POLYNOMIAL APPROXIMATIONS TO FUNCTIONS OF OPERATORS

by

Pravin Singh

Submitted in partial fulfilment of the

requirements for the degree of

Doctor of Philosophy,

in the

Department of Mathematics and Applied Mathematics,

University of Natal

1994

Durban

1994

PREFACE

The theoretical work described in this thesis was carried out in the Department of Mathematics and Applied Mathematics, University of Natal, Durban, from March 1991 to June 1994, under the supervision of Professor Janusz R. Mika.

The studies represent original work by the author and have not been submitted in any form to another University. Where use was made of the work of others it has been duly acknowledged in the text.

ACKNOWLEDGEMENTS

The author is indebted to his supervisor Professor Janusz Mika for his expert guidance. Great appreciation is expressed for the financial support from Professor Mika's FRD grant which made it possible to attend the SAMS and SANUM conferences in 1992. The UDP is acknowledged for their financial support which covered the cost of registration for this degree.

To three beautiful people:

Rookmoney, Sarishka and Shivani

ABSTRACT

To solve the linear equation $Ax = f$, where f is an element of Hilbert space H and A is a positive definite operator such that the spectrum $\sigma(A) \subset [m,M]$, we approximate the inverse operator A^{-1} by an operator V which is a polynomial in A. Using the spectral theory of bounded normal operators the problem is reduced to that of approximating a function of the real variable by polynomials of best uniform approximation. We apply two different techniques of evaluating A^{-1} so that the operator V is chosen either as a polynomial $P_n(A)$ when $P_n(\lambda)$ approximates the function $1/\lambda$ on the interval $[m,M]$ or a polynomial $Q_n(A)$ when $1 - \lambda\, Q_n(\lambda)$ approximates the function zero on $[m,M]$. The polynomials $P_n(A)$ and $Q_n(A)$ satisfy three point recurrence relations, thus the approximate solution vectors $P_n(A)f$ and $Q_n(A)f$ can be evaluated iteratively. We compare the procedures involving $P_n(A)f$ and $Q_n(A)f$ by solving matrix vector systems where A is positive definite. We also show that the technique can be applied to an operator which is not selfadjoint, but close, in the sense of operator norm, to a selfadjoint operator. The iterative techniques we develop are used to solve linear systems arising from the discretization of Freedholm integral equations of the second kind. Both smooth and weakly singular kernels are considered. We show that earlier work done on the approximation of linear functionals $< x,g >$, where $g \in H$, involve a zero order approximation to the inverse operator and are thus special cases of a general result involving an approximation of arbitrary degree to A^{-1}.

LIST OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Here we evaluate the integral using the technique by Atkinson [3] which involves an adaptation of the trapezoidal rule which is summarized below.

Firstly, the interval (a,b) is divided into N equally spaced subintervals and the integral is evaluated over each subinterval.

Hence

$$\int_a^b K(x,y)\ \phi(y)\ dy = \sum_{k=1}^{N} \int_{y_k}^{y_{k+1}} K(x,y)\ \phi(y)\ dy\ , \tag{6.11}$$

where $y_k = a + (k-1)h$, $k = 1, 2, \cdots N+1$ and $h = (b-a)/N$ is the length of each subinterval. On each subinterval (y_k, y_{k+1}) we replace $\phi(y)$ by the linear Lagrange polynomial

$$1/h\ [\ (y-y_k)\ \phi(y_{k+1}) - (y-y_{k+1})\ \phi(y_k)\ ]\ . \tag{6.12}$$

Substituting (6.12) into (6.11) and letting $x = y_i$ $(i = 1,2,\cdots N+1)$, we obtain

$$\int_a^b K(y_i,y)\ \phi(y)\ dy = \sum_{k=1}^{N+1} \omega_{ik}\ \phi(y_k)\ , \tag{6.13}$$

where the weights ω_{ik} are given by

$$\omega_{i1} = -1/h \int_{y_1}^{y_2} (y-y_2)\ K(y_i,y)\ dy$$

$$\omega_{ik} = 1/h \int_{y_{k-1}}^{y_k} (y-y_{k-1})\ K(y_i,y)\ dy - 1/h \int_{y_k}^{y_{k+1}} (y-y_{k+1})\ K(y_i,y)\ dy,\ k = 2,3\cdots N$$

$$\omega_{iN+1} = 1/h \int_{y_N}^{y_{N+1}} (y-y_N)\ K(y_i,y)\ dy.$$

INTRODUCTION

Let A be a bounded normal operator and let the inverse A^{-1} exist. Most of the methods available for solving the linear system $Ax = f$, where f is an element of Hilbert space H, avoid the calculation of A^{-1} since such calculation is expensive. We consider approximations to A^{-1} by polynomials in A since such polynomials are bounded and are easily evaluated.

In chapter 1 we review the spectral theory of functions of normal operators. If A is a bounded normal operator and $f(\lambda)$ is a function that is bounded and continuous on a domain containing the spectrum $\sigma(A)$, then the operator function $f(A)$ is related to the function $f(\lambda)$ by equations (1.6) and (1.7). If A is bounded and positive definite with $\sigma(A) \subset [m,M]$ $(0 < m < M)$ then $f(A) = A^{-1}$ is related to the function $f(\lambda) = 1/\lambda$ in the interval $[m,M]$. For non-selfadjoint operators we would have to consider functions of the complex variable. However, the approximation of such functions is numerically not very attractive. So in this thesis we consider only positive definite operators and operators that deviate little from symmetry.

For positive definite operators A we choose to approximate A^{-1} by a polynomial operator V in A. By the spectral theory discussed in chapter 1 the problem is reduced to that of the approximation of a function of the real variable by polynomials of best uniform approximation. The operator V is chosen as a polynomial $P_n(A)$ of degree n when $P_n(\lambda)$ approximates the function $1/\lambda$ in the interval $[m,M]$ or a polynomial $Q_n(A)$ when $1 - \lambda Q_n(\lambda)$ approximates the function zero in the interval $[m,M]$. These two polynomials $P_n(A)$ and $Q_n(A)$ satisfy three point recurrence relations. The derivation of these two approximations is presented in chapter 2. The theory pertaining to the polynomials $P_n(A)$ is due to Bond and Mika [1].

However Bond and Mika [1] had not fully investigated the application and usefulness of these polynomials. The theory pertaining to the polynomials $Q_n(A)$ represents original work by the author. These polynomials were derived from the Chebyshev polynomials.

In chapter 3 the polynomials $P_n(A)$ and $Q_n(A)$ are used to solve matrix vector systems of the form $Ax = f$, where f is an element of Hilbert space H and A is a positive definite operator with $\sigma(A) \subset [m,M]$. Some of the available techniques for solving such matrix vector systems is presented. We then devise a method to generate the positive definite matrix A. The approximate solution vectors $P_n(A)f$ and $Q_n(A)f$ are calculated by using the three point recurrence relations. The procedures $P_n(A)f$ and $Q_n(A)f$ are compared in detail. From the numerical results it is evident that the method $P_n(A)f$ is slightly superior to the method $Q_n(A)f$. This is mainly due to our modification of the original recursion formula for the polynomials $P_n(A)$. The classical Chebyshev acceleration scheme is discussed in detail (page 26). It is shown that the Chebyshev scheme [eqn (3.17)] yields after n iterations ($n-1$ matrix vector multiplications) the same result as the vector $Q_{n-1}(A)f$, which is evaluated iteratively. The disadvantage of the Chebyshev scheme is that the intermediate results have no meaning and can go off the number representation scale on the computer (see [2]). The polynomial scheme $Q_{n-1}(A)f$ yields an approximate solution after each iteration. This should be considered as a great advantage of the present method.

It is shown in chapter 4 that a non-symmetric matrix vector system can be symmetrized by using the adjoint operator $A^\star$ and solved by using the methods given in chapter 3. If the operator A is non-symmetric but close in the sense of operator norm to the symmetric operator $L = (A+A^\star)/2$ and the operator L is bounded and positive definite, then the system can still be solved by the methods of chapter 3 without using the symmetrizing procedure. The above method is characterized by a greater rate of

convergence than the symmetrizing procedure. Furthermore the solution of such slightly non-symmetric systems by using the polynomials $Q_n(A)$ yields better results than when the polynomials $P_n(A)$ are implemented.

A comprehensive analysis of papers involving the evaluation of linear functionals of the form $< x,g >$, where $g \in H$, is presented in chapter 5. The method in all these papers involve deriving upper and lower bounds to the functional $< x,g >$ in order to avoid calculation of the solution x. We show that results from the literature all involve the zero order $(n = 0)$ approximation to A^{-1}. This point had not been realized by the relevant authors. In addition we show that the use of variational methods to evaluate the linear functional $< x,g >$ is too expensive for applications. It seems best to first evaluate the solution x by using the polynomial methods developed in chapter 3 and then to directly evaluate the functional $< x,g >$.

In chapter 6 we consider the application to Fredholm integral equations of the second kind. Both smooth and weakly singular kernels are considered. If the solution of the integral equation is required at a particular point, we show how it can be evaluated by using the polynomials $P_n(A)$ or $Q_n(A)$. The polynomials $Q_n(A)$ provide better results than the polynomials $P_n(A)$. However, both techniques yield far superior results than existing methods. A discretization of the integral equation yields a slightly non-symmetric matrix vector system. The discretization is accomplished by using Simpson quadrature for smooth kernels and a technique by Atkinson [3] for weakly singular kernels. The method of chapter 4 is applied to calculate the solution. The solution differs little from each other when the polynomials $P_n(A)$ or $Q_n(A)$ are used.

The results in appendix A pertaining to the polynomials $P_n(A)$ are quoted in Bond and Mika [1]. We show in detail how these results are derived. In appendix B we derive

results pertaining to the polynomial $Q_n(A)$. Appendix C contains a review of some of the earlier work done on the approximation to the inverse operator. In appendix D we derive in detail some of the results in chapter 5 on the application to linear functionals.

CHAPTER 1

FUNCTIONS OF NORMAL OPERATORS

Let H be a Hilbert space with an inner product $<.,.>$ and norm $\|.\|$. Let $B(H)$ be the Banach algebra of all bounded linear operators in H and $N(H)$ the set of all normal linear operators in H. Then $N_b(H) = B(H) \cap N(H)$ is the set of all bounded normal linear operators in H.

If $y \in H$, $x \in D(A) \subset H$, where $D(A)$ is the domain of a normal operator A, then by the spectral theory (see [4] and [1]).

$$< Ax, y > = \int_{\sigma(A)} \lambda \; d(< E_A(\lambda)x, y >), \tag{1.1}$$

where $\sigma(A)$ is the spectrum of A and $E_A(\lambda)$ the family of projection operators. Equation (1.1) is usually written in an abbreviated form as

$$Ax = \int_{\sigma(A)} \lambda \; d(E_A(\lambda)x); \qquad x \in D(A),$$

which if $A \in N_b(H)$ and $D(A) = H$ is simply written as

$$A = \int_{\sigma(A)} \lambda \; dE_A(\lambda). \tag{1.2}$$

The domain $D(A)$ is given by

$$D(A) = \left\{ x \in H : \int_{\sigma(A)} |\lambda|^2 \, d(\|E_A(\lambda)x\|^2) < \infty \right\}.$$

If $A \in N_b(H)$ then $\|A\|$ is just the spectral radius

$$\|A\| = spr(A) = sup \{ |\lambda| : \lambda \in \sigma(A) \}.$$

Let $L^{(A)}$ be the space of all complex valued functions of a complex variable continuous on a domain containing $\sigma(A)$, and equip $L^{(A)}$ with the seminorm

$$\|f\|^{(A)} = sup \{ |f(\lambda)| : \lambda \in \sigma(A) \}; \quad f \in L^{(A)}. \tag{1.3}$$

For each $A \in N(H)$ and $f \in L^{(A)}$ one can define an operator $\Psi_A(f) \in N_b(H)$ by the formula

$$\Psi_A(f) = \int_{\sigma(A)} f(\lambda) \, dE_A(\lambda). \tag{1.4}$$

Since [4]

$$\|\Psi_A(f)\| = \|f\|^{(A)}, \tag{1.5}$$

for a fixed $A \in N(H)$ equation (1.4) defines an isometric mapping $\Psi_A : L^{(A)} \to N_b(H)$ and for a fixed $f \in L$ an operator function $\Psi(f) : N(H) \to N_b(H)$, where $A \mapsto \Psi_A(f)$. By relaxing restrictions on f one can define more general operator functions $\Psi(f)$ whose values are unbounded normal operators.

It is customary to abuse slightly the notation and write $\Psi_A(f)$ as $f(A)$ (see [4]).

Hence instead of (1.4) we have

$$f(A) = \int_{\sigma(A)} f(\lambda)\, dE_A(\lambda)\,, \tag{1.6}$$

and from (1.5)

$$\|f(A)\| = sup\,\{\, |f(\lambda)| : \lambda \in \sigma(A)\,\} \tag{1.7}$$

As an example take $f \in L^{(A)}$ defined by $f(\lambda) = (\lambda - \mu)^{-1}$, $\mu \notin \sigma(A)$. From (1.6) $f(A) = (A - \mu I)^{-1}$ is the resolvent operator. If $0 \notin \sigma(A)$, take $\mu = 0$, then $f(\lambda) = \lambda^{-1}$ and $f(A) = A^{-1}$ represents the inverse operator.

CHAPTER 2

METHOD OF APPROXIMATE INVERSE

Consider the linear equation

$$Ax = f, \tag{2.1}$$

where $f \in H$ and A is a bounded positive definite operator from H into itself that is bounded below by $m > 0$ and above by $M > m$ so that the following inequality is satisfied.

$$m < \phi,\phi > \leq < A\phi,\phi > \leq M < \phi,\phi >; \forall\ \phi \in H\,. \tag{2.2}$$

Hence $\sigma(A) \subset [m,M]$. If V is an operator approximating the inverse A^{-1} and $\tilde{x} = Vf$ an approximate solution of (2.1) then the error can be written as

$$x - \tilde{x} = (A^{-1} - V)f = (I - VA)x\,, \tag{2.3}$$

where $x = A^{-1}f$ is the exact solution of (2.1). The above formula suggests two possible approaches to evaluating V (see [1]). It can be chosen as a polynomial $P_n(A)$ when optimizing $\|A^{-1} - V\|$ or a polynomial $Q_n(A)$ when optimizing $\|I - VA\|$. From (1.7) with $f(A) = A^{-1} - P_n(A)$ or $f(A) = I - A\,Q_n(A)$ we have

$$\epsilon_n^P = \|A^{-1} - P_n(A)\| = \sup_{\lambda \in \sigma(A)} |1/\lambda - P_n(\lambda)| \tag{2.4}$$

$$\epsilon_n^Q = \|I - A\,Q_n(A)\| = \sup_{\lambda \in \sigma(A)} |1 - \lambda\,Q_n(\lambda)|\,, \tag{2.5}$$

Thus the problem is reduced to that of approximating functions of the real variable by polynomials of best uniform approximation. As in writing (1.6), here we also use the same symbols P_n or Q_n to denote both the polynomial in A and the polynomial in λ. This, however, should not lead to any confusion.

It has been proved by Chebyshev (see [5]) that for any function $f(\lambda)$ defined and continuous on a closed interval $[m,M]$ and for any non-negative integer n, there exists a unique polynomial $h_n(\lambda)$ of degree n which deviates least from $f(\lambda)$ over the interval $[m,M]$. This polynomial of best approximation is characterized by the fact that in the interval $[m,M]$ the number of consecutive points at which the difference $f(\lambda) - h_n(\lambda)$ with alternate change of signs assumes the maximum value $\|f(\lambda) - h_n(\lambda)\|^{[m,M]}$ is not less than $n+2$. Here

$$\|g(\lambda)\|^{[m,M]} = sup\{\ |g(\lambda)| : \lambda \in [m,M]\ \},$$

similarly as was done in (1.3).

THE POLYNOMIALS $P_n(A)$

Since the norms $\|.\|$ and $\|.\|^{(A)}$ are equivalent by (1.5), the problem of approximating the inverse A^{-1} by a polynomial $P_n(A)$ in the uniform operator norm is equivalent to the problem of the uniform approximation of the function $1/\lambda$ by a polynomial $P_n(\lambda)$ on the interval $[m,M]$.

Chebyshev found an explicit expression for a polynomial of best approximation $\tilde{r}_n(t)$ of arbitrary degree n to the function $1/(u-t)$ for $-1 \leq t \leq 1$ and $u > 1$ [Appendix A].

This is given by

$$\tilde{r}_n(t) = \frac{2\delta}{1-\delta^2}\left[\, 1 + 2\sum_{k=1}^{n-1} \delta^k\, T_k(t) + \frac{2\delta^n}{1-\delta^2}\, T_n(t)\,\right];\ n \geq 1, \tag{2.6}$$

where $\delta = u - \sqrt{u^2 - 1}$ and the optimal error of approximation is

$$\|1/(u-t) - \tilde{r}_n(t)\|^{[-1,1]} = \left[\frac{2\delta}{1-\delta^2}\right]^2 \delta^n = \frac{\delta^n}{u^2-1}. \tag{2.7}$$

Here $T_n(t) = cos\ n(cos^{-1}t)$ is the Chebyshev polynomial of degree n. The error in (2.7) is optimal in the sense that for an arbitrary polynomial $r_n(t) \neq \tilde{r}_n(t)$ we have

$$\|1/(u-t) - r_n(t)\|^{[-1,1]} > \|1/(u-t) - \tilde{r}_n(t)\|^{[-1,1]}.$$

To approximate $f(\lambda) = 1/\lambda$ one has to introduce a new variable

$$t = \frac{M+m-2\lambda}{M-m}, \tag{2.8}$$

such that $t \in [-1,1]$. Since $1/\lambda = a/(u-t)$, where $u = (M+m)/(M-m)$ and $a = 2/(M-m)$, it follows that the polynomial best approximating $1/\lambda$ in $[m,M]$ is $P_n(\lambda) = a\tilde{r}_n(t)$ or

$$P_n(\lambda) = \frac{1}{\sqrt{Mm}}\left[\, 1 + 2\sum_{k=1}^{n-1} \delta^k\, T_k(t) + \frac{2\delta^n}{1-\delta^2}\, T_n(t)\,\right];\ n \geq 1, \tag{2.9}$$

where by expressing δ in terms of m and M we have $\delta = (\sqrt{M}-\sqrt{m})/(\sqrt{M}+\sqrt{m})$ and t

in the right hand side has to be replaced by λ according to (2.8). The optimal error ϵ_n^P is given by.

$$\epsilon_n^P = \frac{a\delta^n}{u^2\text{-}1} = \frac{1}{2}\left[\frac{1}{m} - \frac{1}{M}\right]\delta^n. \tag{2.10}$$

The zero order polynomial $P_0(\lambda)$ for which the optimal error is $\epsilon_0^P = 1/2\ (1/m\text{-}1/M)$ can easily be deduced from the simple sketch in fig 2.1 [see also Appendix A].

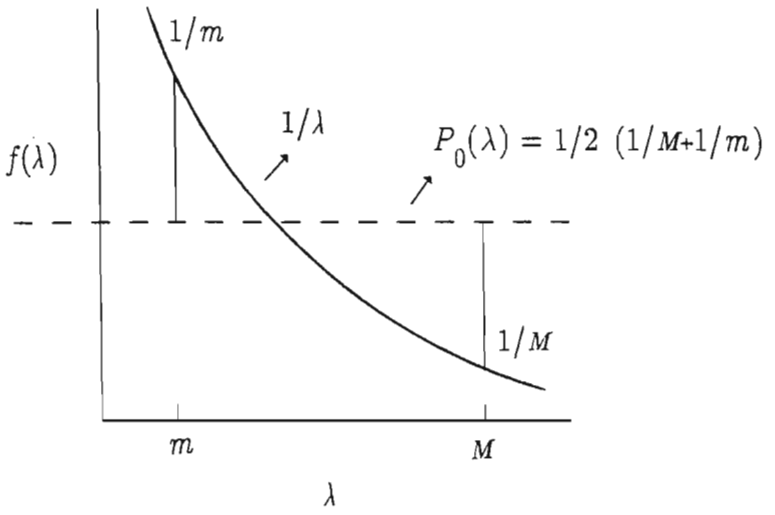

FIG 2.1: Polynomial $P_0(\lambda)$ approximating the function $1/\lambda$.

Thus

$$P_0(\lambda) = \frac{1}{2}\left[\frac{1}{M} + \frac{1}{m}\right].$$

The next polynomial,

$$P_1(\lambda) = \frac{1}{\sqrt{Mm}}\left[1 + \frac{2\delta}{1\text{-}\delta^2}\,T_1(t)\right] = \frac{(\sqrt{M}+\sqrt{m})^2}{2\,mM} - \frac{\lambda}{Mm}$$

can be obtained from equation (2.9) or directly [see Appendix A].

The polynomials $P_n(\lambda)$ satisfy the recurrence relation [see Appendix A]

$$P_{n+2}(\lambda) = 2\delta t\, P_{n+1}(\lambda) - \delta^2\, P_n(\lambda) + 2a\delta. \tag{2.11}$$

Thus the polynomials $P_n(A)$ can be evaluated from the recurrence relation

$$P_{n+2}(A) = 2\delta t_A\, P_{n+1}(A) - \delta^2\, P_n(A) + 2a\delta I, \tag{2.12}$$

where $t_A = uI - aA$, with

$$P_0(A) = \frac{1}{2}\left[\frac{1}{M} + \frac{1}{m}\right] I, \tag{2.13}$$

and

$$P_1(A) = \frac{(\sqrt{M}+\sqrt{m})^2}{2\,mM}\, I - \frac{1}{Mm}\, A\ . \tag{2.14}$$

THE POLYNOMIALS $Q_n(A)$

It is clear that optimization of equation (2.5) requires the polynomial $Q_n(\lambda)$ such that $\lambda\ Q_n(\lambda)$ best approximates the function $f(\lambda) = 1$ in $[m,M]$ or such that

$$Z_{n+1}(\lambda) = 1 - \lambda\ Q_n(\lambda), \tag{2.15}$$

best approximates zero in $[m,M]$. We notice that $Z_{n+1}(\lambda)$ has to satisfy the condition

$$Z_{n+1}(0) = 1\ . \tag{2.16}$$

The best polynomial of degree $n+1$ approximating zero in [-1,1] is $2^{-n}\,T_{n+1}(t)$ for $-1 \leq t \leq 1$. The preceding result can be found in the literature (see [6]). To approximate zero in $[m,M]$ by a polynomial satisfying (2.16), we make the change of variable as in (2.8). Returning to the original variable λ we have

$$2^{-n}\,T_{n+1}\left[\frac{M+m-2\lambda}{M-m}\right]. \tag{2.17}$$

Since we have the additional requirement (2.16), we must have

$$Z_{n+1}(\lambda) = \left[\,T_{n+1}\left[\frac{M+m}{M-m}\right]\,\right]^{-1} T_{n+1}\left[\frac{M+m-2\lambda}{M-m}\right]. \tag{2.18}$$

The maximum error of approximation is given by [Appendix B]

$$\left[\,T_{n+1}\left[\frac{M+m}{M-m}\right]\,\right]^{-1} = \frac{2}{\delta^{n+1}+\delta^{-(n+1)}}, \tag{2.19}$$

where as previously $\delta = (\sqrt{M}-\sqrt{m})/(\sqrt{M}+\sqrt{m})$. Since $Q_n(\lambda)$ is related to $Z_{n+1}(\lambda)$ by equation (2.15), the error of approximation ϵ_n^Q using polynomials $Q_n(\lambda)$ is also given by (2.19).

From (2.15), (2.18) and (2.19), we obtain the expression

$$Q_n(\lambda) = \frac{1}{\lambda}\left[\,1 - \frac{2}{\delta^{n+1}+\delta^{-(n+1)}}\,T_{n+1}\left[\frac{M+m-2\lambda}{M-m}\right]\,\right]. \tag{2.20}$$

The polynomial $Q_0(\lambda)$ can easily be deduced from the simple sketch in fig 2.2 or from (2.20) [see Appendix B].

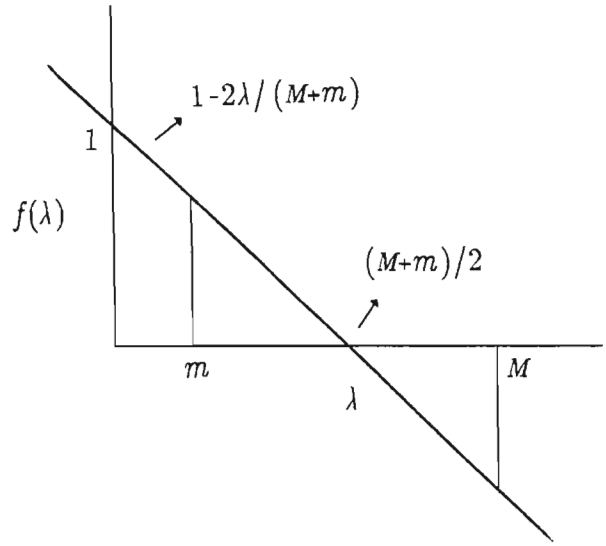

FIG 2.2: Polynomial $Q_0(\lambda)$ such that $1-\lambda Q_0(\lambda)$ approximates the function zero.

Thus

$$Q_0(\lambda) = \frac{2}{M+m}.$$

The next polynomial

$$Q_1(\lambda) = \frac{8(M+m)-8\lambda}{(M+m)^2+4Mm},$$

can be obtained from (2.20) [Appendix B].

The polynomials $Q_n(\lambda)$ satisfy the recurrence relation [Appendix B]

$$Q_{n+2}(\lambda) = Q_n(\lambda) + \frac{1+\delta^{2n+4}}{1+\delta^{2n+6}}\left[\,2\delta t\, Q_{n+1}(\lambda) - 2\delta u\, Q_n(\lambda) + 2a\delta\,\right]. \qquad (2.21)$$

Thus the polynomials $Q_n(A)$ are evaluated from the recurrence relation

$$Q_{n+2}(A) = Q_n(A) + \frac{1+\delta^{2n+4}}{1+\delta^{2n+6}} [\, 2\delta t_A \, Q_{n+1}(A) - 2\delta u \, Q_n(A) + 2a\delta I \,], \quad (2.22)$$

where $t_A = uI - aA$, with

$$Q_0(A) = \frac{2}{M+m} I, \quad (2.23)$$

and

$$Q_1(A) = \frac{8}{(M+m)^2+4Mm} \left[\, (M+m)I - A \, \right] \quad (2.24)$$

After this necessary background we are now in a position to follow the review, presented in Appendix C, of some of the earlier work done on the approximation to the inverse operator.

CHAPTER 3

SOLUTION OF LINEAR SYSTEMS WITH SYMMETRIC MATRICES

We consider the approximate solution of the system of linear equations

$$Ax = f. \tag{3.1}$$

A is a bounded positive definite matrix operator and hence selfadjoint. Among various methods, the approximate solution of (3.1) can be achieved by means of the cyclic iterative scheme

$$x_{k+1} = x_k + R_n(A)\ (f - Ax_k),\ \ k = 0,1,2,\cdots, \tag{3.2}$$

where $R_n(A)$ is implemented iteratively and x_0 is a starting vector. Here we will take $R_n(A)$ as $P_n(A)$ or $Q_n(A)$.

First we summarize some of the other methods (see for instance [7]) available for solving (3.1). The Jacobi method

$$x_{k+1} = (I - D^{-1}A)x_k + D^{-1}f, \tag{3.3}$$

or equivalently

$$x_{k+1} = x_k + D^{-1}(f - Ax_k), \tag{3.4}$$

has been used to solve matrix vector problems. Here D is a diagonal matrix composed of the diagonal elements of A.

The extrapolated Jacobi method

$$x_{k+1} = (I - \omega D^{-1}A)x_k + \omega D^{-1}f, \tag{3.5}$$

or equivalently

$$x_{k+1} = x_k + \omega D^{-1}(f - Ax_k), \tag{3.6}$$

is a modification of the Jacobi technique involving a relaxation parameter ω. Convergence is guaranteed as long as ω is chosen so that $\| I - \omega D^{-1}A \| < 1$.

The stationary Richardson method

$$x_{k+1} = x_k + \omega (f - Ax_k), \tag{3.7}$$

has been used to solve operator equations, in particular integral equations by Kleinman et al [8].

The successive overrelaxation (SOR) method

$$x_{k+1} = [I - (\omega^{-1}D + L_\Delta)^{-1} A]\, x_k + (\omega^{-1}D + L_\Delta)^{-1} f, \tag{3.8}$$

where D is a diagonal matrix composed of the diagonal elements of A and L_Δ is a matrix composed of the lower triangular part of A, has proved particularly useful in solving linear systems arising from difference equations for solution of elliptical partial differential equations (see [7]). For convergence the iteration matrix $I - (\omega^{-1}D + L_\Delta)^{-1}A$ must satisfy the inequality $\|I - (\omega^{-1}D + L_\Delta)^{-1}A\| < 1$ thus

restricting ω to the range $0 < \omega < 2$. We note that (3.8) can be expressed in the equivalent form

$$x_{k+1} = x_k + (\omega^{-1}D + L_\Delta)^{-1} (f - Ax_k) \tag{3.9}$$

The classical Chebyshev iteration method

$$x_{k+1} = x_k + \beta_k(f - Ax_k) : k = 0,1,\cdots n\text{-}1 , \tag{3.10}$$

where the parameters β_k are connected with the roots of the Chebyshev polynomials, has been used since the early fifties to solve equations of the form (3.1) (see [2]). When the scheme of (3.10) is implemented on a computer with a fixed number of digits and when the operator A is ill-posed, a loss of significant digits can occur in the intermediate and final results in x_k and the values of the intermediate iterates may go off the number representation scale in the computer. This instability has inhibited the use of this technique in certain cases. Young [9] has succeeded in reducing computational instabilities by showing how the β_k should be implemented. Lebedev and Finogenov [2] showed how the β_k should be properly ordered to eliminate instability. However, it must be stressed that intermediate values of x_k $(k = 0,1,\cdots n\text{-}1)$ have no meaning, they are just a means to get to the approximate solution x_n. The Chebyshev scheme is discussed in detail on page 26 and we compare it to the polynomial scheme $Q_{n-1}(A)f$.

The technique in (3.7) can be obtained by multiplying (3.1) by a relaxation parameter ω and letting $\omega A = I - B$. Hence one obtains the equation $x = Bx + \omega f$ or the iteration scheme $x_{k+1} = Bx_k + \omega f$, which is equivalent to (3.7) upon substitution of $B = I - \omega A$. Convergence is guaranteed as long as $\| B \| = \| I - \omega A \| < 1$. If A is positive definite and $\sigma(A) \subset [m,M]$ then minimizing $\| I - \omega A \|$ yields the optimal choice for ω, namely $\omega = 2/(M+m)$. Here $A^{-1} \approx \omega I$ with $\omega = Q_0(\lambda)$.

The technique we consider here is derived in a similar way, except that equation (3.1) is multiplied by the polynomial $R_n(A)$ rather than by ω.

If ω is replaced by ω_k in (3.7) then one obtains the general relaxation scheme $x_{k+1} = x_k + \omega_k\ (f - Ax_k)$, where the relaxation parameter ω_k need not be constant at each step. One however has the difficulty of determining suitable ω_k, before each iteration is performed, in order that convergence may occur and be speeded up. In the same spirit $R_n(A)$ could be replaced by $R_n^k(A)$ in (3.2). This indicates that polynomials $R_n(A)$ of different order could be used at each step. Although these \olynomials are readily available to us, we choose for convenience to use the scheme indicated in (3.2) with polynomials of fixed order.

Now the error $e_k = x - x_k$, after k cycles, can be expressed in terms of e_0 by using equation (3.2). Hence

$$\begin{aligned} e_k &= x - x_k \\ &= x - x_{k-1} - R_n(A)\ (Ax - Ax_{k-1}) \\ &= [\ I - AR_n(A)\ ]\ e_{k-1} \\ &= [\ I - AR_n(A)\ ]^k\ e_0\ . \end{aligned} \tag{3.11}$$

Hence we obtain convergence as long as $\| I - AR_n(A) \| < 1$.

We now devise a procedure to check the methods for matrix vector systems of the form (3.1). The $N \times N$ matrix A is generated from an orthogonal transformation using Householder matrices [10]. Hence

$$A = U\ D\ U, \tag{3.12}$$

where $U = I - 2\ w\ w^t$ is the $N \times N$ Householder matrix, with w a chosen real N column vector such that $w^t w = \|w\|_2^2 = 1$ and D is a diagonal $N \times N$ matrix with chosen (positive) eigenvalues. Householder matrices are chosen because of their orthogonal and symmetry property. We choose the vector w to have identical components $1/\sqrt{N}$ .The distribution of eigenvalues λ_k is chosen such that

$$\lambda_k = m + \frac{k\text{-}1}{k+1}(M\text{-}m) : k = 1,2,\cdots N\text{-}1\ , \tag{3.13}$$

with $\lambda_N = M$. It should be noted that similar results are obtained with other distributions of eigenvalues. A suitable vector f is chosen for the right hand side of equation (3.1). Since $A^{-1} = U\,D^{-1}\,U$, the exact solution of (3.1) can be obtained.

The calculation of the vector $P_n(A)\ (f - Ax_k)$ of equation (3.2) can be achieved in two ways. Firstly we can evaluate $P_0(A)$ and $P_1(A)$ and generate $P_n(A)$ from the recurrence relation (2.12). We can then multiply by the residual $(f - Ax_k)$. Alternatively the recurrence relation (2.12) can first be modified by multiplying by the residual $(f - Ax_k)$. We can evaluate the vectors $P_0(A)\ (f - Ax_k)$ and $P_1(A)\ (f - Ax_k)$ and generate the vector $P_n(A)\ (f - Ax_k)$ by using this modification of (2.12). We choose the latter technique since it involves vector iterates and is less expensive (from a computational point of view) than the former technique which involves matrix iterates. The vector $Q_n(A)\ (f - Ax_k)$ is calculated in a similar fashion by using a modification of the recurrence relation (2.22). Throughout this thesis the recurrence relations are used only to evaluate vector iterates. Therefore wherever mention is made of using the recurrence relations, it is to be understood in the above sense, where the necessary modification must first be made.

It is best to use $x_0 = 0$ as a starting value in equation (3.2) since this saves on one

matrix vector multiplication (henceforth called operation). Examining the first iterate $x_1 = x_0 + R_n(A)(f - Ax_0)$, one notes that it involves $n+1$ operations (if $x_0 \neq 0$). Clearly $R_{n+1}(A)f$ would give better results for the same effort.

Consider the direct approach $x_n = P_n(A)f$ and $x_n = Q_n(A)f$ for solving (3.1). The corresponding errors $e_n^P = x - P_n(A)f$ and $e_n^Q = x - Q_n(A)f$ satisfy the inequalities

$$\begin{aligned} \| e_n^P \| &\leq \| A^{-1} - P_n(A) \| \, \| f \| \\ &= \frac{1}{2} \left[\frac{1}{m} - \frac{1}{M} \right] \delta^n \| f \|, \end{aligned} \tag{3.14}$$

and

$$\begin{aligned} \| e_n^Q \| &\leq \| I - A\, Q_n(A) \| \, \| x \| \\ &= \frac{2}{\delta^{n+1} + \delta^{-(n+1)}} \| x \|. \end{aligned} \tag{3.15}$$

Here we have used equations (2.4), (2.5), (2.10) and (2.19). We note that (3.14) gives an a priori error bound and hence the solution can be determined to any prescribed tolerance. The method of solution is quite simple. Starting with $R_0(A)f$ and $R_1(A)f$, generate higher order approximations $R_n(A)f$ by using the recurrence relations (2.12) or (2.22). The results using this technique are presented in table 3.1 . Note that we present actual errors.

As evident from the table, the effects of roundoff error appear sooner using equation (2.12) than (2.22). Whilst theoretically the error should decrease with increasing n, for $n = 23, 24, 25$ the error remains constant using (2.12). This roundoff error is due to the δ^2 coefficient in the second term of equation (2.12). Since $\delta < 1$, we lose significant digits in computing the second term $\delta^2 P_n(A)$ in (2.12). This error accumulates and

TABLE 3.1: Solution of a symmetric matrix vector system

n	Error (P_n) using (2.12)		Error (P_n) using (3.16)		Error (Q_n) using (2.22)	
	$\|\ \|_\infty$	$\|\ \|_2$	$\|\ \|_\infty$	$\|\ \|_2$	$\|\ \|_\infty$	$\|\ \|_2$
10	0.216e-4	0.362e-4	0.216e-4	0.363e-4	0.390e-4	0.530e-4
15	0.142e-6	0.269e-6	0.165e-6	0.317e-6	0.297e-6	0.451e-6
20	0.542e-7	0.784e-7	0.147e-8	0.267e-8	0.241e-8	0.357e-8
23	0.556e-7	0.799e-7	0.901e-10	0.145e-9	0.143e-9	0.195e-9
24	0.556e-7	0.800e-7	0.312e-10	0.599e-10	0.513e-10	0.770e-10
25	0.556e-7	0.800e-7	0.128e-10	0.223e-10	0.213e-10	0.299e-10

Solution of a symmetric 10×10 matrix vector system using the direct approach $P_n(A)f$ and $Q_n(A)f$. Here $[m,M\,] = [1,5]$, $\delta = 0.38197$, the components $f_k(k = 1,\ 2,\cdots 10)$ of f are chosen obey the relation $f_k = 0.1k$ $(k = 1,\ 2,\cdots 10)$, $\|f\|_\infty = 1.0$ and $\|f\|_2 = 1.96214$.

it's effect is evident after several iterations. We therefore propose to rewrite equation (2.12) in the form

$$P_{n+2}(A) = P_n(A) + 2\delta t_A\, P_{n+1}(A) - 2\delta u\, P_n(A) + 2a\delta I, \tag{3.16}$$

where we have used the identity $\delta^2 = 2\delta u - 1$. In table 3.1 we also present results using the relation (3.16) to generate $P_n(A)f$. It is clear that the results are superior using the relation (3.16) instead of (2.12). Comparing the recurrence relations (3.16) and (2.22) for $P_n(A)$ and $Q_n(A)$ respectively, we observe a remarkable similarity because the coefficient $(1+\delta^{2n+4})/(1+\delta^{2n+6})$ in (2.22) is close to unity with large n since $\delta < 1$. Thus it is no wonder that (3.16) and (2.22) behave in a similar fashion. From now on we will therefore use the relation (3.16) for the polynomials $P_n(A)$ as it is computationally more stable than (2.12).

For the selfadjoint case, it is simpler and yields slightly better results to implement a direct rather than cyclic approach. Using a cyclic approach of k cycles with $x_0 = 0$ requires $k(n+1)-1$ operations. Better results should be obtained by using the direct approach $R_{k(n+1)-1}(A)f$ as an approximation to the solution. This is evident by comparing the results presented in tables 3.2 and 3.3, where we solve the same problem by using the direct as well as cyclic approaches. For example, 32 operations (using P_{32}) in table 3.2 yields an error that is roughly 10^2 times smaller than when 32 operations (3 cycles of P_{10}) are implemented in table 3.3 .

We compare the procedures $P_n(A)f$ and $Q_n(A)f$ by examining the results presented in tables 3.4 and 3.5 . It is seen that the results are better using the procedure $P_n(A)f$. Rewriting equation (3.14) and (3.15), we obtain

$$\| e_n^P \| \leq \frac{2\delta}{1+\delta^2} \left\| \frac{1}{2}\left[\frac{1}{m}+\frac{1}{M}\right] f \right\| \delta^n$$

and

$$\| e_n^Q \| \leq \frac{2\delta}{1+\delta^{2n+2}} \| x \| \delta^n.$$

Since $1/2\,(1/m + 1/M)\,f$ is the zero order approximation to x , the upper bound for $\| e_n^P \|$ is less than that for $\| e_n^Q \|$. This gives some indication why $P_n(A)f$ gives better results than $Q_n(A)f$. We note that the errors in tables 3.4 and 3.5 satisfy the inequalities (3.14) and (3.15) for the upper bounds (for the induced norm $\|.\|_2$), except for $n = 25\,(Q_{25})$ and $n = 26$ $(P_{26}$ and $Q_{26})$ in table 3.4. This is due to the effect of roundoff error which occurs sooner for small δ as compared to large δ . We recall the dependence of the recurrence relations on δ . In comparing tables 3.4 and 3.5 we notice a dramatic decrease in the rate of convergence due to large δ , making it necessary for higher order approximations to achieve the same accuracy. However, it must be mentioned that the use of very high order polynomial approximations involve a large

number of successive iterations which can lead to inaccuracies due to roundoff error. For such cases it would be better to implement low order polynomials in a cyclic approach.

TABLE 3.2: Solution of a symmetric matrix vector system

n	Error (P_n)		Error (Q_n)	
	$\|\ \|_\infty$	$\|\ \|_2$	$\|\ \|_\infty$	$\|\ \|_2$
10	0.324e-2	0.701e-2	0.580e-2	0.738e-2
21	0.104e-4	0.243e-4	0.219e-4	0.268e-4
32	0.561e-7	0.894e-7	0.814e-7	0.982e-7
43	0.151e-9	0.298e-9	0.283e-9	0.357e-9

Solution of a symmetric 50×50 matrix vector system using the direct approach $P_n(A)f$ and $Q_n(A)f$. Here $[m,M] = [1,16]$, $\delta = 0.6$, $f_k = 1.0$ $(k = 1, 2, \cdots 50)$, $\|f\|_\infty = 1.0$ and $\|f\|_2 = 7.07106$.

TABLE 3.3: Solution of a symmetric matrix vector system

n	Ops	Error (P_n)		Error (Q_n)	
		$\|\ \|_\infty$	$\|\ \|_2$	$\|\ \|_\infty$	$\|\ \|_2$
10	10	0.324e-2	0.701e-2	0.580e-2	0.738e-2
	21	0.992e-4	0.207e-3	0.384e-4	0.532e-4
	32	0.533e-5	0.779e-5	0.310e-6	0.385e-6
	43	0.149e-6	0.316e-6	0.209e-8	0.279e-8

Solution of the problem presented in table 3.2 using the cyclic technique described in equation (3.2). Ops – denote the number of matrix vector operations at the end of each iteration or cycle.

TABLE 3.4: Solution of a symmetric matrix vector system

n	Error (P_n)		Error (Q_n)	
	$\|\ \|_\infty$	$\|\ \|_2$	$\|\ \|_\infty$	$\|\ \|_2$
0	0.445	0.462	0.428	0.516
1	0.113	0.143	0.159	0.174
2	0.260e-1	0.382e-1	0.289e-1	0.451e-1
5	0.511e-3	0.678e-3	0.535e-3	0.851e-3
10	0.508e-6	0.846e-6	0.782e-6	0.113e-5
15	0.768e-9	0.136e-8	0.977e-9	0.171e-8
20	0.986e-12	0.183e-11	0.134e-11	0.224e-11
25	0.172e-14	0.296e-14	0.233e-14	0.361e-14
26	0.611e-15	0.977e-15	0.833e-15	0.118e-14

Solution of a symmetric 10 × 10 matrix vector system using the direct approach $P_n(A)f$ and $Q_n(A)f$. Here $[m,M] = [1,3]$ $\delta = 0.26795$, $f_k = 0.1k$ $(k = 1, 2, \cdots 10)$, $\|f\|_\infty = 1.0$, $\|f\|_2 = 1.96214$, $\|x\|_\infty = 0.58844$ and $\|x\|_2 = 1.30681$.

TABLE 3.5: Solution of a symmetric matrix vector system.

n	Error (P_n)		Error (Q_n)	
	$\|\ \|_\infty$	$\|\ \|_2$	$\|\ \|_\infty$	$\|\ \|_2$
0	0.834e+2	0.957e+2	0.785e+2	0.980e+2
10	0.593e+1	0.111e+2	0.173e+2	0.217e+2
40	0.181e-1	0.223e-1	0.427e-1	0.535e-1
60	0.336e-3	0.428e-3	0.771e-3	0.966e-3
80	0.671e-5	0.884e-5	0.139e-4	0.175e-4
100	0.930e-7	0.164e-6	0.254e-6	0.315e-6
150	0.374e-11	0.670e-11	0.111e-10	0.137e-10

Solution of a symmetric 10 × 10 matrix vector system using the direct approach $P_n(A)f$ and $Q_n(A)f$. Here $[m,M] = [0.01,1]$, $\delta = 0.81818$, $f_k = 0.1k$ $(k = 1, 2, \cdots 10)$, $\|f\|_\infty = 1.0$ $\|f\|_2 = 1.96214$, $\|x\|_\infty = 78.350031$ and $\|x\|_2 = 100.06323$.

CHEBYSHEV ACCELERATION TECHNIQUE

The Chebyshev acceleration technique has originally been used for systems of linear equations but it can be extended to equations in Hilbert space (see [11]). For the equation (3.1), we consider the general scheme

$$x_{k+1} = x_k + \beta_k(f - Ax_k) : k = 0,1,\cdots n\text{-}1, \tag{3.17}$$

proposed by Richardson [11,12]. Here $\beta_0, \beta_1, \cdots \beta_{n-1}$ are some non $-$ zero numbers and the value of x_0 is assumed given (usually $x_0 = 0$).

The error e_k after k iterations is given by

$$\begin{aligned} e_k &= x - x_k \\ &= (I - \beta_{k-1}A)\ e_{k-1}. \end{aligned}$$

Hence after n steps

$$\begin{aligned} e_n &= \prod_{k=0}^{n-1} (I - \beta_k A)\ e_0 \\ &= H_n(A)\ e_0, \end{aligned} \tag{3.18}$$

where $H_n(A)$ is a polynomial operator function corresponding to the n_{th} degree polynomial $H_n(\lambda) = \prod_{k=0}^{n-1} (1 - \beta_k\lambda)$. Since the inequality $\| e_n \| \leq \| H_n(A) \| \| e_0 \|$ is satisfied, the optimal error is obtained if $H_n(\lambda)$ is chosen as the best polynomial approximating zero in the interval $[m,M]$, with the additional property that $H_n(0) = 1$. But this polynomial has already been found and is given by $Z_n(\lambda)$ from equation (2.18).

Hence

$$H_n(\lambda) = Z_n(\lambda) = \prod_{k=0}^{n-1} (1 - \beta_k \lambda) = \left[T_n\left[\frac{M+m}{M-m} \right] \right]^{-1} T_n\left[\frac{M+m-2\lambda}{M-m} \right], \quad (3.19)$$

so that the optimal values $\beta_0, \beta_1, \cdots \beta_{n-1}$ are equal to the reciprocal of the roots λ_k of the Chebyshev polynomials, given by

$$\lambda_k = \frac{1}{2}\left[M + m - (M - m)\ cos\left[(k + \frac{1}{2}) \frac{\pi}{n} \right] \right], \quad k = 0,1, \cdots\ n-1. \quad (3.20)$$

With $x_0 = 0$ in the standard Chebyshev scheme, one obtains from equation (3.18), with $Z_n(A)$ replacing $H_n(A)$,

$$\begin{aligned} x_n &= [\ I - Z_n(A)\ ]\ A^{-1} f \\ &= S_{n-1}(A)\ f, \end{aligned} \quad (3.21)$$

where $S_{n-1}(A) = [\ I - Z_n(A)\ ]\ A^{-1}$ is a polynomial of degree $n - 1$. Since $Z_n(A) = [I - A\ S_{n-1}(A)]$, it follows from equation (2.15) that $S_{n-1}(A) = Q_{n-1}(A)$. After n iterations of the scheme (3.17) one obtains the approximate solution $x_n = Q_{n-1}(A)f$. Hence the Chebyshev scheme (3.17) yields the same result after n iterations as does the vector $Q_{n-1}(A)f$, which is evaluated iteratively by using the recurrence relation (2.22). It must be noted that both schemes involve the same number $(n - 1)$ of operations (with $x_0 = 0$). The advantage of the polynomial scheme is that each iteration yields an approximate solution. However, with the Chebyshev scheme, each iterate x_k $(k = 1,2, \cdots\ n - 1)$ has no meaning. The approximation is only completed after the n_{th} iteration and it is this last iterate x_n that yields the approximate solution. This leads to instabilities (see [2]) as already discussed.

CONCLUSION

For symmetric systems the solution should be obtained by using the procedure $P_n(A)f$ where the $P_n(A)f$ are evaluated by using the recurrence relation (3.16) to generate vector iterates. It is essential to use relatively low order polynomials in order to minimize roundoff error. The order can be decided on from the magnitude of δ. For large δ one can use large order polynomials. For small δ, the order of the polynomials should be smaller. Recall that δ occurs in the recurrence relations (3.16) and (2.22) for $P_n(A)$ and $Q_n(A)$ respectively, which when implemented iteratively results in powers of δ. For large polynomial orders n the number of iterations are large resulting in greater powers of δ. Since $\delta < 1$, for small δ the value of these powers of δ are close to zero resulting in inexact representation on the computer. This explains why one cannot get errors smaller than the order of 10^{-15}. But usually errors of the order of 10^{-6} are more than sufficient.

For ill-conditioned problems (δ large) it is theoretically necessary to use high order polynomials to achieve the desired accuracy. However, as already explained, this may lead to roundoff errors due to large number of iterations. For such cases it is better to implement a cyclic approach using low polynomial orders (see problem in table 3.3), thereby reducing roundoff error.

In their calculations, Bond and Mika [1] used the recurrence relation (2.12) for the polynomials $P_n(A)$. However we have shown the relation (2.12) leads to inexact representation on the computer. From a computational point of view, we have provided an improved recurrence relation in the form of (3.16) for the polynomials $P_n(A)$.

CHAPTER 4

SOLUTION OF NON-SELFADJOINT SYSTEMS

Non-symmetric systems of the form (3.1) can be symmetrized by multiplying by the adjoint. Thus

$$A^{\star}A\,x = A^{\star}\,f = f' \tag{4.1}$$

and the resulting symmetric system

$$B\,x = f', \tag{4.2}$$

where $B = A^{\star}A$ is positive definite, can be solved as before. However the convergence rate is affected. This can be understood if we make the assumption

$$m^2 < \phi,\phi > \leq < A\phi,A\phi > \leq M^2 < \phi,\phi >;\ \forall\ \phi \in H\,. \tag{4.3}$$

Since $< A\phi,A\phi > = < A^{\star}A\phi,\phi >$, it follows that $\sigma(A^{\star}A) \subset [m^2,M^2]$. It also follows from (4.3) that $m \leq |\lambda| \leq M$, where $\lambda \in \sigma(A)$. The convergence parameter for the system (4.1) is now $\delta^{\star} = (M-m)/(M+m)$ as compared to $\delta = (\sqrt{M}-\sqrt{m})/(\sqrt{M}+\sqrt{m})$ had A been selfadjoint. It is easily verified that $\delta < \delta^{\star}$. Hence convergence is slowed down.

When the deviation from symmetry is small we can avoid the above disadvantage by considering the symmetric operator $L = (A+A^{\star})/2$. Provided that L is positive definite we can obtain the parameters m and M from $\sigma(L)$ and use them to construct the polynomial $R_n(\lambda)$. Substituting the operator A for λ we obtain the

approximation $R_n(A)$. Hence we use the polynomials $R_n(A)$ in the cyclic approach (3.2) to solve the original non-symmetric problem.

The deviation from symmetry is measured by $\| A - L \| = \| (A-A^{\star})/2 \|$. The non-symmetric matrix is generated by adding an upper triangular perturbation to the matrix generated by equation (3.12). When the deviation from symmetry is large it is essential to use the symmetrizing procedure in equation (4.1) to solve the general non-symmetric system. We solve the same non-symmetric problem using both the symmetrizing technique in (4.1) and the cyclic technique described above. These results are presented in tables 4.1 and 4.2 respectively. In table 4.3 we summarize the solution for a non-symmetric problem with a larger deviation from symmetry.

TABLE 4.1: Solution of a non-symmetric matrix vector system.

n	Error (P_n)		Error (Q_n)	
	$\| \|_{\infty}$	$\| \|_2$	$\| \|_{\infty}$	$\| \|_2$
5	0.113e-1	0.155e-1	0.776e-2	0.126e-12
10	0.741e-4	0.146e-3	0.740e-4	0.109e-3
15	0.683e-6	0.127e-6	0.749e-6	0.118e-5
20	0.588e-8	0.127e-7	0.895e-8	0.117e-7
25	0.523e-10	0.119e-9	0.837e-10	0.112e-9
30	0.626e-12	0.121e-11	0.718e-12	0.116e-11

Solution of a non-symmetric 10×10 matrix vector system using the symmetrizing procedure in equation (4.1) and the direct approach $P_n(B)f'$ and $Q_n(B)f'$. Parameters m and M are taken from $\sigma(B)$. Here $[m,M] = [0.203\ ,\ 1.069]$, $\delta = 0.39298$, $\|f'\|_{\infty} = 0.48474$ and $\|f'\|_2 = 1.0396$.

TABLE 4.2: Solution of a non-symmetric matrix vector system.

n	Ops	Error (P_n)		Error (Q_n)	
		$\|\ \|_\infty$	$\|\ \|_2$	$\|\ \|_\infty$	$\|\ \|_2$
5	5	0.169e-3	0.218e-3	0.154e-3	0.192e-3
	11	0.182e-7	0.377e-7	0.110e-7	0.247e-7
	17	0.423e-11	0.707e-11	0.294e-11	0.367e-11
	23	0.583e-15	0.137e-14	0.389e-15	0.636e-15
10	10	0.380e-7	0.833e-7	0.463e-7	0.778e-7
	21	0.255e-15	0.485e-14	0.155e-14	0.294e-14
	32	0.111e-15	0.222e-15	0.222e-15	0.312e-15
20	20	0.788e-14	0.139e-13	0.566e-14	0.114e-13
	41	0.222e-15	0.239e-15	0.111e-15	0.250e-15

Solution of the non-symmetric problem presented in table 4.1 using the cyclic technique described in equation (3.2). Ops – denote the number of matrix vector operations at the end of each iteration or cycle. Parameters m and M are taken from $\sigma(L)$. Here $[m,M\ ] = [0.448\ ,\ 1.033]$, $\delta = 0.20587$, $\|f\|_\infty = 0.71705$, $\|f\|_2 = 1.3309$ and $\|A\text{-}L\ \|_2 = 0.05627$.

TABLE 4.3: Solution of a non-symmetric matrix vector system.

n	Ops	Error (P_n)		Error (Q_n)	
		$\| \|_\infty$	$\| \|_2$	$\| \|_\infty$	$\| \|_2$
10	5	0.143	0.247	0.251e-1	0.443e-1
	21	0.258e-1	0.511e-1	0.526e-3	0.109e-2
	32	0.860e-2	0.112e-1	0.231e-4	0.307e-4
	43	0.125e-2	0.282e-2	0.357e-6	0.859e-6
20	20	0.387e-2	0.700e-2	0.572e-3	0.960e-3
	41	0.176e-4	0.322e-4	0.295e-6	0.629e-6
	62	0.837e-7	0.171e-6	0.229e-9	0.445e-9
	83	0.445e-9	0.953e-9	0.150e-12	0.318e-12

Solution a non-symmetric 10×10 matrix vector system using the cyclic technique described in equation (3.2). Parameters m and M are taken from $\sigma(L)$. Here $[m,M] = [0.55,16.29]$, $\delta = 0.68955$, $\|f\|_\infty = 9.25786$, $\|f\|_2 = 15.9675$ and $\|A\text{-}L\|_2 = 0.50642$.

Comparing tables 4.1 and 4.2, one clearly sees that convergence is very much slower when one uses the symmetrizing procedure of equation (4.1). For example, 25 operations (P_{25}) in table 4.1 yields an error that is roughly 10^5 times larger than that obtained when 23 operations are implemented in table 4.2 (4 cycles of P_5). No real advantage is gained by using cycles of very high order polynomials because of roundoff error. From table 4.3 (for larger deviation from symmetry) it is evident that the polynomials $Q_n(A)$ yield better results than the polynomials $P_n(A)$.

CONCLUSION

For slightly non-symmetric systems one should implement the polynomials $Q_n(A)$ in a cyclic fashion, where the parameters m and M are taken from the $\sigma(L)$, provided that $L = (A+A^{\star})/2$ is positive definite. The relative deviation from symmetry $\|A-L\|_2/\|L\|$ expressed as a percentage is 5.4% for the problem in table 4.2 and 3.1% for the problem in table 4.3. However, the problem in table 4.3 is ill-posed $(M/m \approx 29.62)$ as compared to the problem in table 4.2 $(M/m \approx 2.31)$ thus accounting for the large number of iterations necessary for convergence. If the system is non-symmetric the symmetrizing procedure of equation (4.1) should to be used.

CHAPTER 5

APPLICATION TO LINEAR FUNCTIONALS

Consider equation

$$Ax = f, \tag{5.1}$$

where A is an invertible linear operator in H and f is a given element of H. In some physical applications, the object of primary interest is not the solution $x = A^{-1}f$ of (5.1), but the linear functional defined by the inner product

$$\Omega = < x,g >, \tag{5.2}$$

where $g \in H$ is fixed. Such a functional is usually expressed by a suitable integral.

An approximate value for Ω can be found by evaluating an approximate solution of (5.1) and substituting into (5.2). However, such an approach was considered not effective and in the early 1970s it was suggested to use variational principles by extending the famous variational principle of Courant and Hilbert [13]. Variational approximations to integral quantities like Ω can also be obtained from functionals which have stationary properties that do not necessarily attain extremal values. Such generalized variational principles have been used extensively, for instance, in nuclear reactor physics (see Stacey [14]). Initially the effort was directed at obtaining bounds for $\Omega_f = < x,f >$.

Jensen, Smith and Wilkins ([15] – 1969) considered a positive semi-definite integral

operator A and found the complementary (upper and lower) bounds

$$\frac{<\Phi, f>^2}{<\Phi, A\Phi>} \leq \Omega_f \leq < f, E^{-1}f> - \frac{<\Phi, (AE^{-1}-I)f>^2}{<\Phi, (AE^{-1}-I)A\Phi>}. \tag{5.3}$$

Such a procedure is only possible if A can be written as the sum of two semi-definite operators of which one of them (say E) possesses an inverse. Here Φ is an arbitrary function approximating in some sense the solution.

In 1974, Barnsley and Robinson [16] obtained the bounds [see Appendix D].

$$J_f(\Phi) + \frac{1}{M} \| A\Phi - f\|^2 \leq \Omega_f \leq J_f(\Phi') + \frac{1}{m} \| A\Phi' - f\|^2, \tag{5.4}$$

for the case of positive definite A in a real Hilbert space. Here Φ and Φ' are approximations to the solution of (5.1), $J_f(\Phi) = 2 < \Phi, f > - < \Phi, A\Phi >$ [see Appendix D] and the primes denote that different pairs of trial vectors can be used in the upper and lower bounds. It is assumed that A is bounded below by $m > 0$ and above by $M > m$. Using the adjoint equation [see Appendix D]

$$Ay = g, \tag{5.5}$$

and after several manipulations they were able to obtain complementary bivariational bounds on Ω as a consequence of bounds on Ω_f. Bivariational indicates the approximation to the solution of both the original and adjoint equations. Their bounds are

$$\begin{aligned} \Omega &\leq J + (\tfrac{1}{m} - \tfrac{1}{M}) \| A\Phi - f\| \| A\Psi - g\| + \tfrac{1}{2} (\tfrac{1}{m} + \tfrac{1}{M}) C - \tfrac{1}{2} (\tfrac{1}{m} - \tfrac{1}{M}) |C|, \\ \Omega &\geq J - (\tfrac{1}{m} - \tfrac{1}{M}) \| A\Phi - f\| \| A\Psi - g\| + \tfrac{1}{2} (\tfrac{1}{m} + \tfrac{1}{M}) C + \tfrac{1}{2} (\tfrac{1}{m} - \tfrac{1}{M}) |C|, \end{aligned} \tag{5.6}$$

where $C = C(\Psi,\Phi) = < A\Phi - f, A\Psi - g >$, Ψ is an approximation to the solution of equation (5.5) and $J = J(\Psi,\Phi) = - < \Psi, A\Phi > + < \Phi, g > + < \Psi, f >$ [see Appendix D].

Cole and Pack ([17] – 1975) found families of functionals bounding Ω_f. Consider the quadratic trial functional

$$W(\Phi) = < x, f > + < \Delta x, H' \Delta x >, \tag{5.7}$$

where H' is positive or negative definite and $\Delta x = \Phi - x$ represents the error in approximating x by the trial function Φ . Substituting for Δx into (5.7) we obtain.

$$W(\Phi) = < x, (A + H')x > - < \Phi, H'(2x - \Phi) >. \tag{5.8}$$

It is necessary to choose $H' = AHA - A$ (where H is selfadjoint) in order to eliminate the unknown solution x from (5.8). Hence one obtains the expression

$$W(\Phi) = < \Phi, 2f - A\Phi > + < f - A\Phi, H(f - A\Phi) > . \tag{5.9}$$

For different choices of H in (5.9), Cole and Pack were able to reproduce the results that had been obtained previously. For $H = 0$ one obtains the lower bound $< \Phi, 2f - A\Phi >$ of Courant and Hilbert [13], provided $H' = - A$ is negative definite. If A is bounded below by m and $H = I/m$, one obtains the upper bound $J_f(\Phi) + 1/m \, \|A\Phi - f\|^2$ as derived by Barnsley and Robinson [16] in 1974.

In 1976 Barnsley and Robinson [18] sought bounds associated with unbounded non-selfadjoint operators. They made the assumption that $\exists \; m > 0$ such that $\|A\phi\|^2 \geq m^2 \, \|\phi\|^2 \;\; \forall \phi \in D(A)$.

Introducing the adjoint equation [Appendix D]

$$A^{\star}y = g\,, \tag{5.10}$$

they derived the bounds

$$J - \frac{1}{m}\,\|A\Phi - f\|\;\|A^{\star}\Psi - g\,\| \leq \Omega \leq J + \frac{1}{m}\,\|A\Phi - f\|\;\|A^{\star}\Psi - g\,\|. \tag{5.11}$$

If $A = A^{\star}$ in (5.11), the bounds obtained are weaker than those previously obtained in their 1974 paper [16].

Robinson [19] in 1978 derived bounds for $\Omega + \bar{\Omega}$, associated with the special operator $A = H + i\omega$, where H is positive definite and bounded below by $m > 0$ and ω is real. He obtained the expression

$$\begin{aligned}
&\Omega + \bar{\Omega} \leq J + \bar{J} + \frac{1}{m}\,Re < A\Phi - f, A^{\star}\Psi - g > + \frac{1}{m}\,\|\,A\Phi - f\|\;\|\,A^{\star}\Psi - g\,\|, \\
&\Omega + \bar{\Omega} \geq J + \bar{J} + \frac{1}{m}\,Re < A\Phi - f, A^{\star}\Psi - g > - \frac{1}{m}\,\|\,A\Phi - f\|\;\|\,A^{\star}\Psi - g\,\|.
\end{aligned} \tag{5.12}$$

If $\omega = 0$ then $A = H$ is selfadjoint and the bounds obtained from (5.12) are tighter than those obtained by Barnsley and Robinson in 1974 [16] and Cole and Pack in 1975 [17] for selfadjoint operators in real Hilbert spaces.

In 1979, Barnsley and Robinson [20] derived the bounds

$$\begin{aligned}
&\Omega \leq J + \tfrac{1}{2}\,(\frac{1}{m} + \frac{1}{M}) < A\Phi - f, A\Psi - g > + \tfrac{1}{2}\,(\frac{1}{m} - \frac{1}{M})\;\|A\Phi - f\|\;\|A\Psi - g\,\|, \\
&\Omega \geq J + \tfrac{1}{2}\,(\frac{1}{m} + \frac{1}{M}) < A\Phi - f, A\Psi - g > - \tfrac{1}{2}\,(\frac{1}{m} - \frac{1}{M})\;\|A\Phi - f\|\;\|A\Psi - g\,\|,
\end{aligned} \tag{5.13}$$

where A is selfadjoint and bounded below by $m > 0$ and above by $M > m$.

In 1985, Mika, Cole and Pack [21] were the first to show that the bounds to Ω are linked to the approximation of the inverse operator A^{-1}. For positive definite operators, they showed that the best approximation of zero order is $1/2\,(1/m + 1/M)\,I$ with associated error $1/2\,(1/m - 1/M)$. Using this approximation to A^{-1} they obtained the bounds of Barnsley and Robinson [eqn (5.13)]. They also derived bounds for Ω when the operator A was non-selfadjoint. This involves symmetrizing equation (5.1) by multiplying by the adjoint $A^{\star}$ and making the assumption that $m^2 \|\phi\|^2 \leq \|A\phi\|^2 \leq M^2 \|\phi\|^2 \;\; \forall \phi \in H$, where $M > m > 0$. Hence they derived the bounds

$$\Omega \leq J + \frac{1}{2}\left[\frac{1}{m^2} + \frac{1}{M^2}\right] < g\text{-}A^{\star}\Psi, A^{\star}f\text{-}A^{\star}A\Phi > + \frac{1}{2}\left[\frac{1}{m^2} - \frac{1}{M^2}\right] \|g\text{-}A^{\star}\Psi\| \; \|A^{\star}f\text{-}A^{\star}A\Phi\|\,,$$

$$\Omega \geq J + \frac{1}{2}\left[\frac{1}{m^2} + \frac{1}{M^2}\right] < g\text{-}A^{\star}\Psi, A^{\star}f\text{-}A^{\star}A\Phi > - \frac{1}{2}\left[\frac{1}{m^2} - \frac{1}{M^2}\right] \|g\text{-}A^{\star}\Psi\| \; \|A^{\star}f\text{-}A^{\star}A\Phi\|\,, \tag{5.14}$$

as a consequence of bounds involving selfadjoint operators.

In 1986, Robinson and Yuen [22] derived bounds associated with the special operator $A = I + \lambda K$, where K is an integral operator. They introduced the selfadjoint operator $L = 1/2\,(A + A^{\star})$ and assumed that it was positive definite. They obtained the bounds

$$\begin{aligned}\Omega \leq\; & J + \tfrac{1}{2} <A\Phi - f, L^{-1}(A^{\star}\Psi - g)> \\ & + \tfrac{1}{2} <A\Phi - f, L^{-1}(A\Phi - f)>^{1/2} <A^{\star}\Psi - g, L^{-1}(A^{\star}\Psi - g)>^{1/2},\end{aligned}$$

$$\begin{aligned}\Omega \geq\; & J + \tfrac{1}{2} <A\Phi - f, L^{-1}(A^{\star}\Psi - g)> \\ & - \tfrac{1}{2} <A\Phi - f, L^{-1}(A\Phi - f)>^{1/2} <A^{\star}\Psi - g, L^{-1}(A^{\star}\Psi - g)>^{1/2}\end{aligned} \tag{5.15}$$

Again we see the appearance of the inverse operator. If L^{-1} is unknown and L is bounded below by $m > 0$, they obtained the bounds

$$\Omega \leq J + \frac{1}{2m} < A\Phi - f, A^{\star}\Psi - g > + \frac{1}{2m} \| A\Phi - f \| \, \| A^{\star}\Psi - g \|,$$

$$\Omega \geq J + \frac{1}{2m} < A\Phi - f, A^{\star}\Psi - g > - \frac{1}{2m} \| A\Phi - f \| \, \| A^{\star}\Psi - g \|. \tag{5.16}$$

It was only in 1987 that Mika [23] realized that all the bounds derived previously were special cases of the general result [see Appendix D].

$$\Omega \leq J + < V(f - A\Phi), g - A^{\star}\Psi > + \| A^{-1} - V \| \, \| g - A^{\star}\Psi \| \, \| f - A\Phi \|,$$

$$\Omega \geq J + < V(f - A\Phi), g - A^{\star}\Psi > - \| A^{-1} - V \| \, \| g - A^{\star}\Psi \| \, \| f - A\Phi \|. \tag{5.17}$$

Here V is an approximation to the inverse operator A^{-1}. The third term represents the error which is cubic as compared to earlier work where the error is quadratic, because of the presence of the factor $\| A^{-1} - V \|$. Optimization of the bounds in (5.17) involves minimization of the residuals $\| f - A\Phi \|$ and $\| g - A^{\star}\Psi \|$ as well as the optimal approximation of A^{-1} by V. Comparing (5.17) with the bounds obtained previously, one can identify the different approximations to the inverse operator. These approximations, all of zero order, are summarized in table 5.1 .

Computation of the approximate solutions Φ and Ψ in (5.17) by minimum residual techniques can be expensive. Although the zero order approximations $R_0(A)f$ and $R_0(A)g$ for Φ and Ψ respectively require no effort to compute, one still has to perform two operations in evaluating the residuals $f - A\Phi$ and $g - A^{\star}\Psi$. It seems best to set $\Phi = \Psi = 0$ and concentrate on the approximation of the inverse operator.

TABLE 5.1: Earlier zero order approximations to A^{-1}.

Author	Equation no.	$V \approx A^{-1}$	Error
Barnsley & Robinson (1974)	(5.6)	I/m and I/M	$1/m - 1/M$
Barnsley & Robinson (1976)	(5.11)	0	$1/m$
Robinson(1978)	(5.12) $\begin{bmatrix}\omega = 0 \\ \text{real } H\end{bmatrix}$	$I/2m$	$1/2m$
Barnsley & Robinson (1979)	(5.13)	$1/2\,(1/m+1/M)I$	$1/2\,(1/m-1/M)$
Robinson & Yuen (1986)	(5.15) $\begin{bmatrix}L^{-1} = I \\ K+K^{\star} = 0\end{bmatrix}$	$I/2$	$1/2$
	(5.16)	$I/2m$	$1/2m$

Different zero order approximations to the inverse operator A^{-1} deduced by comparing (5.17) with the bounds obtained previously.

The operator V can be taken as one of the polynomials $R_n(A)$ discussed in chapters 2 and 3. If say $\Psi = 0$, then from (5.17) one obtains the approximation

$$\Omega \approx < \Phi, g > + < V(f - A\Phi), g > , \qquad (5.18)$$

where we have ignored the error term. If further $\Phi = R_k(A)f$ and $V = R_n(A)$, then

$$\Omega \approx < (R_k(A) + R_n(A) - A\, R_k(A)\, R_n(A))f, g >. \tag{5.19}$$

Now $R_k(A) + R_n(A) - A\, R_k(A)\, R_n(A)$ is a polynomial of degree $n + k + 1$ approximating the inverse operator A^{-1}, but the best such polynomial is obviously $R_{k+n+1}(A)$. Hence

$$< R_{k+n+1}(A)f,g > \tag{5.20}$$

is a better approximation to the linear functional Ω involving the same number of operations. We present some simple results in table 5.2 confirming the above.

From (5.17) with $\Phi = \Psi = 0$ and ignoring the error term one obtains the approximation $\Omega = < x,g > \approx < Vf,g > = < R_n(A)f,g >$, which essentially involves finding the approximate solution $R_n(A)f$ first and then computing the functional.

TABLE 5.2: Calculation of a linear functional.

Ops	(n,k)	Error using eq (5.19)	Error using eq (5.20)
1	(0,0)	0.148e+1	0.208
3	(1,1)	0.171e-1	-0.731e-2
4	(2,1)	-0.147e-2	0.355e-3
5	(2,2)	0.159e-2	0.635e-3

Calculation of a functional using equations (5.19) and (5.20). The system is a simple 10×10 matrix vector system. The functional is the dot product having an exact value of 13.415. Here $[m,M\,] = [1,2]$ and $\delta = 0.17157$. Here we have used the polynomials $P_n(A)$.

CONCLUSION

Variational techniques were used earlier in order to save on the computation of the true solution, since such computation was expensive. Since the calculation of the true solution by the method of approximate inverse is simple (iterative) and less expensive, it seems to dispense with the need for variational methods in this context.

CHAPTER 6

APPLICATION TO FREDHOLM INTEGRAL EQUATIONS

Consider the Fredholm integral equation of the second kind

$$\phi(x) + \lambda \int_a^b K(x,y)\ \phi(y)\ dy = f(x),\ a \leq x \leq b\ . \tag{6.1}$$

Here $\phi(x)$ is the unknown solution, λ a real constant and $K(x,y)$ represents the kernel.

Equation (6.1) can be cast in operator form

$$(A\ \phi)(x) = f(x)\ , \tag{6.2}$$

where $A = I + \lambda\ K$.

We now discuss some of the methods available for solving (6.1). The method of expansion of solution (see [24]) involves assuming a special form for the solution $\phi(x)$ of (6.1). If we write

$$\phi(x) \approx \phi_N(x) = \sum_{k=0}^{N} c_k^{(N)}\ h_k(x)\ , \tag{6.3}$$

where the set $\{h_k(x)\}$ is complete in $\mathcal{L}^2(a,b)$, then with an appropriate choice of the $c_k^{(N)}$, and for sufficiently large N, we may approximate $\phi(x)$ as closely as we please by $\phi_N(x)$. The problem is then one of determining the coefficients $c_k^{(N)}$. This

may involve a least squares approximation where the $c_k^{(N)}$ are chosen such that $\| (A\phi_N)(x) - f(x) \|_2$ is minimized. However this leads to the evaluation of N^2 triple integrals which is expensive for large N.

Piessens and Branders [25] used the expansion method of the form (6.3) where they took $h_k(x) = T_k(x)$ – the Chebyshev polynomials of order k. Firstly the interval $[a,b]$ is mapped by a linear transformation to the interval $[-1,1]$ because of the orthogonality of the Chebyshev polynomials on $[-1,1]$ with weight function $(1 - x^2)^{-1/2}$. With $\phi(x) = \Sigma_0^N c_k^{(N)} T_k(x)$ in (6.1) we obtain

$$\sum_{k=0}^{N} c_k^{(N)} [\, T_k(x) + \lambda\, I_k(x) \,] = f(x)\,, \tag{6.4}$$

where

$$I_k(x) = \int_{-1}^{1} K(x,y)\; T_k(y)\; dy \tag{6.5}$$

Substituting $N+1$ values of x $(x_k,\ k = 0,1,\cdots N)$ results in a linear system, the solution of which gives the coefficients $c_k^{(N)}$. The values x_k are chosen equidistantly between -1 and 1. The evaluation of $I_k(x_j)$ $(k,j = 0,1,\cdots N)$ is done by means of a three point recurrence relation for $I_k(x)$. This recurrence relation saves on computational effort. However it must be stressed that the derivation of a recurrence is only possible for special kernels. For arbitrary kernels it is essential to use numerical integration to evaluate $I_k(x_j)$ and this is very expensive.

The Nystrom or quadrature method (see [24]) entails approximating the integral in (6.1) by a N-point quadrature rule.

Hence

$$\int_a^b K(x,y)\ \phi(y)\ dy = \sum_{k=1}^{N} \omega_k\, K(x,y_k)\ \phi(y_k)\ , \tag{6.6}$$

where ω_k represents the relevant weight functions. Letting $x = y_i$ in (6.1) and using (6.6) one obtains the discretization

$$\phi(y_i) + \lambda \sum_{k=1}^{N} \omega_k\, K(y_i,y_k)\ \phi(y_k) = f(y_i)\ ,\ i = 1,2,\cdots N\,. \tag{6.7}$$

Solving the linear system in (6.7) by Gaussian elimination involves $O(N^3)$ multiplications, whilst the effort using (if possible) the Neumann series (Picard iteration) is of the order $O(N^2)$. Thus for N large Gaussian elimination can lead to surprisingly large computing times. This simple comparison shows that iterative schemes are worth investigating. We shall now investigate some.

If A is positive definite the solution $\phi(x) = A^{-1}f(x)$ of (6.2) can be obtained by using the polynomial methods $P_n(A)$ or $Q_n(A)$ discussed in chapter 2 to approximate the inverse operator A^{-1}. If the polynomials $P_n(A)$ are used for example, one needs to evaluate

$$P_0 f(x) = \frac{1}{2}\left[\frac{1}{m} + \frac{1}{M}\right] f(x)\ , \tag{6.8}$$

$$P_1 f(x) = \frac{(\sqrt{M}+\sqrt{m})^2}{2\,mM}\ f(x) - \frac{1}{mM}\ (I+\lambda K)\ f(x)\,, \tag{6.9}$$

and generate higher order approximations $P_n f(x)$ using the recurrence relation (3.16). This method is usually possible up to some low order due to difficulty in evaluating the integrals $Kf(x)$, $K^2 f(x)$, $K^3 f(x)$ $\cdots$, whether this evaluation is done analytically (in special cases) or numerically. The preceding technique is quite useful for obtaining pointwise solutions.

POINTWISE SOLUTIONS

To calculate the solution $\phi(x)$ of (6.1) at some fixed point $x = x^\star$, we first calculate $R_0 f(x^\star)$ and $R_1 f(x^\star)$ and then generate higher order approximations $R_n f(x^\star)$ using the recurrence relations (2.22) or (3.16). Here $R_n f(x^\star)$ denotes $R_n f(x)$ evaluated at $x^\star$. Since $A = I + \lambda K$, this involves the calculation of the integrals $Kf(x^\star)$, $K^2 f(x^\star)$, $K^3 f(x^\star), \cdots$, where as before $K^n f(x^\star)$ denotes $K^n f(x)$ evaluated at $x^\star$. These integrals are calculated analytically where possible or numerically by using an appropriate quadrature rule. We illustrate the above by solving the following problem (see [20]).

I1 $K(x,y) = \begin{cases} x(1-y) & \text{for } x \leq y \\ y(1-x) & \text{for } x \geq y \end{cases}$

$\lambda = 1.0$

$f(x) = x^2$

$a = 0,\ b = 1.0$

with exact pointwise solution

$\phi(0.5) = 2.5\ sech(0.5) - 2 = 0.217047209925.$

Results

The results, summarized in table 6.1, are quite impressive using low polynomial orders n. This is because the convergence parameter $\delta = 0.02412$ is close to zero. The polynomial

TABLE 6.1: Pointwise solution of an integral equation.

n	Error (P_n) $\phi(0.5)$-$P_n f(0.5)$	Error (Q_n) $\phi(0.5)$-$Q_n f(0.5)$
0	-0.215e-1	-0.209e-1
1	0.429e-3	0.404e-3
2	-0.719e-5	-0.670e-5
3	0.351e-6	0.343e-6
4	-0.121e-7	-0.744e-8

Solution of problem I1 using the procedures $P_n(A)f$ and $Q_n(A)f$. Here $[m\ M] = [1\ ,\ 1.10132]$, $\delta = 0.02412$, $Kf(x) = (x\text{-}x^4)/12$ and $K^2f(x) = (4x\text{-}5x^3\text{+}x^6)/360$. A 20 interval Simpson quadrature is used to evaluate $K^3f(0.5)$ and $K^4f(0.5)$.

procedure $Q_n(A)f$ gives better results than the procedure $P_n(A)f$. Barnsley and Robinson [20] solved this problem by calculating upper and lower bounds to $\phi(0.5)$. They obtained the bounds $0.217047 \leq \phi(0.5) \leq 0.217048$ from which $\phi(0.5) \approx 0.2170475$. This represents an absolute error of approximately 0.3×10^{-6} which is comparable with our result for $n = 3$. While Barnsley and Robinson [20] used much numerical effort to obtain their bounds their method did not allow for an improvement of the solution. We, however, can improve the solution by choosing n larger. Thus for $n = 4$ (Q_4) we obtain the absolute error 0.744×10^{-8}.

If one requires the solution $\phi(x)$ at few abscissae then the method just implemented is advisable. However, if the behaviour of the solution $\phi(x)$ in the interval $[a,b]$ is required then this would necessitate the calculation of a large number of function values

of $\phi(x)$. In such a case the procedure of pointwise solutions is not recommended since it would require the calculation of a large number of integrals numerically which is a very expensive process. The alternative approach is a full discretization of (6.1).

SOLUTION BY DISCRETIZATION

Smooth Kernels

We solve the following problems with a symmetric kernel.

I2 $K(x,y) = \mid x - y \mid$

$\lambda = 1.0$

$f(x) = 1 + x - sin(x)$

$a = 0,\ b = \pi/2$

with exact solution $\phi(x) = sin(x)$

I3 $K(x,y) = \mid x - y \mid$

$\lambda = 1.0$

$f(x) = 0.1x^5 + x^3 - 0.25x + 0.2$

$a = 0,\ b = 1.0$

with exact solution $\phi(x) = x^3$

For the above we use the Simpson quadrature technique to approximate the integral.

This leads to the discretization

$$\phi(x) + \lambda h/3 \left[K(x,a) + K(x,b) + 2 \sum_{p=1}^{n-1} K(x,a+2ph)\ \phi(a+2ph) \right.$$

$$\left. + 4 \sum_{p=1}^{n} K(x,a+(2p-1)h)\ \phi(a+(2p-1)h) \right] = f(x)\ , \qquad (6.10)$$

with $x = a + qh$, $q = 0, 1, 2, \cdots 2n$, $h = (b-a)/2n$, where $2n = N$ represents the number of discretization intervals.

Weakly Singular Kernels

We solve the following problems:

I4 $K(x,y) = |\ x - y\ |^{-\frac{1}{2}}$

$\lambda = 1.0$

$f(x) = 2x^2 \left[(1+x)^{\frac{1}{2}} + (1-x)^{\frac{1}{2}} \right] + 4x/3 \left[(1-x)^{\frac{3}{2}} - (1+x)^{\frac{3}{2}} \right]$

$+ 0.4 \left[(1+x)^{\frac{5}{2}} + (1-x)^{\frac{5}{2}} \right] + x^2$

$a = -1.0,\ b = 1.0$

with exact solution $\phi(x) = x^2$

I5 $K(x,y) = |\ x - y\ |^{-\frac{1}{2}}$

$\lambda = 0.5$

$f(x) = x^2$

$a = -1.0,\ b = 1.0$

where the exact solution is unknown.

Here we evaluate the integral using the technique by Atkinson [3] which involves an adaptation of the trapezoidal rule which is summarized below.

Firstly, the interval (a,b) is divided into N equally spaced subintervals and the integral is evaluated over each subinterval.

Hence

$$\int_a^b K(x,y)\ \phi(y)\ dy = \sum_{k=1}^{N} \int_{y_k}^{y_{k+1}} K(x,y)\ \phi(y)\ dy\ , \tag{6.11}$$

where $y_k = a + (k-1)h$, $k = 1, 2, \cdots N+1$ and $h = (b-a)/N$ is the length of each subinterval. On each subinterval (y_k, y_{k+1}) we replace $\phi(y)$ by the linear Lagrange polynomial

$$1/h\ [\ (y-y_k)\ \phi(y_{k+1}) - (y-y_{k+1})\ \phi(y_k)\ ]\ . \tag{6.12}$$

Substituting (6.12) into (6.11) and letting $x = y_i$ $(i = 1,2,\cdots N+1)$, we obtain

$$\int_a^b K(y_i,y)\ \phi(y)\ dy = \sum_{k=1}^{N+1} \omega_{ik}\ \phi(y_k)\ , \tag{6.13}$$

where the weights ω_{ik} are given by

$$\omega_{i1} = -1/h \int_{y_1}^{y_2} (y-y_2)\ K(y_i,y)\ dy$$

$$\omega_{ik} = 1/h \int_{y_{k-1}}^{y_k} (y-y_{k-1})\ K(y_i,y)\ dy - 1/h \int_{y_k}^{y_{k+1}} (y-y_{k+1})\ K(y_i,y)\ dy,\ k = 2,3\cdots N$$

$$\omega_{iN+1} = 1/h \int_{y_N}^{y_{N+1}} (y-y_N)\ K(y_i,y)\ dy.$$

Hence, with $x = y_i$ in equation (6.1) and using (6.13), one obtains the discretization

$$\phi(y_i) + \lambda \sum_{k=1}^{N+1} \omega_{ik}\, \phi(y_k) = f(y_i),\ i = 1,2, \cdots N+1\ . \qquad (6.14)$$

We note that for the kernel $K(x,y) = \mid x - y \mid^{-\frac{1}{2}}$, the weights can be calculated analytically.

Both discretization procedures mentioned above lead to slightly non-symmetric systems, which are solved by using the cyclic technique discussed in chapter four for similar systems.

Results

The results are quoted for polynomials of order ten, namely P_{10} and Q_{10} and are summarized in tables 6.2 to 6.14 . Note that we present actual errors. The results for smooth and weakly singular kernels yield average errors of the order of magnitude 10^{-3} for $N = 70$. We notice that the errors do not damp after the second iteration or cycle which clearly shows that the final error is essentially the discretization error. To improve the results we would have to increase N . This improvement is clearly seen in tables 6.5 and 6.12 . From the captions to tables 6.2, 6.3 and 6.4 we observe that the deviation from symmetry $\|A - L\|_2$ decreases with increasing N . In fact this pattern is maintained throughout our results. Thus if our solution converges for small N then we are guaranteed convergence for larger N because the discretized system tends more to a symmetric one with increasing N. We also note that using the polynomials $P_n(A)$ and $Q_n(A)$ give equally good results. Our results for problem I5 (tables 6.12, 6.13 and 6.14)

agree remarkably well with that as quoted in the paper by Piessens and Branders [25], except at $x = 0.99313$. However at this particular abscissa our result lies between that calculated by various other authors and the revised solution of Piessens and Branders [25]. The simple approach of full discretization of the integral equation (6.1) together with the solution of the resulting linear algebraic system by the iterative methods we have discussed, yields quite acceptable results.

CONCLUSION

In cases where the solution $\phi(x)$ of (6.1) is required at few selected points (abscissae), a solution by discretization of (6.1) is unnecessary. The polynomial procedure $Q_n(A)f(x)$ provides a simple yet effective method of evaluating such pointwise solutions. If the behaviour of $\phi(x)$ on the interval $[a,b]$ is required then a full discretization of (6.1) is unavoidable.

The use of Gauss quadrature techniques in place of the Simpson technique in connection with smooth kernels, would yield better results at the expense of more effort. Using the midpoint rule instead of Simpson's rule for symmetric kernels would yield symmetric systems, however this would require large discretizations to achieve the same error. For weakly singular kernels, approximating $\phi(y)$ on each subinterval by a higher order Lagrange polynomial is possible, thus reducing discretization error. However, the calculation of the weights are far more complicated and expensive.

We have presented here a brief motivation for the use of polynomial methods to solve Fredholm integral equations of the second kind numerically. The subject of integral equations is vast and requires special attention. A full analysis of numerical methods used for solving integral equations is beyond the scope of this thesis.

TABLE 6.2: Solution of an integral equation.

	x	Error(P_n)	Error(Q_n)
cycle 1	0	0.484e-2	0.482e-2
	0.942478	0.374e-3	0.362e-3
	1.256637	-0.243e-3	-0.258e-3
	1.570796	0.214e-3	0.193e-3
$\|\ \|_\infty$		0.847e-2	0.849e-2
$\|\ \|_2$		0.142e-1	0.142e-1
cycle 2	0	0.482e-2	0.482e-2
	0.942478	0.364e-3	0.362e-3
	1.256637	-0.250e-3	-0.250e-3
	1.570796	0.206e-3	0.206e-3
$\|\ \|_\infty$		0.847e-2	0.847e-2
$\|\ \|_2$		0.142e-1	0.142e-1

Solution of problem I2. Here $[m,M\ ] = [0.46,\ 1.91]$ $\delta = 0.34160$, $N = 10$ and $\|A - L\|_2 = 0.23252$.

TABLE 6.3: Solution of an integral equation.

	x	Error(P_n)	Error(Q_n)
cycle 1	0	0.122e-2	0.120e-2
	0.942478	0.937e-4	0.804e-4
	1.256637	-0.656e-4	-0.814e-4
	1.570796	0.472e-4	0.249e-4
$\|\ \|_\infty$		0.208e-2	0.209e-2
$\|\ \|_2$		0.490e-2	0.492e-2
cycle 2	0	0.120e-2	0.120e-2
	0.942478	0.814e-4	0.814e-4
	1.256637	-0.747e-4	-0.747e-4
	1.570796	0.357e-4	0.357e-4
$\|\ \|_\infty$		0.208e-2	0.208e-2
$\|\ \|_2$		0.491e-2	0.491e-2

Solution of problem I2. Here $[m,M\ ] = [0.46,\ 1.91]$ $\delta = 0.34160$, $N = 20$ and $\|A\ \text{-}\ L\ \|_2 = 0.19256$.

TABLE 6.4: Solution of an integral equation.

	x	Error(P_n)	Error(Q_n)
cycle 1	0	0.117e-3	0.988e-4
	0.942478	0.194e-4	0.605e-5
	1.256637	-0.356e-5	-0.124e-4
	1.570796	0.152e-4	-0.731e-5
$\|\ \|_\infty$		0.161e-3	0.179e-3
$\|\ \|_2$		0.730e-3	0.760e-3
cycle 2	0	0.980e-4	0.980e-4
	0.942478	0.641e-5	0.641e-5
	1.256637	-0.640e-5	-0.640e-5
	1.570796	0.253e-5	0.252e-5
$\|\ \|_\infty$		0.171e-3	0.171e-3
$\|\ \|_2$		0.739e-3	0.739e-3

Solution of problem I2. Here $[m,M\ ] = [0.46,\ 1.91]$ $\delta = 0.34160$, $N = 70$ and $\|A\ -\ L\ \|_2 = 0.15848$.

TABLE 6.5: Discretization error.

x	$N = 10$	$N = 20$	$N = 40$	$N = 70$
0	0.482e-2	0.120e-2	0.300e-3	0.980e-4
0.314159	0.347e-2	0.862e-3	0.215e-3	0.702e-4
0.942478	0.364e-3	0.814e-4	0.198e-4	0.641e-5
1.256637	-0.250e-3	-0.748e-4	-0.194e-4	-0.640e-5
1.570796	0.206e-3	0.357e-4	0.795e-5	0.252E-5
$\|\ \|_\infty$	0.847e-2	0.208e-2	0.522e-3	0.171e-3
$\|\ \|_2$	0.142e-1	0.491e-2	0.172e-2	0.739e-3

Solution of Problem I2, illustrating the effect of discretization error at selected abscissae. The discretized system is solved exactly using several cycles.

TABLE 6.6: Solution of an integral equation.

	x	Error(P_n)	Error(Q_n)
cycle 1	0	0.405e-3	0.405e-3
	0.4	0.187e-3	0.187e-3
	0.6	0.638e-4	0.638e-4
	1.0	-0.173e-4	-0.173e-4
$\|\ \|_\infty$		0.244e-2	0.244e-2
$\|\ \|_2$		0.277e-2	0.277e-2
cycle 2	0	0.405e-3	0.405e-3
	0.4	0.187e-3	0.187e-3
	0.6	0.638e-4	0.638e-4
	1.0	-0.173e-4	-0.173e-4
$\|\ \|_\infty$		0.244e-2	0.244e-2
$\|\ \|_2$		0.277e-2	0.277e-2

Solution of problem I3. Here $[m,M\ ] = [0.78,\ 1.37]$ $\delta = 0.13989$, $N = 10$ and $\|A\ -\ L\ \|_2 = 0.09424$.

TABLE 6.7: Solution of an integral equation.

	x	Error(P_n)	Error(Q_n)
cycle 1	0	0.102e-3	0.102e-3
	0.4	0.483e-4	0.483e-4
	0.6	0.178e-4	0.178e-4
	1.0	-0.223e-5	-0.223e-5
$\|\ \|_\infty$		0.719e-3	0.719e-3
$\|\ \|_2$		0.101e-2	0.101e-2
cycle 2	0	0.102e-3	0.102e-3
	0.4	0.483e-4	0.483e-4
	0.6	0.178e-4	0.178e-4
	1.0	-0.223e-5	-0.223e-5
$\|\ \|_\infty$		0.719e-3	0.719e-3
$\|\ \|_2$		0.101e-2	0.101e-2

Solution of problem I3. Here $[m,M\ ] = [0.78,\ 1.37]$ $\delta = 0.13989$, $N = 20$ and $\|A\ -\ L\ \|_2 = 0.07804$.

TABLE 6.8: Solution of an integral equation.

	x	Error(P_n)	Error(Q_n)
cycle 1	0	0.256e-4	0.256e-4
	0.4	0.122e-4	0.122e-4
	0.6	0.457e-5	0.457e-5
	1.0	-0.433e-6	-0.433e-6
$\|\ \|_\infty$		0.194e-3	0.194e-3
$\|\ \|_2$		0.359e-3	0.359e-3
cycle 2	0	0.256e-4	0.256e-4
	0.4	0.122e-4	0.122e-4
	0.6	0.457e-5	0.457e-5
	1.0	-0.433e-6	-0.433e-6
$\|\ \|_\infty$		0.194e-3	0.194e-3
$\|\ \|_2$		0.359e-3	0.359e-3

Solution of problem I3. Here $[m,M\ ] = [0.78,\ 1.37]$ $\delta = 0.13989$, $N = 40$ and $\|A\ -\ L\ \|_2 = 0.06864$.

TABLE 6.9: Solution of an integral equation.

	x	Error(P_n)	Error(Q_n)
cycle 1	-1.0	0.112e-2	0.114e-2
	-0.4	0.125e-2	0.126e-2
	0.4	0.125e-2	0.126e-2
	1.0	0.112e-2	0.114e-2
$\|\ \|_\infty$		0.126e-2	0.127e-2
$\|\ \|_2$		0.560e-2	0.562e-2
cycle 2	-1.0	0.110e-2	0.110e-2
	-0.4	0.125e-2	0.125e-2
	0.4	0.125e-2	0.125e-2
	1.0	0.110e-2	0.110e-2
$\|\ \|_\infty$		0.127e-2	0.127e-2
$\|\ \|_2$		0.560e-2	0.560e-2

Solution of problem I4. Here $[m,M\ ] = [1.4,\ 4.8]$ $\delta = 0.29865$, $N = 20$ and $\|A - L\|_2 = 0.18492$.

TABLE 6.10: Solution of an integral equation.

	x	Error(P_n)	Error(Q_n)
cycle 1	-1.0	0.354e-3	0.380e-3
	-0.4	0.319e-3	0.323e-3
	0.4	0.318e-3	0.323e-3
	1.0	0.354e-3	0.380e-3
$\|\ \|_\infty$		0.354e-3	0.380e-3
$\|\ \|_2$		0.201e-2	0.204e-2
cycle 2	-1.0	0.274e-3	0.274e-3
	-0.4	0.320e-3	0.320e-3
	0.4	0.320e-3	0.320e-3
	1.0	0.274e-3	0.274e-3
$\|\ \|_\infty$		0.323e-3	0.323e-3
$\|\ \|_2$		0.200e-2	0.200e-2

Solution of problem I4. Here $[m,M\ ] = [1.4,\ 4.8]$ $\delta = 0.29865$, $N = 40$ and $\|A - L\|_2 = 0.14146$.

TABLE 6.11: Solution of an integral equation.

	x	Error(P_n)	Error(Q_n)
cycle 1	-1.0	0.222e-3	0.258e-3
	-0.4	0.104e-3	0.109e-3
	0.4	0.104e-3	0.109e-3
	1.0	0.222e-3	0.258e-3
$\|\ \|_\infty$		0.222e-3	0.258e-3
$\|\ \|_2$		0.897e-3	0.945e-3
cycle 2	-1.0	0.892e-4	0.892e-4
	-0.4	0.106e-3	0.106e-3
	0.4	0.106e-3	0.106e-3
	1.0	0.891e-4	0.892e-4
$\|\ \|_\infty$		0.106e-3	0.106e-3
$\|\ \|_2$		0.870e-3	0.870e-3

Solution of problem I4. Here $[m,M\ ] = [1.4,\ 4.8]$ $\delta = 0.29865$, $N = 70$ and $\|A$ - $L\|_2 = 0.11264$.

TABLE 6.12: Discretization error.

x	$N = 10$	$N = 20$	$N = 40$	$N = 70$
-1.0	0.438e-2	0.110e-2	0.274e-3	0.892e-4
-0.4	0.487e-2	0.125e-2	0.320e-3	0.106e-3
0	0.492e-2	0.127e-2	0.323e-3	0.106e-3
0.4	-0.487e-2	0.125e-2	0.320e-3	0.106e-3
1.0	0.438e-2	0.110e-2	0.274e-3	0.892e-4
$\|\ \|_\infty$	0.492e-2	0.127e-2	0.323e-3	0.106e-3
$\|\ \|_2$	0.157e-1	0.560e-2	0.200e-2	0.870e-3

Solution of Problem I4, illustrating the effect of discretization error at selected abscissae. The discretized system is solved exactly using several cycles.

TABLE 6.13: Solution of an integral equation.

x	$N = 20$	$N = 30$	$N = 50$	$N = 70$
0.99313	0.6630540	0.6626137	0.6598361	0.6567397
0.96397	0.5901404	0.5761526	0.5520521	0.5483424
0.74633	0.2513220	0.2507034	0.2510444	0.2510025
0.51087	0.0662370	0.0670994	0.0671376	0.0671728
0.07653	-0.0784614	-0.0784229	-0.0782376	-0.0780721

Solution of problem I5, showing values of the solution $\phi(x)$ at selected abscissae. The solution is first obtained at evenly spaced points. Then the values at the above abscissae are calculated using the fact that $\phi(x)$ is assumed a linear function on each subinterval. Here $[m,M\ ] = [1.14,\ 2.9]$ and $\delta = 0.22927$.

TABLE 6.14: Solution of an integral equation.

x	Ullman	Schlitt	Cohen – Ickovic	P & B	P & B Revised
0.99313	0.63303	0.63130	0.62997	0.62856	0.701180
0.96397	0.55114	0.54817	0.54650	0.55111	0.545315
0.74633	0.25311	0.25121	0.24995	0.25121	0.250216
0.51087	0.06767	0.06737	0.06631	0.06733	0.066658
0.07653	-0.07907	-0.07790	-0.07882	-0.07799	-0.078524

Solution of problem I5, by various authors, as quoted in Piessens and Branders (P & B) [25]. The last column shows a revised solution by Piessens and Branders [25].

APPENDIX A

The results in this thesis, pertaining to the polynomials $P_n(A)$, are quoted in Bond and Mika [1]. Here we show in detail how these results are derived.

DERIVATION OF $\tilde{r}_n(t)$

The polynomial $\tilde{r}_n(t)$ of best approximation to $1/(u-t)$ for $-1 \leq t \leq 1$ and $u > 1$, has been found by Chebyshev. We state the results from Meinardus [5] and use it to show how $\tilde{r}_n(t)$ can be written explicitly in terms of the Chebyshev polynomials. Following [5] we define a function

$$\phi_n(t) = \tilde{r}_n(t) - \frac{1}{u-t} = \frac{\Gamma}{2}\left[v^n \frac{\delta - v}{1-\delta v} + v^{-n} \frac{1-\delta v}{\delta - v} \right], \qquad \text{(A1)}$$

where $t = 1/2\ (v + v^{-1})$, $|v| = 1$, $u = 1/2\ (\delta + \delta^{-1})$ or $\delta = u - \sqrt{u^2 - 1} < 1$ and Γ is the optimal error of approximation given by

$$\Gamma = \left[\frac{2\delta}{1-\delta^2} \right]^2 \delta^n = \frac{\delta^n}{u^2 - 1}. \qquad \text{(A2)}$$

With $v = e^{i\theta}$ (since $|v| = 1$), we have $t = cos\ \theta = Re(v)$ for $0 \leq \theta \leq \pi$. Hence as t runs through the interval [–1,1] from left to right, v describes the upper half of the circle $|v| = 1$.

That Γ describes the optimal error can easily be seen if one rewrites $\phi_n(t)$ in trigonometric form. Since

$$\left| \frac{\delta - v}{1-\delta v} \right| = 1,$$

we let

$$\frac{\delta - v}{1 - \delta v} = e^{i\eta}. \tag{A3}$$

Then

$$\begin{aligned} \phi_n(t) &= \frac{\Gamma}{2}\left[\, e^{in\theta}\, e^{i\eta} + e^{-in\theta}\, e^{-i\eta}\,\right] \\ &= \Gamma\, cos(n\theta + \eta) \\ &= \frac{\delta^n}{u^2 - 1}\, \Psi_n(t), \end{aligned} \tag{A4}$$

where

$$\Psi_n(t) = cos(n\theta + \eta)\ . \tag{A5}$$

From (A1) and (A2) we have

$$\phi_n(t) = \frac{2\delta^{n+2}}{(1-\delta^2)^2}\left[\, v^n \frac{\delta - v}{1 - \delta v} + v^{-n}\frac{1 - \delta v}{\delta - v}\,\right]. \tag{A6}$$

Let

$$\psi_n(v) = v^n \frac{\delta - v}{1 - \delta v},$$

then using $\bar{v} = e^{-i\theta} = v^{-1}$, we obtain

$$v^{-n}\frac{1 - \delta v}{\delta - v} = v^{-n}\frac{v^{-1} - \delta}{\delta v^{-1} - 1} = \bar{v}^n \frac{\delta - \bar{v}}{1 - \delta\bar{v}} = \psi_n(\bar{v}).$$

But $\psi_n(\bar{v}) = \overline{\psi_n(v)}$, hence (A6) becomes

$$\phi_n(t) = \frac{4\delta^{n+2}}{(1-\delta^2)^2}\, Re\,[\,\psi_n(v)\,].$$

Using $1/(1-\delta v) = \Sigma_{k=0}^{\infty} (\delta v)^k \ (|\delta v| < 1)$, $\psi_n(v)$ can be expressed as

$$\psi_n(v) = v^n(\delta - v) \sum_{k=0}^{\infty} (\delta v)^k.$$

Thus $\phi_n(t)$ can be simplified as follows

$$\begin{aligned}
\phi_n(t) &= \left[\frac{2\delta}{1-\delta^2} \right]^2 \delta^n \, Re \left[v^n(\delta - v) \sum_{k=0}^{\infty} \delta^k \, v^k \right] \\
&= \left[\frac{2\delta}{1-\delta^2} \right]^2 Re \left[\sum_{k=0}^{\infty} \delta^{n+k+1} \, v^{n+k} - \sum_{k=0}^{\infty} \delta^{n+k} \, v^{n+k+1} \right] \\
&= \left[\frac{2\delta}{1-\delta^2} \right]^2 Re \left[\sum_{k=n}^{\infty} \delta^{k+1} \, v^k - \sum_{k=n+1}^{\infty} \delta^{k-1} \, v^k \right], \qquad \text{(A7)}
\end{aligned}$$

where we have used the dummy variable $k' = k + n$ in the first summation and $k' = k + n + 1$ in the second summation and omitted primes in the final expression.

Since $v^k = e^{ik\theta}$, $Re(v^k) = \cos k\theta = \cos k(\cos^{-1} t) = T_k(t)$, the Chebyshev polynomial of degree k. Hence $\phi_n(t)$ in (A7) can be expressed in terms of the Chebyshev polynomials by

$$\phi_n(t) = \left[\frac{2\delta}{1-\delta^2} \right]^2 \left[\sum_{k=n}^{\infty} \delta^{k+1} \, T_k(t) - \sum_{k=n+1}^{\infty} \delta^{k-1} \, T_k(t) \right]. \qquad \text{(A8)}$$

We now express $1/(u-t)$ in terms of the Chebyshev polynomials.

$$\frac{1}{u-t} = \frac{1}{\frac{1}{2}(\delta+\delta^{-1})-\frac{1}{2}(v+v^{-1})}$$

$$= \frac{2\delta v}{(v-\delta)(1-\delta v)} \,. \tag{A9}$$

Using partial fractions (A9) can be simplified. Hence

$$\frac{1}{u-t} = \frac{2\delta}{1-\delta^2}\left[\frac{1}{1-\delta v} + \frac{\delta}{v-\delta}\right]$$

$$= \frac{2\delta}{1-\delta^2}\left[\frac{1}{1-\delta v} + \frac{\delta v^{-1}}{1-\delta v^{-1}}\right]$$

$$= \frac{2\delta}{1-\delta^2}\left[\frac{1}{1-\delta v} + \frac{\delta \bar{v}}{1-\delta \bar{v}}\right]$$

$$= \frac{2\delta}{1-\delta^2}\left[\sum_{k=0}^{\infty} (\delta v)^k + \delta\bar{v} \sum_{k=0}^{\infty} (\delta\bar{v})^k\right]$$

$$= \frac{2\delta}{1-\delta^2}\left[1 + \sum_{k=1}^{\infty} \delta^k v^k + \sum_{k=0}^{\infty} (\delta\bar{v})^{k+1}\right]$$

$$= \frac{2\delta}{1-\delta^2}\left[1 + \sum_{k=1}^{\infty} \delta^k\, v^k + \sum_{k=1}^{\infty} \delta^k\, \bar{v}^k\right]$$

$$= \frac{2\delta}{1-\delta^2}\left[1 + \sum_{k=1}^{\infty} \delta^k (v^k + \bar{v}^k)\right]$$

$$= \frac{2\delta}{1-\delta^2}\left[1 + 2\sum_{k=1}^{\infty} \delta^k\, Re(v^k)\right]$$

$$= \frac{2\delta}{1-\delta^2}\left[1 + 2\sum_{k=1}^{\infty} \delta^k\, T_k(t)\right] \tag{A10}$$

Hence from (A1),(A8) and (A10)

$$\tilde{r}_n(t) = \frac{1}{u-t} + \phi_n(t)$$

$$= \frac{2\delta}{1-\delta^2}\left[1 + 2\sum_{k=1}^{\infty}\delta^k T_k + \frac{2\delta}{1-\delta^2}\sum_{k=n}^{\infty}\delta^{k+1} T_k - \frac{2\delta}{1-\delta^2}\sum_{k=n+1}^{\infty}\delta^{k-1} T_k\right]$$

$$= \frac{2\delta}{1-\delta^2}\left[1 + 2\sum_{k=1}^{n-1}\delta^k T_k + 2\delta^n T_n + 2\sum_{k=n+1}^{\infty}\delta^k T_k + \frac{2\delta}{1-\delta^2}\delta^{n+1} T_n\right.$$
$$\left. + \frac{2\delta}{1-\delta^2}\sum_{k=n+1}^{\infty}\delta^{k+1} T_k - \frac{2\delta}{1-\delta^2}\sum_{k=n+1}^{\infty}\delta^{k-1} T_k\right]$$

$$= \frac{2\delta}{1-\delta^2}\left[1 + 2\sum_{k=1}^{n-1}\delta^k T_k + \left[2\delta^n + \frac{2\delta}{1-\delta^2}\delta^{n+1}\right] T_n\right.$$
$$\left. + \sum_{k=n+1}^{\infty}\left[2\delta^k + \frac{2\delta}{1-\delta^2}\delta^{k+1} - \frac{2\delta}{1-\delta^2}\delta^{k-1}\right] T_k\right]$$

$$= \frac{2\delta}{1-\delta^2}\left[1 + 2\sum_{k=1}^{n-1}\delta^k T_k + \frac{2\delta^n}{1-\delta^2} T_n\right], \qquad \text{(A11)}$$

since

$$2\delta^k + \frac{2\delta}{1-\delta^2}\delta^{k+1} - \frac{2\delta}{1-\delta^2}\delta^{k-1} = 0\,.$$

DERIVATION OF $P_0(\lambda)$ AND $P_1(\lambda)$

We first derive expressions for $\tilde{r}_0(t)$ and $\tilde{r}_1(t)$ from first principles.

Derivation of $\tilde{r}_0(t)$

From (A5)

$$\Psi_0(t) = cos\ \eta\ , \tag{A12}$$

where from (A3)

$$\begin{aligned} e^{i\eta} &= \frac{\delta - v}{1 - \delta v} \\ &= \left[\frac{\delta - v}{1 - \delta v}\right]\left[\frac{1 - \delta \bar{v}}{1 - \delta \bar{v}}\right] \\ &= \frac{2\delta - \delta^2 e^{-i\theta} - e^{i\theta}}{1 - 2\delta\, cos\theta + \delta^2}\ , \end{aligned} \tag{A13}$$

where we have used $v = e^{i\theta}$. Taking the real part of (A13), using $t = cos\ \theta$ and $\delta^2 = 2u\delta - 1$, we obtain

$$cos\ \eta = \frac{1 - ut}{u - t}\ . \tag{A14}$$

From equations (A1), (A4), (A12) and (A14) we have

$$\begin{aligned} \tilde{r}_0(t) &= \frac{1}{u-t} + \phi_0(t) \\ &= \frac{1}{u-t} + \frac{1-ut}{(u^2-1)(u-t)} \\ &= \frac{u}{u^2-1} \end{aligned} \tag{A15}$$

Derivation of $\tilde{r}_1(t)$

Similarly from (A5)

$$\Psi_1(t) = cos(\theta + \eta)$$

$$= cos\ \theta\ cos\ \eta - sin\ \theta\ sin\ \eta\ . \qquad \text{(A16)}$$

Taking the imaginary part of (A13) and simplifying as before, we obtain

$$sin\ \eta = \frac{\delta^2\text{-}1}{2\delta\,(u\text{-}t)}\ sin\ \theta\ .$$

But $(\delta^2 - 1)/2\delta = u - 1/\delta = \delta - u = -\sqrt{u^2 - 1}$, so that

$$sin\ \eta = -\frac{\sqrt{u^2\text{-}1}}{u\text{-}t}\ sin\ \theta\ . \qquad \text{(A17)}$$

Using $t = cos\ \theta$ and substituting for $cos\ \eta$ and $sin\ \eta$ from (A14) and (A17) into (A16) we obtain

$$\Psi_1(t) = \frac{1}{u\text{-}t}\left[\ t(1 - ut) + (1 - t^2)\sqrt{u^2 - 1}\ \right]\ . \qquad \text{(A18)}$$

From (A1), (A4) and (A18)

$$\tilde{r}_1(t) = \frac{1}{u\text{-}t} + \frac{\delta}{u^2\text{-}1}\,\Psi_1(t)$$

$$= \frac{1}{(u\text{-}t)\,(u^2\text{-}1)}\,[\ u^2 - 1 + \delta\,(u - t)\ \Psi_1(t)\ ]$$

$$= \frac{1}{(u\text{-}t)\,(u^2\text{-}1)}\left[u^2 - 1 + \left[u - \sqrt{u^2 - 1}\right]\left\{\ t(1 - ut) + (1 - t^2)\sqrt{u^2 - 1}\ \right\}\right],$$

where $\delta = u - \sqrt{u^2 - 1}$ has been substituted. After some simplification it can be shown that

$$\tilde{r}_1(t) = \frac{1}{\sqrt{u^2 - 1}} + \frac{t}{u^2 - 1}, \tag{A19}$$

which is the same as from (A11) for $n = 1$.

We recall from chapter 2 that $P_n(\lambda) = a\ \tilde{r}_n(t)$, where $a = 2/(M - m)$ and $t = u - a\lambda$, with $u = (M + m)/(M - m)$. Thus using (A15) and (A19) we obtain

$$P_0(\lambda) = \frac{1}{2}\left[\frac{1}{M} + \frac{1}{m}\right], \tag{A20}$$

and

$$P_1(\lambda) = \frac{(\sqrt{M}+\sqrt{m})^2}{2mM} - \frac{\lambda}{Mm}. \tag{A21}$$

DERIVATION OF RECURRENCE RELATION

From (A5)

$$\begin{aligned}
\Psi_{n+2} &= cos[(n+2)\theta + \eta] \\
&= cos[(n+1)\theta + \eta + \theta] \\
&= cos[(n+1)\theta + \eta]\ cos\ \theta - sin[(n+1)\theta + \eta]\ sin\ \theta \\
&= t\,\Psi_{n+1} - \frac{1}{2}\left[\ cos(n\theta + \eta) - cos[(n+2)\theta + \eta]\ \right].
\end{aligned}$$

Here we have used $t = cos\ \theta$ and the well known trigonometric identity

$$sin\ A\ sin\ B = 1/2\ [\ cos(A - B) - cos(A + B)\ ].$$

Hence we obtain the relation

$$\Psi_{n+2} = 2t\,\Psi_{n+1} - \Psi_n \tag{A22}$$

From equations (A1), (A4) and (A22) we have,

$$\begin{aligned}
\tilde{r}_{n+2} &= \frac{1}{u-t} + \frac{\delta^{n+2}}{u^2-1}\,\Psi_{n+2} \\
&= \frac{1}{u-t} + \frac{\delta^{n+2}}{u^2-1}\,(2t\,\Psi_{n+1} - \Psi_n) \\
&= \frac{1}{u-t} + 2\delta t\,\frac{\delta^{n+1}}{u^2-1}\,\Psi_{n+1} - \delta^2\,\frac{\delta^n}{u^2-1}\,\Psi_n \\
&= \frac{1}{u-t} + 2\delta t\left[\frac{1}{u-t} + \frac{\delta^{n+1}}{u^2-1}\,\Psi_{n+1} - \frac{1}{u-t}\right] \\
&\quad - \delta^2\left[\frac{1}{u-t} + \frac{\delta^n}{u^2-1}\,\Psi_n - \frac{1}{u-t}\right] \\
&= 2\delta t\,\tilde{r}_{n+1} - \delta^2\,\tilde{r}_n + \frac{1-2\delta t+\delta^2}{u-t}\,.
\end{aligned}$$

But $1 - 2\delta t + \delta^2 = 2\delta\,(u-t)$, hence

$$\tilde{r}_{n+2} = 2\delta t\,\tilde{r}_{n+1} - \delta^2\,\tilde{r}_n + 2\delta\,. \tag{A23}$$

Multiplying by $a = 2/\,(M-m)$, one obtains the recurrence relation

$$P_{n+2}(\lambda) = 2\delta t\,P_{n+1}(\lambda) - \delta^2\,P_n(\lambda) + 2a\delta. \tag{A24}$$

for $P_n(\lambda)$.

APPENDIX B

Here we derive results pertaining to the polynomials $Q_n(A)$. These results represent original work.

DERIVATION OF ERROR

From equation (2.18), the polynomial $Z_n(\lambda)$ of degree n best approximating zero in the interval $[m,M]$, with $Z_n(0) = 1$, is given by

$$Z_n(\lambda) = \left[T_n\left[\frac{M+m}{M-m} \right] \right]^{-1} T_n\left[\frac{M+m-2\lambda}{M-m} \right] , \tag{B1}$$

with maximum error of approximation

$$\left[T_n\left[\frac{M+m}{M-m} \right] \right]^{-1} . \tag{B2}$$

With $u = (M+m)/(M-m)$,

$$\begin{aligned} T_n\left[\frac{M+m}{M-m} \right] &= T_n(u) \\ &= cos\ n(cos^{-1}u) \\ &= cos\ nz , \end{aligned} \tag{B3}$$

where $z = cos^{-1}u$ is in general complex.

Now

$$\begin{aligned} u &= \cos z \\ &= \frac{e^{iz}+e^{-iz}}{2} \\ &= \frac{\beta+\beta^{-1}}{2}, \end{aligned} \tag{B4}$$

where $\beta = e^{iz}$. Hence β satisfies the quadratic equation

$$\beta^2 - 2u\beta + 1 = 0 , \tag{B5}$$

with roots β_1 and β_2 given by

$$\begin{aligned} \beta_1 &= u - \sqrt{u^2-1} , \\ \beta_2 &= u + \sqrt{u^2-1} . \end{aligned} \tag{B6}$$

In fact $\beta_1\beta_2 = 1$ from (B5) or directly from (B6). Substituting $u = (M+m)/(M-m)$ in (B6) it is easy to show that

$$\beta_1 = \frac{\sqrt{M}-\sqrt{m}}{\sqrt{M}+\sqrt{m}} = \delta .$$

Hence from (B3)

$$\begin{aligned} T_n(u) &= \cos nz \\ &= \frac{e^{inz}+e^{-inz}}{2} \\ &= \frac{\beta_1^n+\beta_1^{-n}}{2} \\ &= \frac{\delta^n+\delta^{-n}}{2} \end{aligned} \tag{B7}$$

We note that if $\beta_2 = \delta^{-1}$ is taken in place of β_1, the expression (B7) remains unchanged.

Hence the maximum error of approximation is given by

$$T_n^{-1}(u) = \frac{2}{\delta^n + \delta^{-n}}. \tag{B8}$$

DERIVATION OF $Q_0(\lambda)$

Letting $n = 0$ in equation (2.20) one obtains the expression

$$Q_0(\lambda) = \frac{1}{\lambda}\left[\, 1 - \frac{2}{\delta + \delta^{-1}}\, T_1\left[\, \frac{M+m-2\lambda}{M-m} \,\right] \right].$$

substituting $T_1(t) = t = (M+m-2\lambda)/(M-m)$ and $2/(\delta+\delta^{-1}) = (M-m)/(M+m)$ in the above equation we obtain

$$\begin{aligned} Q_0(\lambda) &= \frac{1}{\lambda}\left[\, 1 - \frac{M+m-2\lambda}{M+m} \,\right] \\ &= \frac{2}{M+m}. \end{aligned} \tag{B9}$$

DERIVATION OF $Q_1(\lambda)$

Letting $n = 1$ in equation (2.20) one obtains the expression

$$Q_1(\lambda) = \frac{1}{\lambda}\left[\, 1 - \frac{2}{\delta^2 + \delta^{-2}}\, T_2\left[\, \frac{M+m-2\lambda}{M-m} \,\right] \right].$$

Using

$$T_2(t) = 2t^2 - 1$$

$$= \frac{(M+m)^2 + 4Mm - 8(M+m)\lambda + 8\lambda^2}{(M-m)^2}$$

and

$$\frac{2}{\delta^2+\delta^{-2}} = \frac{(M-m)^2}{(M+m)^2 + 4Mm},$$

$Q_1(\lambda)$ can be expressed as

$$Q_1(\lambda) = \frac{1}{\lambda}\left[1 - \left\{1 - \frac{8(M+m)\lambda + 8\lambda^2}{(M+m)^2 + 4Mm}\right\}\right]$$

$$= \frac{8(M+m) - 8\lambda}{(M+m)^2 + 4Mm}. \qquad \text{(B10)}$$

DERIVATION OF RECURRENCE RELATION

We derive a recurrence relation for the polynomials $Q_n(\lambda)$, by using the recurrence relation

$$T_{n+2}(t) = 2tT_{n+1}(t) - T_n(t), \qquad \text{(B11)}$$

for the Chebyshev polynomials.

We rewrite (B1) in the form

$$Z_n(\lambda) = \frac{T_n(t)}{T_n(u)},$$

where $t = (M + m - 2\lambda)/(M - m) = u - a\lambda$, with $u = (M + m)/(M - m)$ and $a = 2/(M - m)$. Hence

$$\begin{aligned} Z_{n+2}(\lambda) &= \frac{T_{n+2}(t)}{T_{n+2}(u)} \\ &= \frac{2tT_{n+1}(t) - T_n(t)}{T_{n+2}(u)} \\ &= 2t\,\frac{T_{n+1}(t)}{T_{n+1}(u)}\,\frac{T_{n+1}(u)}{T_{n+2}(u)} - \frac{T_n(t)}{T_n(u)}\,\frac{T_n(u)}{T_{n+2}(u)} \\ &= 2t\,Z_{n+1}(\lambda)\,\frac{T_{n+1}(u)}{T_{n+2}(u)} - Z_n(\lambda)\,\frac{T_n(u)}{T_{n+2}(u)}. \end{aligned} \tag{B12}$$

From the recurrence relation (B11) we obtain

$$\frac{T_n(u)}{T_{n+2}(u)} = 2u\,\frac{T_{n+1}(u)}{T_{n+2}(u)} - 1 .$$

With the above substitution (B12) becomes

$$Z_{n+2}(\lambda) = Z_n(\lambda) + \frac{T_{n+1}(u)}{T_{n+2}(u)}\,[\,2t\,Z_{n+1}(\lambda) - 2u\,Z_n(\lambda)\,] .$$

Substituting $Z_k(\lambda) = 1 - \lambda\ Q_{k-1}(\lambda)$ $(k = n,\ n+1,\ n+2)$, into the above equation we

obtain

$$\lambda\ Q_{n+1}(\lambda) = \lambda\ Q_{n-1}(\lambda) + \frac{T_{n+1}(u)}{T_{n+2}(u)}\left[\ 2t\lambda\ Q_n(\lambda) - 2u\lambda\ Q_{n-1}(\lambda) + 2(u-t)\ \right].$$

Using $u - t = a\ \lambda$ and (from B8)

$$\frac{T_{n+1}(u)}{T_{n+2}(u)} = \frac{\delta^{n+1} + \delta^{-(n+1)}}{\delta^{n+2} + \delta^{-(n+2)}} = \delta\left[\frac{1+\delta^{2n+2}}{1+\delta^{2n+4}}\right],$$

the above equation can be simplified to read

$$Q_{n+1}(\lambda) = Q_{n-1}(\lambda) + \left[\frac{1+\delta^{2n+2}}{1+\delta^{2n+4}}\right]\left[\ 2\delta t\ Q_n(\lambda) - 2\delta u\ Q_{n-1}(\lambda) + 2a\delta\ \right],$$

which can more conveniently be written as

$$Q_{n+2}(\lambda) = Q_n(\lambda) + \left[\frac{1+\delta^{2n+4}}{1+\delta^{2n+6}}\right]\left[\ 2\delta t\ Q_{n+1}(\lambda) - 2\delta u\ Q_n(\lambda) + 2a\delta\ \right]. \qquad \text{(B13)}$$

APPENDIX C

In this appendix we review some of the earlier work done on the approximation to the inverse operator A^{-1}. Here A is not necessarily a positive definite operator.

In 1985, Mika, Cole and Pack [21] considered the approximation of the inverse operator A^{-1} by approximations of the form $z_0 I$, where A is a linear positive definite operator in the sense that the inequality

$$m < \phi,\phi > \leq < A\phi,\phi > \leq M < \phi,\phi >$$

is satisfied $\forall \phi \in H$. Using $A^{-1} \approx z I$ we have

$$\| A^{-1} - z I \| = \sup_{\lambda \in [m, M]} \left| \frac{1}{\lambda} - z \right| . \qquad \text{(C1)}$$

Since the supremum must occur at the end points m and M, we have

$$\| A^{-1} - z I \| = max \left\{ \left| \frac{1}{m} - z \right|, \left| \frac{1}{M} - z \right| \right\} = \gamma(z) . \qquad \text{(C2)}$$

Minimization of (C2) with respect to z yields the optimal value

$$z_0 = \frac{1}{2} \left[\frac{1}{m} + \frac{1}{M} \right]. \qquad \text{(C3)}$$

This is illustrated in figure C.1 overleaf.

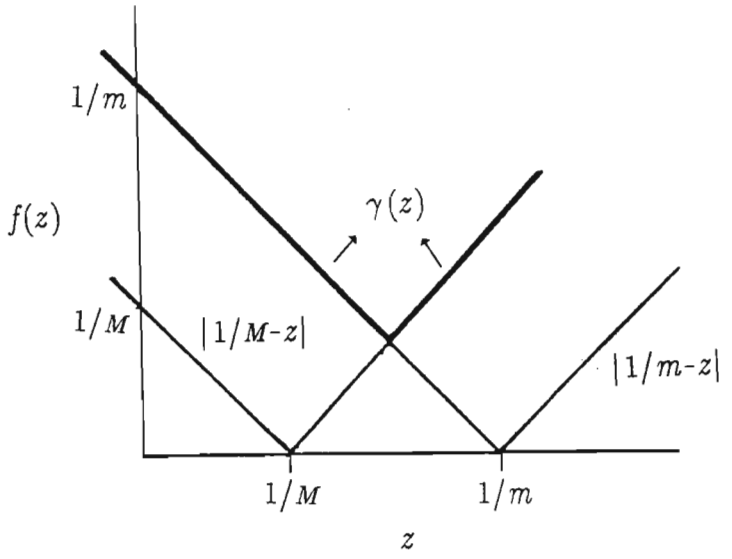

FIG C.1: Zero order approximation to A^{-1} .

The quantity $\gamma(z_0)$ yields the optimal error of approximation

$$\gamma_0 = \frac{1}{2}\left[\frac{1}{m} - \frac{1}{M}\right]. \tag{C4}$$

For non-selfadjoint operators A , they made the assumption

$$m^2 \|\phi\|^2 \leq \|A\phi\|^2 \leq M^2 \|\phi\|^2, \forall \phi \in D(A) ,$$

which is equivalent to $m^2 \|\phi\|^2 \leq < A^\star A\phi, \phi > \leq M^2 \|\phi\|^2$. Now

$$A^{-1} = A^{-1} (A^\star)^{-1} A^\star = (A^\star A)^{-1} A^\star.$$

Since $A^\star A$ is positive definite, the best zero order approximation to $(A^\star A)^{-1}$ is $1/2\,(1/m^2 + 1/M^2)\, I$.

Hence A^{-1} approximated to first order in $A^{\star}$ is given by

$$\frac{1}{2}\left[\frac{1}{m^2}+\frac{1}{M^2}\right]A^{\star}. \qquad \text{(C5)}$$

It must be stated that the preceding approximation is due to the best approximation of $(A^{\star}A)^{-1}$ by approximations of the form $z_0 I$ and is hence not the best approximation of A^{-1}.

Mika and Pack [26] in 1986, sought approximations of the form $z_0 I$ to A^{-1}, where $z_0 \in \mathbb{C}(z)$ – the space of complex numbers, and A is a bounded linear and normal operator in a complex Hilbert space H. Their technique involved locating the spectrum $\sigma(A^{-1})$ from the spectrum $\sigma(A)$ and hence finding $z = z_0$ that minimizes

$$\| A^{-1} - z I \| = \sup_{\lambda \in \sigma(A)} \left| \frac{1}{\lambda} - z \right|. \qquad \text{(C6)}$$

For positive definite A with $\sigma(A) \subset [m,M]$, $M > m > 0$, they obtained the well known result in (C3) and (C4).

For $A = S + i\omega$ (ω a real constant), with S positive definite and $\sigma(S) \subset [m,M]$, $\sigma(A^{-1})$ lies on the circle centred at $-i/2\omega$ with radius $1/2\omega$ between points $1/(M+i\omega)$ and $1/(m+i\omega)$. From the illustration in figure C.2 it is clear that z_0 is the centre of the chord joining these two points.

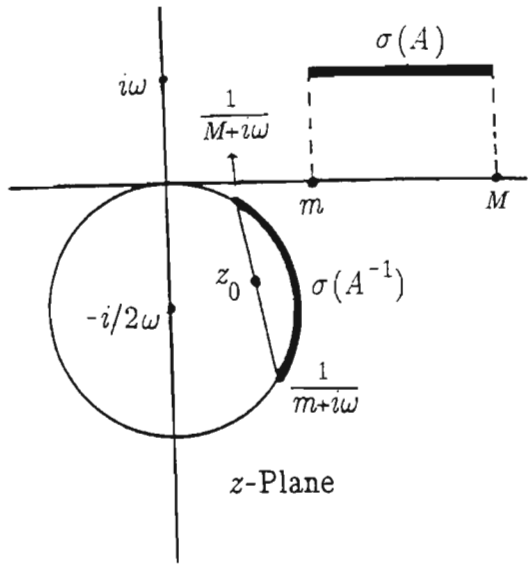

FIG C.2: Zero order approximation to A^{-1} .

Thus

$$z_0 = \frac{1}{2}\left[\frac{1}{M+i\omega} + \frac{1}{m+i\omega}\right] \tag{C7}$$

and

$$\gamma_0 = \frac{1}{2}\left|\frac{1}{M+i\omega} - \frac{1}{m+i\omega}\right| . \tag{C8}$$

If $\sigma(A)$ lies within a circle of radius ρ with centre $\xi = r\,e^{i\phi}$, $\rho < r$, then $\sigma(A^{-1})$ is contained in a circle of which the centre is z_0 and the radius is γ_0 . An example is the operator $A = \xi I + K$ with $\| K \| \leq \rho < r$. This is shown in figure C.3 .

Thus

$$z_0 = \frac{1}{2}\left[\frac{1}{r-\rho} + \frac{1}{r+\rho}\right] e^{-i\phi} = \frac{r\;e^{-i\phi}}{r^2-\rho^2} \tag{C9}$$

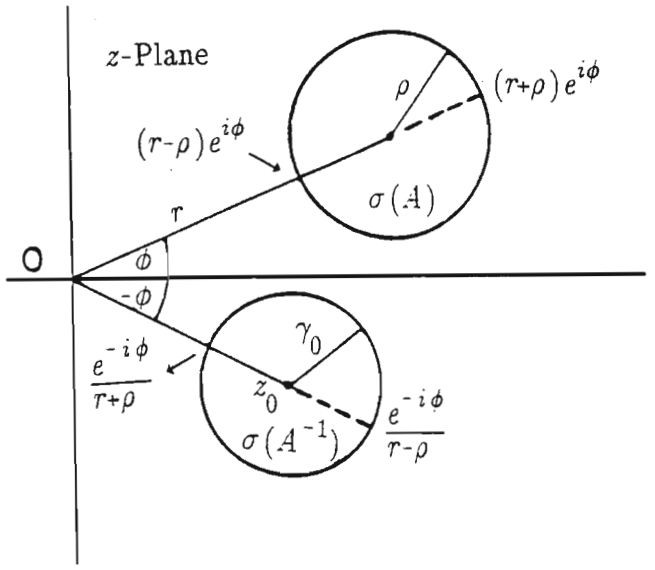

FIG C.3: Zero order approximation to A^{-1} .

and

$$\gamma_0 = \frac{1}{2}\left[\frac{1}{r-\rho} - \frac{1}{r+\rho}\right] = \frac{\rho}{r^2-\rho^2} . \qquad \text{(C10)}$$

If A is such that $\sigma(A) = \{\lambda : Re\,\lambda \leq -k\}$, then $\sigma(A^{-1})$ is the circle centred at $z_0 = -1/2k$ with radius $\gamma_0 = 1/2k$. (see figure C.4).

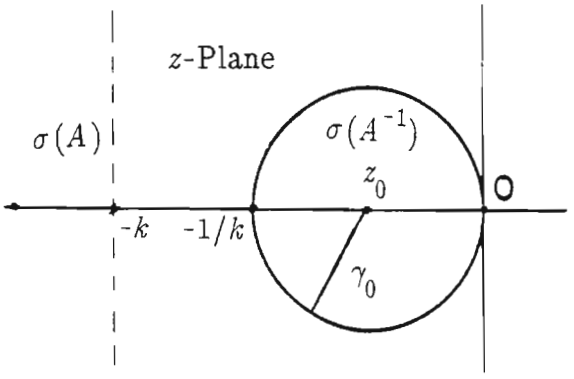

FIG C.4: Zero order approximation to A^{-1} .

Mika and Pack [26] also considered polynomial approximations $P_n(A)$ to the inverse operator A^{-1} for positive definite A. They found the zero and first order approximations together with the associated errors and stated the necessity of numerical algorithms to find higher order approximations. Thus they obtained

$$P_1(A) = \frac{(\sqrt{M}+\sqrt{m})^2}{2\,mM} - \frac{A}{Mm} \tag{C11}$$

with associated error

$$\gamma_1 = \frac{(\sqrt{M}-\sqrt{m})^2}{2\,mM} \,. \tag{C12}$$

For operators of the form $A = I + K$, where K is positive definite with $\sigma(K) \subset [0,M]$, $M < 1$, they showed the superiority of γ_1 over γ_{neu}, the error associated with the first order Neumann expansion of $(I+K)^{-1}$. In fact they showed that γ_1 can be reduced to less than $1/8$ of γ_{neu}.

Using the spectral theory of unbounded normal operators, Lamb, Mika and Roach [27], in 1987, found polynomial approximations for resolvents and semigroups in terms of inverses and resolvents respectively. Consider the resolvent $(I+A)^{-1}$, where A is unbounded and selfadjoint with $\sigma(A) \subset [1/c,\infty)$ $(0 < c < 1)$. Since $\sigma[(I+A)^{-1}] \subset (0, c/(c+1)]$, the zero order approximation to $(I+A)^{-1}$ is given by

$$\frac{c}{2(c+1)}\, I \,, \tag{C13}$$

with associated error

$$\frac{c}{2(c+1)}\,. \tag{C14}$$

Since polynomials in A are unbounded, it is not possible to find higher order approximations in terms of A. Because A^{-1} is bounded, higher order approximations in terms of A^{-1} are possible. Thus the first order approximation to $(I+A)^{-1}$ is represented by the polynomial

$$P_1(A^{-1}) = \left[\frac{2+c}{2(1+c)} - \frac{1}{\sqrt{1+c}}\right] I + \frac{A^{-1}}{1+c}\,, \tag{C15}$$

with the error given by

$$\frac{1}{2}\left[1-(1+c)^{-\frac{1}{2}}\right]^2. \tag{C16}$$

Using a Neumann expansion one can obtain the second order expression

$$(I+A)^{-1} = (I+A^{-1})^{-1}\,A^{-1} \approx A^{-1} - A^{-2}.$$

It was shown in [27] that the optimal polynomial $P_1(A^{-1})$ given in (C15) is superior to the Neumann expansion of second order for most values of c. However, for values of c close to zero the Neumann expansion is better, but at the expense of using a non optimal second order polynomial.

The problem of the zero order approximation of inverses of arbitrary operators A whose spectrum $\sigma(A)$ is in general complex, is obviously complicated by the difficulty in

locating $\sigma(A)$ and hence $\sigma(A^{-1})$. One can imagine the difficulty of obtaining higher order approximations to inverses of operators whose spectrum is complex.

In 1992 Bond and Mika [1] derived the polynomials $P_n(A)$ discussed in chapter 2. They obtained the three point recurrence relation (2.12) thus providing an effective means for evaluating these polynomials. However they did not investigate the application and usefulness of these polynomials.

APPENDIX D

In this appendix we derive in detail some of the results for linear functionals presented in chapter 5.

DERIVATION OF $J_f(\Phi)$ AND BOUNDS FOR Ω_f

Bounds for $\Omega_f = < x,f >$ in equation (5.4) can easily be calculated by using the trial function $\Phi = x + \Delta x$, where x is the solution of equation (5.1) and Δx represents the error in approximating x by Φ. Here we take a real Hilbert space and hence symmetric inner products and consider positive definite A, writing

$$< x,f > = < \Phi,f > - < \Delta x,f > . \tag{D1}$$

Now

$$\begin{aligned} < \Delta x,f > &= < \Delta x,Ax > \\ &= < \Delta x,A\Phi - A\Delta x > \\ &= < \Delta x,A\Phi > - < \Delta x,A\Delta x > \\ &= < A\Delta x,\Phi > - < A\Delta x,\Delta x > \\ &= < A\Phi - f,\Phi > - < A\Delta x,\Delta x > . \end{aligned} \tag{D2}$$

Substituting (D2) into (D1), one obtains

$$< x,f > = J_f(\Phi) + < A\Delta x,\Delta x > , \tag{D3}$$

where

$$J_f(\Phi) = < \Phi, 2f - A\Phi > . \tag{D4}$$

Since it is assumed that A is bounded below by $m > 0$ and above by $M > m$, the condition

$$m < \phi, \phi > \leq < A\phi, \phi > \leq M < \phi, \phi >$$

is satisfied $\forall \phi \in H$.

Hence

$$\begin{aligned} < A\Delta x, \Delta x > &= < A^{1/2}\Delta x, A^{1/2}\Delta x > \\ &\leq 1/m < A(A^{1/2}\Delta x), A^{1/2}\Delta x > \\ &= 1/m < A\Delta x, A\Delta x > \\ &= 1/m \, \| A\Phi - f \|^2 . \end{aligned} \tag{D5}$$

Similarly, it can be shown that

$$< A\Delta x, \Delta x > \geq 1/M \, \| A\Phi - f \|^2 . \tag{D6}$$

Using (D5) and (D6) in (D3) we obtain the result

$$J_f(\Phi) + 1/M \, \| A\Phi - f \|^2 \leq \Omega_f \leq J_f(\Phi) + 1/m \, \| A\Phi - f \|^2 , \tag{D7}$$

as in equation (5.4).

Now

$$
\begin{aligned}
J_f(\Phi) &= <\Phi, 2f - A\Phi> \\
&= <\Phi, f> + <\Phi, f - A\Phi> \\
&= <\Phi, f> + <\Phi, A(x - \Phi)> \\
&= <\Phi, f> + <\Phi - x, A(x - \Phi)> + <x, A(x - \Phi)> \\
&= <Ax, x> - <A(x - \Phi), x - \Phi> . \qquad \text{(D8)}
\end{aligned}
$$

Since A is positive definite, it follows from (D8) that $J_f(\Phi)$ is maximum when $\Phi = x$.

The functional $J_f(\Phi)$ can also be derived by using variational calculus [28]. Here the linear functionals are defined by a suitable integral with respect to the variable z in the interval (a,b). We depart from convention and use the symbol Δ instead of δ to denote increments and first order variations. This is to avoid confusion since the symbol δ has already been used in chapter 2 to denote a convergence parameter.

The basic problem is to find a functional $J_f(\Phi)$ $(\Phi = \Phi(z))$ such that increments $\Delta\Phi$ in Φ yield only second order increments in the functional $J_f(\Phi)$. This is represented by the equation

$$
J_f(\Phi + \Delta\Phi) = J_f(\Phi) + O(\Delta\Phi^2) . \qquad \text{(D9)}
$$

Terms involving $O(\Delta\Phi)$ in $J_f(\Phi + \Delta\Phi) - J_f(\Phi)$ represent the first variation ΔJ_f of $J_f(\Phi)$. The requirement that this first variation disappears results in an equation called the Euler equation.

If the equation

$$A\Phi = f \tag{D10}$$

is to be an Euler equation of of the functional $J_f(\Phi)$, then the first variation ΔJ_f must have the form

$$\begin{aligned}
\Delta J_f = \frac{\partial J_f}{\partial \Phi} \Delta\Phi &= \int_a^b (f - A\Phi)\, \Delta\Phi \; dz \\
&= < f - A\Phi, \Delta\Phi > \\
&= < f, \Delta\Phi > - < A\Phi, \Delta\Phi > \\
&= \Delta < f, \Phi > - 1/2\, \Delta < A\Phi, \Phi > \\
&= \frac{1}{2} \Delta \left[< 2f, \Phi > - < A\Phi, \Phi > \right] .
\end{aligned} \tag{D11}$$

From (D11) it follows that $J_f(\Phi)$ can be taken as

$$J_f(\Phi) = < 2f, \Phi > - < A\Phi, \Phi > = < \Phi, 2f - A\Phi > ,$$

which is the same as (D4).

DERIVATION OF $J(\Psi,\Phi)$ AND ADJOINT EQUATION

We derive the functional $J(\Psi,\Phi)$ by again using variational calculus [28]. For non-selfadjoint A it is necessary to introduce an additional trial function $\Psi = \Psi(z)$. Hence, if equation (D10) is to be the Euler equation of the functional $J(\Psi,\Phi)$, the first variation $\Delta J^{(\Psi)}$ with respect to Ψ must have the form

$$\Delta J^{(\Psi)} = \frac{\partial J}{\partial \Psi}\,\Delta\Psi = \int_a^b (f - A\Phi)\,\Delta\Psi\; dz$$

$$= < f - A\Phi, \Delta\Psi > . \qquad \text{(D12)}$$

Integrating (D12) with respect to Ψ yields

$$J(\Psi,\Phi) = \int_a^b [\; (f - A\Phi)\,\Psi + h(\Phi,z)\;]\; dz\,, \qquad \text{(D13)}$$

where $h(\Phi,z)$ is a constant of integration. Now the first variation $\Delta J^{(\Phi)}$ of equation (D13) with respect to Φ yields

$$\Delta J^{(\Phi)} = \frac{\partial J}{\partial \Phi}\,\Delta\Phi = \int_a^b [\; - A\Delta\Phi\;\Psi + \frac{\partial h}{\partial \Phi}\,\Delta\Phi\;]\; dz$$

$$= - < A\Delta\Phi, \Psi > + < \frac{\partial h}{\partial \Phi}, \Delta\Phi > . \qquad \text{(D14)}$$

If $h(\Phi,z) = g(z)\;\Phi$, then from (D14)

$$\Delta J^{(\Phi)} = - < A\Delta\Phi, \Psi > + < g, \Delta\Phi >$$

$$= - < \Delta\Phi, A^{\star}\Psi > + < \Delta\Phi, g >$$

$$= - < \Delta\Phi, A^{\star}\Psi - g > . \qquad \text{(D15)}$$

In the above we have assumed real inner products. Since the first variation $\Delta J^{(\Phi)}$ must also disappear, we obtain the dual or adjoint equation

$$A^{\star}\Psi = g\,. \qquad \text{(D16)}$$

From (D13) we obtain

$$J(\Psi,\Phi) = <\Phi,g> + <f,\Psi> - <A\Phi,\Psi> . \qquad \text{(D17)}$$

DERIVATION OF BOUNDS FOR Ω

Bounds for $\Omega = <x,g>$ in equation (5.17) can easily be obtained by using the trial functions $\Phi = x + \Delta x$ and $\Psi = y + \Delta y$. Here we recall that x is the solution of the equation $Ax = f$ and y is the solution of the dual or adjoint equation $A^{\star}y = g$. Hence

$$<x,g> = <\Phi,g> - <\Delta x, A^{\star}y> . \qquad \text{(D18)}$$

Now

$$\begin{aligned} <\Delta x, A^{\star}y> &= <\Delta x, A^{\star}\Psi - A^{\star}\Delta y> \\ &= <A\Delta x,\Psi> - <\Delta x, A^{\star}\Delta y> \\ &= <A\Phi - f,\Psi> - <\Phi - x, A^{\star}\Psi - g> . \end{aligned} \qquad \text{(D19)}$$

Substituting (D19) into (D18), one obtains

$$<x,g> = J(\Psi,\Phi) + <x-\Phi, g - A^{\star}\Psi> , \qquad \text{(D20)}$$

where $J(\Psi,\Phi)$ is given by (D17). Using $x - \Phi = A^{-1}(f - A\Phi)$, the last term in (D20) can be rewritten as

$$\begin{aligned} <x-\Phi, g - A^{\star}\Psi> = &<(A^{-1} - V)(f - A\Phi), g - A^{\star}\Psi> \\ &+ <V(f - A\Phi), g - A^{\star}\Psi> , \end{aligned} \qquad \text{(D21)}$$

where V is an operator approximating the inverse A^{-1}. Hence (D20) becomes

$$< x,g > = J(\Psi,\Phi) + < V(f - A\Phi),g - A^{\star}\Psi > \\ + < (A^{-1} - V)(f - A\Phi),g - A^{\star}\Psi > . \qquad \text{(D22)}$$

Using the Schwarz inequality

$$|< (A^{-1} - V)(f - A\Phi),g - A^{\star}\Psi >| \leq \| A^{-1} - V \| \| f - A\Phi \| \| g - A^{\star}\Psi \| ,$$

one obtains the result

$$\Omega \leq J + < V(f - A\Phi),g - A^{\star}\Psi > + \| A^{-1} - V \| \| g - A^{\star}\Psi \| \| f - A\Phi \|, \\ \Omega \geq J + < V(f - A\Phi),g - A^{\star}\Psi > - \| A^{-1} - V \| \| g - A^{\star}\Psi \| \| f - A\Phi \|, \qquad \text{(D23)}$$

of (5.17). Here J denotes $J(\Psi,\Phi)$.

REFERENCES

[1] R.A.B. Bond and J.R. Mika: *Method of Approximate Inverse for Solving Equations in Hilbert Space*, Quaestiones Mathematicae **15(1)**, 53-71, 1992.

[2] V.I. Lebedev and S.A. Finogenov: *Ordering of the Iterative Parameters in the* Cyclical Chebyshev Iterative Method, USSR Comput. Math. and Math. Phys. **11**(2), 155-170, 1971.

[3] K.E. Atkinson: *A Survey of Numerical Methods for the Solution of Fredholm Integral Equations of the Second Kind*, Philadelphia : Society for Industrial and Applied Mathematics, 1976.

[4] W. Rudin: *Functional Analysis*, McGraw–Hill Inc. , USA, 1973.

[5] G. Meinardus: *Approximation of Functions: Theory and Numerical Methods*, Springer, Berlin, 1967.

[6] G.A. Watson: *Approximation Theory and Numerical Methods*, John Wiley & Sons Ltd, New York, 1980

[7] J.R. Westlake: *A Handbook of Numerical Matrix Inversion and Solution of Linear Equations*, John Wiley & Sons Inc., New York, 1968.

[8] R.E. Kleinman, G.F. Roach, L.S. Schultz, J. Shirron, P.M. Van den Berg: *An Over-Relaxation method for the Iterative Solution of Integral Equations in Scattering Problems*, Wave Motion **12** , 161-170, 1990.

[9] D. Young: *On Richardson's Method for Solving Linear Systems with Positive Definite Matrices*, J. Math. and Phys. , **32**(4), 243-255, 1954.

[10] J.H. Wilkinson: *The Algebraic Eigenvalue problem*, Clarendon Press, Oxford, 1965.

[11] G.I. Marchuk: *Methods of Numerical Mathematics*, 2nd edition, Springer, New York, 1982.

[12] D.K. Faddeev and V.N. Faddeeva: *Computational Methods of Linear Algebra*, W.H. Freeman and Company, San Franscisco, 1963.

[13] A. Courant and D. Hilbert: *Methods of Mathematical Physics*, Interscience, New York, 1953.

[14] W.M. Stacey: *Variational Methods in Nuclear Reactor Physics*, Academic Press, New York, 1974.

[15] H. Hojgaard Jensen, H. Smith and J.W. Wilkins: *Upper and Lower Bounds on Transport Coefficients arising from a Linearized Boltzmann Equation*, Physical Review, **185**(1), 323-337, 1969.

[16] M.F. Barnsley and P.D. Robinson: *Bivariational Bounds*, Proc. R. Soc. London **A 338**, 527-533, 1974.

[17] R.J. Cole and D.C. Pack: *Some Complementary Bivariational Principles for Linear Integral Equations of Fredholm Type,* Proc. R. Soc. London **A 347**, 239-252, 1975.

[18] M.F. Barnsley and P.D. Robinson: *Bivariational Bounds Associated with Non-Seladjoint Linear Operators,* Proc. R. Soc. Edinburgh, **75 A**, 9, 109-118, 1975/76.

[19] P.D. Robinson: *New Variational Bounds on Generalized Polarizabilities,* J. Math. Phys., **19**(3), 694-699, 1978.

[20] M.F. Barnsley and P.D. Robinson: *Pointwise Bivariational Bounds on Solutions of Fredholm Integral Equations.,* Siam J. Numer. Anal., **16**(1), 135-144, 1979.

[21] J. Mika, D.C. Pack and R.J. Cole: *Optimal Bounds for Bilinear Forms Associated with Linear Equations,* Math. Meth. in the Appl. Sci., **7**, 518-531, 1985.

[22] P.D. Robinson and P.K. Yuen: *Bivariational Methods for Linear Integral Equations with Non−symmetric Kernels,* Siam J. Numer. Anal., **23**(6),1230-1240, 1986.

[23] J.R. Mika: *Spectral Method of Approximating Normal Operators in Hilbert Spaces,* Notices of the S.A. Mathematical Society, **19/3**, 200-215, 1987.

[24] L.M. Delves and J.L. Mohamed: *Computational Methods for Integral Equations,* Cambridge University Press, Cambridge, 1985.

[25] R. Piessens and M. Branders: *Numerical Solutions of Integral Equations of Mathematical Physics, using Chebyshev Polynomials.* Journal of Computational Physics **21**, 178-196, 1976.

[26] J.R. Mika, D.C. Pack: *Approximation to Inverses of Normal Operators,* Proc. Roy. Soc. Edinburgh **103 A**, 335-345, 1986 .

[27] W. Lamb, J.R. Mika and G.F. Roach; *Approximate Solutions of Problems Involving Normal Operators,* Journal of Math. Analysis and Appl., **126(1)**, 209-222, 1987 .

[28] M. Becker: *The Principles and Applications of Variational Methods,* M.I.T. press, Massachusetts, 1964 .