters at $\lambda_n = .01, \lambda_1 = 1.0$; $N = 50$.

Number of Iterations			
k	Noise Level		
	10^{-6}	10^{-3}	10^{-1}
0	4	106	34 [†]
1	3	8	37 [†]
2	3	8	56 [†]

[†] GCR stalled out (stepsize too small).

Table 5.5
Two Clusters at $\lambda_n = .001, \lambda_1 = 1.0$; $N = 50$.

Number of Iterations			
k	Noise Level		
	10^{-6}	10^{-3}	10^{-1}
0	9	857 [†]	35 [†]
1	4	16	21 [†]
2	3	16	32 [†]

[†] GCR stalled out (stepsize too small).

Section 4.4 discussed the case where the eigenvalues were contained in two distinct clusters. For the special case where the matrix has only two eigenvalues the analysis predicts the error bound

$$\frac{\|r_2\|}{\|r_0\|} \leq \frac{\delta}{4} ((3 + \delta)\kappa(A) + 1), \quad (5.1.2)$$

where $\delta = \frac{2\epsilon}{\lambda_1}$. Using this error bound, the number of iterations required to reduce the norm of the residual by 10^{-b} may be computed. If we denote by p the

minimum number of iterations to reduce the norm of the residual by 10^{-6} , then

$$p = 2 \cdot \frac{\log(10^{-6})}{\log \left[\frac{\delta}{4} ((3 + \delta)\kappa(A) + 1) \right]}$$

These values are tabulated in Table 5.6. For certain combinations of the condition number and the normalized error the bound in equation (5.1.2) is greater than 1, so that the predicted number of iterations is meaningless; these values are not displayed. Overall though the predicted number of iterations match very well against the actual number of iterations taken.

Table 5.6
Predicted Number of Iterations for the 2 Cluster Cases.

Number of Iterations			
Condition Number	Noise Level		
	10^{-6}	10^{-3}	10^{-1}
10	4	7	...
100	4	15	...
1000	5

Table 5.7
Three Clusters at $a=0.1$, $b=0.5$, $c=1.0$; $N = 9$.

Number of Iterations			
k	Noise Level		
	10^{-6}	10^{-3}	10^{-1}
1	11	11	20
2	4	7	20
3	4	7	20
5	4	6	20

Table 5.8
Three Clusters at $a=0.1$, $b=0.5$, $c=1.0$; $N = 30$.

Number of Iterations			
k	Noise Level		
	10^{-6}	10^{-3}	10^{-1}
1	11	11	30
2	4	7	25
3	4	7	24
5	4	7	23

These numerical experiments demonstrate that the GCR(k) method applied to the perturbed problem behaves very much like CR applied to the symmetric problem. In addition these test cases point out that the convergence rate of the GCR(k) method does not improve by saving more directions than is necessary to build up a k-th degree polynomial, where k is the number of clusters. For example, the GCR(1) method builds a quadratic polynomial, so that for the two cluster case saving 1 direction is sufficient.

In the third test case, the eigenvalues of the matrix A are uniformly distributed in the interval $[\lambda_n, \lambda_1]$. This purpose of this test case is to determine the effect of the number of saved directions on the convergence rate. Tables 5.9-5.11 illustrate the effect on the convergence rate for various numbers of saved directions and condition numbers.

Table 5.0
Evenly Spaced Eigenvalues $\in [1.0, 10.0]$; $N=10$.

Number of Iterations			
k	Noise Level		
	10^{-6}	10^{-3}	10^{-1}
1	30	30	30
2	27	27	27
3	23	23	23
4	19	19	19
5	21	21	21

Number of iterations for CR on symmetric problem = 10.

Table 5.10
Evenly Spaced Eigenvalues $\in [1.0, 10.0]$; $N=50$.

Number of Iterations			
k	Noise Level		
	10^{-6}	10^{-3}	10^{-1}
1	31	31	31
2	26	26	26
3	24	24	24
4	23	23	23
5	22	22	22
10	21	21	21

Number of iterations for CR on symmetric problem = 20.

Table 5.11
Evenly Spaced Eigenvalues $\in [1.0, 100.0]$; $N=50$.

Number of Iterations			
k	Noise Level		
	10^{-6}	10^{-3}	10^{-1}
1	203	203	201
2	143	143	143
3	114	114	113
4	97	97	97
5	85	85	86
10	67	67	66

Number of iterations for CR on symmetric problem = 34.

The number of iterations required for the CR method to converge on the symmetric problem is also given for comparison. These results can also be compared with the predicted number of iterations for the symmetric case. Using the standard Chebyshev bound (see Section 4.2), the predicted number of iterations to reduce the initial norm of the residual by 10^{-6} is 23 for a matrix with a $\kappa(A) = 10$, and 73 for a matrix with a $\kappa(A) = 100$.

In these test cases the noise level does not affect the convergence behavior of the GCR(k) method. This phenomenon can be explained as follows. The unperturbed matrix A has all simple eigenvalues. In this case, the eigenvalues and eigenvectors of A are both continuous functions of the perturbation (see Wilkinson (1965)). Furthermore, the noise level is small enough that the perturbed matrix also has all simple eigenvalues. Therefore, the matrix $A(\epsilon)$ has a complete set of eigenvectors. Applying Theorem 3.2, we obtain the bound

$$\|r_i\|_2 \leq \kappa(T) \min_{q_i \in P_i} \max_{\lambda \in \lambda(A)} |q_i(\lambda)| \cdot \|r_0\|_2 \quad (5.1.3)$$

where T is the matrix whose columns are the eigenvectors of A . The matrix T is a perturbation of an orthogonal matrix since A is symmetric, which implies that T is probably well conditioned. Therefore, as long as the noise level is not larger than half the separation distance between eigenvalues, the matrix T should remain well-conditioned, which implies that the bound in (5.1.3) should not change by much.

The fourth test case investigates the effect of an isolated large eigenvalue on the convergence rate. In the case of an isolated large eigenvalue the condition number of the linear system may predict a convergence rate much larger than the one observed. As Cline (1976) has shown, the convergence rate really depends on the *effective* condition number, that is the condition number of the matrix if the large isolated eigenvalue were removed. The purpose of this test case is to find out if this property is preserved for the case of a small nonsymmetric perturbation. In this test case the matrix A has $n-1$ uniformly distributed eigenvalues in $[1,10]$, and 1 eigenvalue at 100. Thus the condition number of A is equal to 100, but the effective condition number is equal to 10. Table 5.12 shows that the isolated large eigenvalue does slow down the convergence rate, but only as expected from the analysis for the symmetric case. Comparing Tables 5.10-5.12 we see that the test case with the isolated large eigenvalue (Table 5.12) is converging at almost the same rate as the test case

with a condition number = 10 (Table 5.10). This effect is even more pronounced as the number of saved directions increases.

Table 5.12
One Isolated Large Eigenvalue at $\lambda = 100$; $N = 50$.

k	Number of Iterations		
	Noise Level		
	10^{-6}	10^{-3}	10^{-1}
1	50	50	55
2	35	35	41
3	28	24	30
4	25	27	28
5	23	24	24
10	21	21	21

The last test cases use a small variation of Jordan blocks. In these test cases the matrix A is formed by setting the diagonal elements equal to 1, and the superdiagonal elements equal to α , that is,

$$J = \begin{bmatrix} \lambda & \alpha & 0 & \dots & 0 \\ 0 & \lambda & \alpha & \dots & 0 \\ 0 & 0 & \lambda & \dots & \dots \\ \dots & \dots & \dots & \dots & \alpha \\ 0 & 0 & 0 & \dots & \lambda \end{bmatrix}$$

This is an extreme case of a nonsymmetric matrix in the sense that it has exactly 1 eigenvector regardless of the size of the matrix. Table 5.13 records the results for the test cases where $\alpha = 0.1$, and $\alpha = 0.5$. These test matrices produce results very similar to test case 1, where we had small perturbations to

the identity matrix.

Table 5.13
Number of Iterations to Converge versus α ; $N=10$.

Number of Iterations		
k	α	
	0.1	0.5
1	6	15
2	6	14
3	6	14
4	6	13
5	6	13
10	6	10

Table 5.14 displays the results for $\alpha = 1$, and various dimensions. The last row of this table displays the number of iterations necessary to converge using the standard CR algorithm, which in this case is equivalent to Orthomin(1). These results verify that taking more directions does not necessarily improve the convergence rate. Another point to notice is that the convergence rate does not improve substantially until we use the GCR(n) method.

Table 5.14
Number of Iterations to Converge for $\alpha = 1$, Jordan Blocks.

Number of Iterations				
k	N			
	5	10	20	50
1	26	41	63	119
2	29	40	67	126
3	33	58	71	133
4	5	50	76	135
5	5	54	75	146
10	-	10	81	154
CR	26	41	66	140

In Table 5.15, we tabulate various properties of the test matrices which can be used to predict the rate of convergence.

Table 5.15
Spectral Properties of Jordan Blocks.

α	N	$\ A\ $	$\kappa(A)$	$\ R\ $	δ
0.1	10	1.096	1.21	0.096	0.18
0.1	50	1.100	1.22	0.100	0.18
0.5	10	1.480	2.84	0.480	0.64
0.5	50	1.499	2.99	0.499	0.67
1.0	5	1.866	13.03	0.866	0.92
1.0	10	1.960	48.37	0.960	0.92
1.0	20	1.989	178.1	0.989	1.00
1.0	50	2.0	1054	1.000	1.00

Here δ is the normalized error given by $2\|R\| / \|A\|$.

Notice that for the cases with $\alpha = .1$, the condition number of the symmetric part of A is close to 1, which implies that the GCR(k) method

applied to this test case should behave similarly to the test cases where the GCR(k) method was applied to small perturbations of the identity matrix. For the cases with $\alpha = 0.5$, the bounds in Tables 4.2-4.6 predict that GCR(2) is probably optimal, since the error bounds B_k achieve a minimum for $k=2$. Unfortunately the cases with $\alpha = 1.0$ yield error bounds greater than 1.0, so that we cannot use Tables 4.2-4.6 to predict a convergence rate.

CHAPTER 6

Conclusions

In this study we analyze the behavior of conjugate residual methods for almost symmetric linear systems. The conjugate residual method, which is a popular method for the solution of symmetric positive definite systems is presented and shown to have a convergence rate which depends on the $\sqrt{\kappa(A)}$. We also present the GCR methods, proposed by Eisenstat, Elman, and Schultz. Their convergence rate for the GCR methods depends on the $\kappa(A)$, and is similar to the steepest descent bound. The main result of this study is a new convergence theorem for the application of the GCR methods to almost symmetric linear systems. This theorem shows that the GCR methods have a convergence rate for the perturbed problem which is a small perturbation of the convergence rate for the CR method applied to the unperturbed problem. We also give several applications for special distributions of eigenvalues, which show that the GCR methods on the perturbed problem behave similarly to the CR method on the symmetric problem. In addition, some of the analysis indicates that the clustering of the eigenvalues determines how many previous directions to save in the GCR(k) methods.

There are still some questions left unanswered. The analysis used in this study is a perturbational analysis, and like most analyses of this type it works best for small perturbations. We remark that for large perturbations the error bounds predicted by the theory are meaningless. There is still a question of whether the error bounds derived in this study can be sharpened for large perturbations. Another interesting question relates to roundoff error. Since roundoff error may be considered a small nonsymmetric perturbation to a symmetric operator, it may be possible to apply this work to develop a roundoff error analysis for the conjugate gradient methods.

BIBLIOGRAPHY

- Arnoldi, W.E. [1951]. The principle of minimized iteration in the solution of the matrix eigenvalue problem. *Quart. Appl. Math.*, 9:17-29.
- Axelsson, O. [1979]. A generalized conjugate direction method and its application on a singular perturbation problem. *Proceedings of 8th Biennial Numerical Analysis Conference held at Dundee, Scotland, June 26-29, 1979. Reproduced in Lecture Notes in Mathematics No. 778, pp. 1-12, Springer-Verlag 1980.*
- Axelsson, O. [1980]. Conjugate gradient type methods for unsymmetric and inconsistent systems of linear equations. *Linear Alg. Appl.*, 29:1-16.
- Axelsson, O. and Barker, V.A. [1984]. *Finite Element Solution of Boundary Value Problems*. Academic Press, Orlando, Florida.
- Broyden, C.G. [1965]. A class of methods for solving nonlinear simultaneous equations. *Math. Comp.*, 19:577-593.
- Chandra, R. [1978]. Conjugate Gradient Methods for Partial Differential Equations. *Doctoral Dissertation, Department of Computer Science, Yale University*. Also available as Research Report No. 129.
- Cline, A.K. [1976]. Several observations on the use of conjugate gradient methods. *ICASE Report 76-22. NASA Langley Research Center, Hampton, Virginia*.
- Concus, P. and Golub, G.H. [1976]. A generalized conjugate gradient method for nonsymmetric systems of linear equations. *Technical Report STAN-CS-76-535, Department of Computer Science, Stanford University*.
- Daniel, J.W. [1967]. The conjugate gradient method for linear and nonlinear operator equations. *SIAM J. Numer. Anal.*, 4:10-26.
- Dennis, J.E. Jr., [1984]. Private communication.

- Duff, I.S. [1977]. A survey of sparse matrix research. *Proc. IEEE*, 65:500-535.
- Eisenstat, S.C. [1982]. A note on the generalized conjugate gradient method. *Technical Report No. 228, Department of Computer Science, Yale University.*
- Eisenstat, S.C., Elman, H.C., and Schultz, M.H. [1983]. Variational iterative methods for nonsymmetric systems of linear equations. *SIAM J. Numer. Anal.*, 20:345-357.
- Elman, H.C. [1982]. Iterative Methods for Large Sparse Nonsymmetric Systems of Linear Equations. *Doctoral Dissertation, Department of Computer Science, Yale University.* Also available as Research Report No. 229.
- Engeli, M., Ginsburg, T., Rutishauser, H. and Stiefel, E. [1959]. Refined iterative methods for computation of the solution and the eigenvalues of self-adjoint boundary value problems. *Mitteilungen aus dem Institut für angewandte Mathematik, Birkhäuser Verlag, Basel, Stuttgart.*
- Fox, L. and Parker I.B. [1968]. *Chebyshev Polynomials in Numerical Analysis.* Oxford University Press, London.
- Gay, D.M. [1970]. Some convergence properties of Broyden's method. *SIAM J. Numer. Anal.*, 16:623-630.
- Greenbaum, A. [1981]. Behavior of the conjugate gradient algorithm in finite precision arithmetic. *Report UCRL 85752, Lawrence Livermore Laboratory, Livermore, California.*
- Hestenes, M.R. and Stiefel, E. [1952]. Method of conjugate gradients for solving linear systems. *J. Res. Nat. Bureau Standards*, 49:409-436.
- Jea, K.C. [1982]. Generalized Conjugate Gradient Acceleration of Iterative Methods. *Doctoral Dissertation, Department of Mathematics, University of Texas.*
- Jennings, A. [1977]. Influence of the eigenvalue spectrum on the convergence rate of the conjugate gradient method. *J. Inst. Math. Appl.*, 20:61-72.
- Kato, T. [1982]. *A Short Introduction to Perturbation Theory for Linear Operators.* Springer Verlag, New York.

- Lanczos, C. [1950]. An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *J. Res. Nat. Bureau Standards*, 45:255-282.
- Lanczos, C. [1961]. *Applied Analysis.* Prentice Hall, Englewood Cliffs, New Jersey.
- Luenberger, D.G. [1973]. *Introduction to Linear and Nonlinear Programming.* Addison-Wesley, Reading, Massachusetts.
- Manteuffel, T.A. [1977]. The Tchebyshev iteration for nonsymmetric linear systems. *Numer. Math.*, 28:307-327.
- Manteuffel, T.A. [1978]. Adaptive procedure for estimating parameters for the nonsymmetric Tchebychev iteration. *Numer. Math.*, 31:183-208.
- Paige, C.C. [1972]. Computational variants of the Lanczos method for the eigenproblem. *J. Inst. Math. Appl.*, 10:373-381.
- Paige, C.C. [1976]. Error analysis of the Lanczos algorithm for tridiagonalizing a symmetric matrix. *J. Inst. Math. Appl.*, 18:341-349.
- Paige, C.C. [1980]. Accuracy and effectiveness of the Lanczos algorithm for the symmetric eigenproblem. *Linear Alg. Appl.*, 34:235-258.
- Paige, C.C. and Saunders, M.A. [1975]. Solution of sparse indefinite systems of equations and least squares problems. *SIAM J. Numer. Anal.*, 12:617-629.
- Paige, C.C. and Saunders, M.A. [1982]. LSQR: An algorithm for sparse linear equations and sparse least squares. *ACM Trans. Math. Soft.*, 8:43-71.
- Saad, Y. [1980]. Variations on Arnoldi's method for computing eigenelements of large unsymmetric matrices. *Linear Alg. Appl.*, 34:269-295.
- Saad, Y. [1981]. Krylov subspace methods for solving large unsymmetric linear systems. *Math. Comp.*, 37:105-126.
- Saad, Y. [1982]. The Lanczos biorthogonalization algorithm and other oblique projection methods for solving large unsymmetric systems. *SIAM J. Numer. Anal.*, 19:485-506.

- Saad, Y. [1983]. GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems. *Technical Report No. 254, Department of Computer Science, Yale University.*
- Saad, Y. and Schultz, M.H. [1985]. Conjugate gradient-like algorithms for solving nonsymmetric linear systems. *Math. Comp.*, 44:417-424.
- Stewart, G.W. [1975]. The convergence of the method of conjugate gradients at isolated extreme points in the spectrum. *Numer. Math.*, 24:85-93.
- Stiefel, E. [1955]. Relaxationmethoden bester Strategie zur Lösung linearer Gleichungssysteme. *Comm. Math. Helv.*, 29:157-179.
- Symes, W.W., [1982]. Computational continuation for solutions of wave equations. *Unpublished manuscript.*
- Symes, W.W., [1985]. Stability properties for the velocity inversion problem. *To appear in Proc. of the SEG/SIAM/SPE Symposium, Houston, Texas, January 1985.*
- Vinsome, P.K.W. [1976]. ORTHOMIN - An iterative method for solving sparse sets of simultaneous linear equations. *Proc. Fourth SPE Symposium on Reservoir Simulation, Los Angeles, pp. 149-160.*
- Widlund, O. [1978]. A Lanczos method for a class of nonsymmetric systems of linear equations. *SIAM J. Numer. Anal.*, 15:801-812.
- Wilkinson, J.H. [1965]. *The Algebraic Eigenvalue Problem.* Oxford University Press, London.
- Wozniakowski, H. [1980]. Roundoff error analysis of a new class of conjugate gradient algorithms. *Lin. Alg. Appl.*, 29:507-529.
- Young, D.M. and Jea, K.C. [1980]. Generalized conjugate-gradient acceleration of non-symmetrizable iterative methods. *Linear Alg. Appl.*, 34:159-194.