AN ABSTRACT OF THE THESIS OF

WALTER ARTHUR YUNGEN (Name) for the Master of Science (Degree)

in Mathematics (Major) presented on _____ (Date)

Title RIGOROUS COMPUTER INVERSION OF SOME LINEAR OPERATORS

Redacted for privacy

Abstract approved _____ (Dr. Joel Davis)

In this thesis we consider computer techniques for inverting $n \times n$ matrices and linear Fredholm integral operators of the second kind. We develop techniques which allow us to prove the existence of and find approximations to inverses for the above types of operators. In addition, we are able to bound rigorously the error in the approximations. These techniques were implemented in the form of computer programs and some numerical results are given.

Rigorous Computer Inversion of Some Linear Operators

by

Walter Arthur Yungen

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Master of Science

June 1968

APPROVED:

Redacted for privacy

Assistant Professor of Mathematics

In Charge of Major

Redacted for privacy

Chairman of Department of Mathematics

Redacted for privacy

Dean of Graduate School

Date thesis is presented July 31, 1967

Typed by Carol Baker for Walter Arthur Yungen

ACKNOWLEDGEMENT

I am indebted to my major professor, Dr. Joel Davis, for the considerable interest and encouragement given to me during the preparation of this thesis. Also, I wish to thank Dr. A. T. Lonseth, chairman of the Department of Mathematics, for his interest and support. This work was supported in part by the A.E.C. under contract No. A.T. (45-1) - 1947.

Finally, I express my appreciation to my wife whose patience, encouragement, and support has been invaluable.

TABLE OF CONTENTS

RIGOROUS COMPUTER INVERSION OF SOME LINEAR OPERATORS

I. INTRODUCTION

In this thesis we consider rigorous computer inversion of two families of linear operators, finite dimensional linear operators (i.e. matrices) and linear Fredholm integral operators of the second kind. Our goal is to determine the existence of an inverse and to find an approximation to the inverse with rigorous error bounds on the results.

For each family, we discuss a method of inversion, including a test for existence of the inverse. Also, we find rigorous error bounds on the results, including both the truncation error of the method and the roundoff error occurring in the computations. Rounding error is treated in detail, beginning with bounds for the error occurring in each arithmetic operation performed by the computer. In considering the truncation error for each method, we give only the needed results, drawing heavily from Faddeeva [6] and the work of Anselone and Moore [2].

In Chapter II, we introduce much of the basic notation and give some needed results from numerical analysis. We develop in Chapter III bounds for the roundoff error in each arithmetic operation to be used and apply the results to bounding the error in such compound operations as finding inner products. In Chapter IV we

give a method for inversion of $n \times n$ matrices which includes a test for existence of the inverse and also yields a bound on the truncation error. We also apply the results of Chapter III for bounding the roundoff error in order to give rigorous error bounds on the results. In Chapters V and VI we obtain approximate solutions with rigorous error bounds for linear Fredholm integral equations of the second kind. We also discuss some of the practical problems of implementing this method.

Using the methods discussed in this thesis, we have written computer programs that determine the existence of inverses and find numerical solutions, including rigorous error bounds. Some results from these programs are given. Finally, possible improvements and extensions will be discussed briefly in Chapter VII.

II. NOTATION AND BACKGROUND

The purpose of this chapter is to provide a brief sketch of the notation, definitions, and results needed in later chapters. The two major topics here will be linear operators and numerical quadrature.

The following material on spaces and operators is essentially as presented in Kolmogorov and Fomin [9]. We will be concerned with complete normed linear spaces over the real number field, in particular, with the following two spaces:

(i) the space $C[0,1]$ of continuous functions on $[0,1]$ with norm

$$\|x\| = \text{Max } \{ |x(t)| : 0 \leq t \leq 1 \}, \tag{2.1}$$

and

(ii) Euclidean n-space, R^n, with norm

$$\|r\| = \text{Max } \{ |r_i| : 1 \leq i \leq n \} . \tag{2.2}$$

Using the metric $\rho(x,y) = \|x-y\|$, the above spaces can be considered metric spaces with the resultant concepts of convergence and completeness. Now we can make the following definitions.

Definition 2.1

A Banach (or B-) space is a complete normed linear space. It can be shown that the above examples of spaces are complete and hence Banach spaces (see [9]). From this point on, let S, T denote Banach spaces.

Definition 2.2

A linear operator (linear function) is a function $K: S \to T$ having the following property: $\forall\ s, s' \in S$ and $a \in R$,

$$K(s+as') = K(s) + aK(s'). \tag{2.3}$$

Henceforth, we let K and K_n $(n = 1, 2, 3, \cdots)$ denote linear operators mapping S into T.

Definition 2.3

A bounded linear operator K is a linear operator with the property that there exists a constant M such that

$$\|Ks\| \leq M \cdot \|s\|, \quad \text{for all } s \in S. \tag{2.4}$$

It should be noted that for linear operators, continuity and boundedness are equivalent (see [9]).

Definition 2.4.

The norm $\|K\|$ of an operator K is the greatest lower bound of the numbers M satisfying (2.4), or equivalently,

$$\|K\| = \text{Sup}\{\|Ks\| / \|s\| : s \epsilon S, \quad s \neq 0\}. \tag{2.5}$$

It is easily seen that norms of linear operators have the following properties:

(i) $\|K_1 s\| \leq \|K_1\| \cdot \|s\|$, where $s \epsilon S$,

(ii) $\|K_1 K_2\| \leq \|K_1\| \cdot \|K_2\|$,

and

(iii) $\|K_1 + K_2\| \leq \|K_1\| + \|K_2\|$.

Finally we consider inverses.

Definition 2.5

The linear operator K is said to have an inverse if, for every $t \epsilon T$, the equation

$$Ks = t, \quad s \epsilon S, \tag{2.6}$$

has a unique solution s, i.e. if K is a 1-1, onto operator. The operator which takes each $t \epsilon T$ into the respective solution

of (2.6) is called the inverse of K, or K^{-1}.

In Banach spaces, the operator K^{-1}, if it exists, has the following properties [9]:

(i) K^{-1} is linear,

and

(ii) if K is bounded, then K^{-1} is bounded.

We conclude our discussion of operators with the following theorem which is a special case of a more general theorem proven in [9].

Theorem 2.1

Let I be the identity operator and K_1 be any bounded operator such that

$$\| K_1 \| < 1. \tag{2.7}$$

Then the operator $K = I + K_1$ has an inverse.

In Chapters IV and V we will consider particular linear operators on R^n and $C[0, 1]$.

We now want to consider the approximation of a definite integral by a weighted sum over the interval, i.e.

$$\int_0^1 x(t)dt \approx \sum_{i=1}^{n} \omega_{ni} x(t_{ni}), \quad n \geq 1, \tag{2.8}$$

where the t_{ni} are abscissas belonging to the interval $[0, 1]$ and

the ω_{ni} are real weights associated with the t_{ni}. The notation indicates that the t_{ni} and ω_{ni} depend upon the number n of abscissas used. The formula used in a particular case may depend upon considerations such as the form of the function x, the accuracy required, and the computational tool available (i.e. calculator, computer, etc.).

In later work we will require that the error in (2.8) converge to zero as n becomes large. This brings up the question, under what conditions will this be true? The following theorem from Berezin and Zhidkov [3] will help answer that.

Theorem 2.2

The necessary and sufficient conditions for

$$\sum_{i=1}^{n} \omega_{ni} x(t_{ni}) \rightarrow \int_0^1 x(t)dt \quad \text{as} \quad n \rightarrow \infty, \quad x \in C[0,1]. \tag{2.9}$$

are that this occurs for any polynomial and that

$$\sum_{i=1}^{n} |\omega_{ni}| \leq M < +\infty, \quad n \geq 1. \tag{2.10}$$

For a proof of this theorem see [3]. The second condition is automatically satisfied when the ω_{ni} are positive.

Now let us examine some particular formulas that have the simplifying feature of equal weights for all abscissas. This feature is helpful in the analysis of roundoff error in applying the formula.

The first to be considered is the repeated midpoint rule on the interval $[0, 1]$. This formula has abscissas $t_{ni} = (2i-1)/2n$ and weights $\omega_{ni} = 1/n$, $i = 1, 2, \cdots, n$. For functions having a continuous first derivative satisfying $|x'(t)| \leq L_1$, for $0 \leq t \leq 1$, it can be shown that

$$\left| \int_0^1 x(t)dt - \frac{1}{n} \sum_{i=1}^{n} x(t_{ni}) \right| \leq L_1/4n . \tag{2.11}$$

Also, it can be shown that L_1 can be replaced by a Lipschitz constant for x, i.e. a number L such that

$$|x(s)-x(t)| \leq L|s-t|, \quad s,t \in S. \tag{2.12}$$

For functions having a continuous second derivative satisfying $|x''(t)| \leq L_2$, for $0 \leq t \leq 1$, we can find that

$$\left| \int_0^1 x(t)dt - \frac{1}{n} \sum_{i=1}^{n} x(t_{ni}) \right| \leq L_2/24n^2. \tag{2.13}$$

As in the previous case, we can replace L_2 by a Lipschitz constant for x'. Krylov [11] shows that (2.11) is the minimum error

bound attainable under the conditions specified with any abscissas and weights.

The other formula to be considered is due to Chebychev. Again, this formula has equal weights. The formula exists only for $n = 1, 2, \cdots, 7, 9$. It takes the form

$$\int_{-1}^{1} x(t)dt = \frac{2}{n} \sum_{i=1}^{n} x(t_{ni}) + E_n, \tag{2.14}$$

where

$$E_n = \begin{cases} C_n x^{(n+1)}(\xi)/(n+1)! & \text{for } n \text{ odd,} \\ C_n x^{(n+2)}(\xi)/(n+2)! & \text{for } n \text{ even, and } \xi \epsilon [-1,1]. \end{cases}$$

The abscissas t_{ni} and constants C_n, $n \leq 6$, along with a thorough discussion of this formula, are given in Hildebrand [8]. Although this formula lacks the convenience of existing for all n, it can be generalized by subdividing the interval and applying the formula for $n = 1, 2, \cdots, 7,$ or 9 to each subdivision. In later work we use the Chebychev 5 point rule repeated r times on the interval $[0, 1]$. This has the form

$$\int_{0}^{1} f(s)ds = \frac{1}{m} \sum_{j=1}^{r} \{ \sum_{i=1}^{5} f(t_{ij}^{*}) \} + E'_m, \tag{2.15}$$

where

$$E'_m = \frac{40,625 f^{(6)}(\xi)}{13,934,592 m^6}, \quad \xi \in (0,1), \tag{2.16}$$

$$t^*_{ij} = (t_{5i} + 2j - 1)/2r, \quad 1 \le i \le 5, \quad 1 \le j \le r,$$

and

$$m = 5r.$$

In some cases we do not have a sufficient number of derivatives to use the above error term. If we assume that the function f to be integrated has a Lipschitz constant L on $[0,1]$, we can show that the error term in (2.15) satisfies

$$|E'_m| \le .254L/m. \tag{2.17}$$

We note that this is only slightly different from the bound in (2.11). From the given error terms and the fact that derivatives of polynomials are bounded, it can be seen that the conditions of Theorem 2.2 are satisfied by the above formulas. This concludes our discussion of quadrature formulas.

III. ROUNDOFF ERROR

In this chapter we consider the rounding errors that occur in the arithmetic operations performed by the computer (CDC 3300) and methods of accounting for them in the solution of problems. Much additional general information on rounding errors can be found in Wilkinson [16]. Two general methods will be given for producing answers and their associated error bounds. The first method is that of bounding the error in each operation and keeping a tabulation of the accumulated error at each stage in the computation. The second method involves the use of automatic interval arithmetic. In the first method we will develop bounds for each operation used and then show how these errors propagate.

We begin with a brief discussion of the floating point hardware characteristics for the CDC 3300, which henceforth will be called the computer or the machine. The following information on hardware, and more, can be found in [4] and [5]; however, no discussion of rounding errors is given. This computer is a binary machine having 24 bit words. The internal representation of a floating point number requires two words, with a 36 bit coefficient, an 11 bit exponent, and a sign bit. Several facts are pertinent to our later analysis: (i) the coefficient is normalized $\frac{1}{2} \leq$ coef. < 1, (ii) rounding takes place before normalization, and (iii) truncation

occurs during normalization. We denote the rounding error by E_1, the truncation error by E_2, the total error by E, and the machine version of an exact quantity Q by $\overline{Q}$.

Let us now consider the operation of addition performed on the numbers $A = a \cdot 2^p$ and $B = b \cdot 2^q$, under the assumptions that $A, B \neq 0$, $\frac{1}{2} \leq |a|, |b| < 1$, and $d \geq 0$, where $d = q-p$. We examine the error by cases determined by the quantity d.

Case $d = 0$: Since there is no need for shifting to equalize exponents, there will be no rounding error (i.e. $E_1 = 0$). If the signs of A and B differ, there will be no normalization error (i.e. $E_2 = 0$). If the signs agree, depending on whether the bit shifted off in normalization is 0 or 1, $E_2 = 0$ or $E_2 = -(\text{sign } A) \cdot 2^{-36+q}$. Thus the error bound for this case is

$$|E| = |E_2| \leq 2^{-36+q} = 2^{-36}|B|/|b| \leq 2^{-35}|B|.$$

Case $d = 1$: Since A must be shifted 1 bit to equalize exponents, $E_1 = 0$ or $E_1 = (\text{sign } A) \cdot 2^{-37+q}$. Again, if $AB < 0$, $E_2 = 0$; but if $AB > 0$, $E_2 = 0$ or $E_2 = -(\text{sign } A) \cdot 2^{-36+q}$. Thus the error bound for this case is

$$|E| = |E_1 + E_2| \leq 2^{-36+q} \leq 2^{-35}|B|.$$

Case $d \geq 2$: Again, since A must be shifted d bits, $E_1 = (\text{sign } A)e$ with

$$-(1-2^{-d+1})2^{-37+q} \leq e \leq 2^{-37+q}.$$

If $AB < 0$, $E_2 = 0$; but if $AB > 0$, $E_2 = 0$ only when $|b| < 1-2^{-d}$ and $|E_2| \leq 2^{-36+q}$ otherwise. Thus for this case it can be shown

$$|E| = |E_1 + E_2| < 2^{-35}|B|.$$

In general, we can prove that the error in the addition of A and B satisfies

$$|E| \leq 2^{-35}\text{Max}\{|A|, |B|\}. \qquad (3.1).$$

The same result holds true for subtraction.

For the operation of multiplication on the above A and B, we have $AB = ab \cdot 2^{p+q}$, with $\frac{1}{4} \leq |ab| < 1$. It can be seen that $|E_1| \leq 2^{-37+p+q}$ and $E_2 = 0$. Thus the error bound for multiplication satisfies

$$|E| \leq 2^{-35}|\overline{AB}|. \qquad (3.2)$$

Similarly, it can be shown that the error bound for division satisfies

$$|E| \leq 2^{-35}|\overline{A/B}|. \qquad (3.3)$$

In certain key places in calculations it is sometimes advantageous to use a higher precision arithmetic. Hence we comment also on the CDC 3300 triple precision floating point, DFP(3). The internal representation requires three words, with 47 bits for the coefficient, 24 bits for the exponents, and a sign bit. DFP(3) uses a normalized coefficient, $\frac{1}{2} \leq$ coef. < 1, but has no rounding feature. For additional information see [4]. By an analysis similar to that for floating point, without rounding, one can find the following bounds for the operations:

$$\begin{aligned} &\text{Addition(Subtraction)} \quad - \quad |E| \leq 2^{-45} \text{Max}\,\{|A|, |B|\}, \\ &\text{Multiplication(Division)} - \quad |E| \leq 2^{-45}\, |\overline{AB}| \ (\text{or} \ |\overline{A/B}|). \end{aligned} \tag{3.4}$$

In conversion from DFP(3) to floating point there is no rounding, hence the error bound for conversion of A satisfies

$$|E| \leq 2^{-35}|A|. \tag{3.5}$$

Now let us apply these results to the problem of bounding the error in a compound operation. Let A and B be numbers which are subject to error and let e(Q) denote the absolute value of the error in the quantity Q. In the following analysis we will use the following rules which are easily found using the methods of Wilkinson [16] and our previous work:

$$\text{(i)}\quad e(A \pm B) \leq e(A) + e(B) + 2^{-35}\text{Max}\{|A|, |B|\},$$

$$\text{(ii)}\quad e(AB) \leq |A|e(B) + |B|e(A) + e(A)e(B) + 2^{-35}|\overline{AB}|, \qquad (3.6)$$

and

$$\text{(iii)}\quad e(A/B) \leq \{e(A) + |A|e(B)/|B|\}/\{|B| - e(B)\} + 2^{-35}|\overline{A/B}|,$$

assuming that $|B| > e(B)$.

We consider first the problem of bounding the error in an inner product calculation. Let $H = \{h_i\}$ and $R = \{r_i\}$ be n-dimensional vectors stored in the computer. Let $Q_i = \overline{h_i r_i}$ be the machine value of $h_i r_i$, $S_i = \overline{\sum_{\ell=1}^{i} Q_\ell}$ be the machine value of the partial sum, and $e(S_i)$ be an error bound on the partial sum. Then we find

$$e(S_1) \leq 2^{-35}|Q_1|,$$

$$e(S_i) \leq e(S_{i-1}) + 2^{-35}|Q_i| + 2^{-35}\text{Max}\{|Q_i|, |S_{i-1}|\},$$

$$1 < i \leq n,$$

and hence

$$e(S_n) \leq 2^{-35}\sum_{i=1}^{n}|Q_i| + 2^{-35}\sum_{i=1}^{n-1}\text{Max}\{|S_i|, |Q_{i+1}|\}. \qquad (3.7)$$

A quantity needed in Chapter IV is a bound on the maximum row sum of an $n \times n$ matrix $P = \{p_{ij}\}$, where the p_{ij} are

subject to error $e(p_{ij})$. For row i, $1 \le i \le n$, let $S_\ell^{(i)} = \sum_{j=1}^{\ell} p_{ij}$ be the machine value of the partial sum and $e(S_\ell^{(i)})$ be the associated error bound. Then we have, similar to above,

$$e(S_1^{(i)}) = e(p_{i1}),$$

$$e(S_\ell^{(i)}) \le e(S_{\ell-1}^{(i)}) + e(p_{i\ell}) + 2^{-35}\text{Max}\{|S_{\ell-1}^{(i)}|, |p_{i\ell}|\},$$

$$1 < \ell \le n,$$

and hence

$$e(S_n^{(i)}) \le \sum_{j=1}^{n} e(p_{ij}) + 2^{-35} \sum_{\ell=1}^{n-1} \text{Max}\{|S_\ell^{(i)}|, |p_{i\ell+1}|\}. \tag{3.8}$$

Hence we can then write for the maximum row sum, Q,

$$Q \le \text{Max}\{S_n^{(i)} : 1 \le i \le n\} + \text{Max}\{e(S_n^{(i)}) : 1 \le i \le n\}. \tag{3.9}$$

We note that the error in calculating error bounds must be accounted for in order to keep the bounds rigorous. Conveniently, multiplication by powers of 2 on positive quantities result in answers that are at least as large as the exact product. Also, addition

of positive quantities may be adjusted to preserve rigorous bounds by forcing a round up at each step of the error calculation.

The second method of bounding the error in a series of operations involves the use of interval arithmetic, as discussed in the work of R. E. Moore [12]. This method uses a pair of numbers $A = [a, b]$ to represent the lower and upper bounds, respectively, of a compact interval containing an exact but unknown number c. For any arithmetic operation $\circ$ and intervals A, B we define the compact interval $A \circ B$ by

$$A \circ B = \{c \circ d : c \in A,\ d \in B\}.$$

At each stage of a computation the partial result is an interval guaranteed to contain the answer at that stage. A set of interval arithmetic operations has been implemented for the CDC 3300 at the OSU Computer Center. For details see Computer Center Report 67-1 [14]. These are easy to use and allow production of rigorous interval answers without significant complication of a program. Although this method is much easier to implement, both of the methods discussed here have been found to be practical ways of producing rigorous error bounds on computed quantities.

IV. FINITE DIMENSIONAL LINEAR OPERATOR INVERSION

In this chapter we consider linear operators over finite dimensional normed linear spaces. Suppose we have such a space S with scalar field R and basis $e_1, e_2, \cdots, e_n$. As noted before, we will primarily be interested in the space R^n with norm (2.2) and the usual basis

$$\{e_j = (\delta_{1j}, \delta_{2j}, \cdots, \delta_{nj}): \quad 1 \leq j \leq n\}.$$

Consider the linear operator K on S and the elements of S, Ke_i, $1 \leq i \leq n$. Let us write these in terms of the original basis,

$$\begin{aligned} Ke_1 &= a_{11}e_1 + a_{21}e_2 + \cdots + a_{n1}e_n \\ Ke_2 &= a_{12}e_1 + a_{22}e_2 + \cdots + a_{n2}e_n \\ &\vdots \\ Ke_n &= a_{1n}e_1 + a_{2n}e_2 + \cdots + a_{nn}e_n, \quad a_{ij} \in R . \end{aligned} \tag{4.1}$$

Now consider the $n \times n$ matrix

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} . \qquad (4.2)$$

It is shown in Faddeeva [6] that this matrix uniquely defines the linear operator K, and each such $n \times n$ matrix defines a linear operator. Thus we will consider the problem of inversion of finite dimensional linear operators synonymous with the problem of inversion of $n \times n$ matrices.

Recalling Definition 2.4, it can be seen that the choice of vector norm induces a corresponding choice of norm for matrices. In Faddeeva [6] it is shown that the norm (2.2) for vectors induces the following norm on matrices

$$\|A\| = \text{Max} \{ \sum_{j=1}^{n} |a_{ij}| : 1 \leq i \leq n \}. \qquad (4.3)$$

Definition 2.5 yields a definition for the inverse of a matrix. We shall now discuss finding the inverse of a matrix by the Hotelling Method. This method is discussed in Faddeeva [6] under the heading "Correction of the Elements of an Inverse." This is an iterative

method which easily gives a measure of the accuracy of the result.

Let us consider matrix A for which we want the inverse A^{-1}, and let B_0 be an approximation to A^{-1}. The first step is to form the matrix

$$S_0 = I - AB_0 . \tag{4.4}$$

Since $B_0 \approx A^{-1}$ it is reasonable to assume $\|S_0\| < 1$. Now if we write (4.4) as

$$AB_0 = I - S_0 \tag{4.5}$$

and apply Theorem 2.1, we see that $\|S_0\| < 1$ implies that AB_0 has an inverse. Then from

$$(AB_0)(AB_0)^{-1} = A(B_0(AB_0)^{-1}) = I \tag{4.6}$$

we see that A has a right inverse. For matrices this is sufficient to guarantee the existence of A^{-1}. The following method will result in a series of approximations to A^{-1} which will converge to A^{-1}. Form the sequence of matrices

$$B_i = B_{i-1}(I+S_{i-1}), \quad S_i = I-AB_i, \quad 1 \leq i \leq j. \tag{4.7}$$

It is shown in Faddeeva [6] that $S_j = S_0^{2^j}$ and hence

$$B_j = A^{-1}(I-S_0^{2^j}) . \qquad (4.8)$$

Then by the assumption $\|S_0\| < 1$, $B_j \to A^{-1}$. From this Faddeeva [6] derives

$$\|B_j-A^{-1}\| \leq \|B_0\| \|S_0\|^{2^j}/(1-\|S_0\|). \qquad (4.9)$$

Hence, theoretically, we can approximate A^{-1} to any degree of accuracy desired.

Using the method described, it is possible, beginning with only a reasonable approximation, to conclude the existence of the inverse for a given matrix, get a better approximation to the inverse, and bound the discrepancy between the inverse and its approximation. In order to do these things we need rigorous bounds for $\|B_0\|$ and $\|S_0\|$.

We consider B_0 to be exact, and from Chapter III we have a method for getting a rigorous bound for $\|B_0\|$. It remains then to find a rigorous bound for $\|S_0\|$, where S_0 is as defined in (4.4). Observe that the calculation of each element of AB_0 is just an inner product calculation as discussed in Chapter III. Hence finding $\|S_0\|$ is a simple application of techniques from a previous discussion. Recall that for $I_{ij} = \overline{\sum_{\ell=1}^{n} a_{i\ell} \cdot b_{\ell j}}$, we have an expression for the error

$$e(I_{ij}) \leq 2^{-35}\{\sum_{\ell=1}^{n} \overline{|a_{i\ell} \cdot b_{\ell j}|} + \sum_{\ell=1}^{n-1} \text{Max}\{|\overline{\sum_{m=1}^{\ell} a_{im} \cdot b_{mj}}|, \overline{|a_{i\ell+1} \cdot b_{\ell+1j}|}\}\}, \tag{4.10}$$

where $A = \{a_{ij}\}$ and $B_0 = \{b_{ij}\}$. The elements of S_0 on the diagonal have additional error due to the modification by the identity matrix, hence

$$e(1-I_{ii}) \leq e(I_{ii}) + 2^{-35} \text{Max}\{|I_{ii}|, 1\}. \tag{4.11}$$

We now have a bound for the error in each element and can now apply the previously mentioned method to get a bound for $\|S_0\|$. It remains only to combine these results rigorously using (4.9) to get a bound for $\|B_j - A^{-1}\|$.

As mentioned in Chapter III, there is another approach to the problem of rigorous bounds - interval arithmetic. A technique for finding matrix inverses rigorously using both interval arithmetic and the Hotelling method is discussed by Hansen [7]. We will outline the method from that paper, giving little justification. The following definitions are needed:

Definition 4.1

An interval matrix, indicated by a superscript I, is a matrix of compact intervals.

We note that with each real matrix $M = \{m_{ij}\}$ we can associate, in a natural way, an interval matrix $M^I = \{[m_{ij}, m_{ij}]\}$.

Definition 4.2

Consider $N = \{n_{ij}\}$, $M^I = \{[p_{ij}, q_{ij}]\}$, and $T^I = \{[u_{ij}, v_{ij}]\}$. We define the following relations:

(i) $N \prec M^I$ iff $n_{ij} \in [p_{ij}, q_{ij}]$ for $1 \leq i, j \leq n$,

and

(ii) $M^I \prec T^I$ iff $[p_{ij}, q_{ij}] \subset [u_{ij}, v_{ij}]$ for $1 \leq i, j \leq n$.

Definition 4.3

We define the inverse of A^I by

$$(A^I)^{-1} = \{a_{ij}\}, \tag{4.12}$$

where

$$a_{ij} = \{b_{ij} : A \prec A^I, \ A^{-1} = \{b_{ij}\}\}, \tag{4.13}$$

if A^{-1} is defined for all $A \prec A^I$.

It can be shown that the a_{ij} are compact intervals.

Definition 4.4

Consider $M^I = \{[p_{ij}, q_{ij}]\}$. We define the norm of M^I by

$$\|M^I\| = \text{Max}\ \{\ \|M\| : M \prec M^I\}, \tag{4.14}$$

or, equivalently,

$$\|M^I\| = \text{Max}\{\sum_{j=1}^{n} \text{Max}\{|p_{ij}|, |q_{ij}|\}: 1 \le i \le n\}. \tag{4.15}$$

Definition 4.5

With M^I as above, the width, $W(M^I)$, of an interval matrix M^I is $\|\{|q_{ij}-p_{ij}|\}\|$.

Definition 4.6

With M^I as above, we define the center of M^I, M^c, by

$$M^c = \{m_{ij} : m_{ij} = (p_{ij} + q_{ij})/2\}. \tag{4.16}$$

We are now ready to discuss the method given by Hansen [7] for approximating the inverse $(A^I)^{-1}$. Actually, it is impossible to find $(A^I)^{-1}$ exactly because of roundoff error, but we can find a matrix C^I such that $(A^I)^{-1} \prec C^I$. First, find an approximation B to $(A^c)^{-1}$. Let I^I and B^I be the interval matrices associated with I and B. Calculate

$$E^I = I^I - A^I B^I, \tag{4.17}$$

and find a bound for $\|E^I\|$ using (4.15). If $\|E^I\| \le E_0 < 1$, we are assured that the inverse $(A^I)^{-1}$ exists and may proceed to find C^I. Let

$$S^I = I^I + E^I + (E^I)^2 + \cdots + (E^I)^{\ell}. \tag{4.18}$$

It can be shown that

$$\| (I-E^I)^{-1} - S^I \| \leq \| E^I \|^{\ell+1} / (1-\| E^I \|) = b(\ell). \tag{4.19}$$

Choose ℓ so that

$$b(\ell) \approx \text{Min} \{ | e_{ij} - d_{ij} | : 1 \leq i, j \leq n, \quad E^I = \{ [d_{ij}, e_{ij}] \} , \tag{4.20}$$

and calculate S^I using (4.18). Now define P^I to be the matrix having all elements equal to $[-b(\ell), b(\ell)]$ and calculate

$$C^I = B^I(S^I + P^I), \tag{4.21}$$

which satisfies $(A^I)^{-1} \prec C^I$.

As mentioned in [7], for actual computation it is best to rewrite (4.18) and (4.21) in the form

$$G^I = E^I(I^I + E^I(I^I + \cdots + E^I) \cdots) = S^I - I^I, \tag{4.22}$$

and

$$C^I = B^I + B^I(G^I + P^I) . \tag{4.23}$$

Notice also that P^I need not be formed as a matrix but instead we may form $G^I + P^I$ by direct modification of G^I. These techniques allow us to save space and get narrow, still rigorous, intervals.

We conclude this chapter with some comments on the

precision of the methods outlined above. If we perform the calculation of (4.7) in the form

$$B_i = B_{i-1} + B_{i-1}S_{i-1}, \tag{4.24}$$

it is apparent that we are just making a correction to our current approximation. This correction should become smaller with further iteration. Experience indicates that eventually the order of magnitude of the corrections approaches that of roundoff error and further refinement is impossible. Obviously, it is advantageous for us to calculate S_i as accurately as possible, using higher precision arithmetic if available. In light of the above observation, we make the rule that iteration should be stopped if

$$\|\overline{S_i}\| \geq \|\overline{S_{i-1}}\|. \tag{4.25}$$

We briefly consider the width, $W(C^I)$, of C^I in (4.23). Part of $W(C^I)$ is caused by $W(G^I)$ and $W(P^I)$. Also, the operations of addition and multiplication contribute to $W(C^I)$. Our experience indicates that $W(G^I)$ and $W(P^I)$ are the primary contributors to $W(C^I)$. With the proper choice of ℓ in (4.20) we can minimize $W(P^I)$. Also, both $W(G^I)$ and $W(P^I)$ are affected by the width of E^I. Hence, we find that it is advantageous to calculate E^I in a manner that introduces as little interval width

as possible and still preserve rigor. Here it would be beneficial to use higher precision interval arithmetic. However, at the present time this is unavailable.

V. FREDHOLM INTEGRAL OPERATOR INVERSION - PART I

In this chapter we will consider inversion of the Fredholm integral operator of the second kind. Much of the notation and results in this chapter are from [1] and [2].

Let V be the interval $[0, 1]$. We consider the Fredholm integral equation of the second kind

$$x(s) - \int_0^1 k(s,t)x(t)dt = y(s), \quad s \in V, \tag{5.1}$$

where x, y, and k are continuous on V, V, and $V \times V$ respectively and Riemann integration is used.

Definition 5.1

Define the integral operator $K : C(V) \to C(V)$ by

$$(Kf)(s) = \int_0^1 k(s,t)f(t)dt, \quad s \in V, \ f \in C(V). \tag{5.2}$$

We will require that k be continuous on $V \times V$ as above and, in addition, that there exists a real Q satisfying

$$\text{Max } \{|k(s,t)| : s,t \in V\} = Q. \tag{5.3}$$

Then, using norm

$$\|f\| = \text{Max}\{|f(s)| : s \in V\} \tag{5.4}$$

on $C(V)$, the operator K is bounded,

$$\|K\| = \text{Sup}\{\|Kf\| / \|f\| : f \neq 0\} \leq \text{Max}\{\int_0^1 |k(s,t)|dt : s \in V\} \leq Q. \tag{5.5}$$

We can now write (5.1) in operator form

$$(I - K)x = y . \tag{5.6}$$

On the computer, we will use the following approximation for (5.1):

$$x_n(s) - \sum_{i=1}^{n} \omega_{ni} k(s, t_{ni}) x_n(t_{ni}) = y(s), \quad n \geq 1, \quad s \in V, \tag{5.7}$$

where the ω_{ni} and t_{ni} are weights and abscissas for one of the quadrature formulas on V discussed in Chapter II.

Definition 5.2

Define the operator $K_n : C(V) \to C(V)$ by

$$(K_n f)(s) = \sum_{i=1}^{n} \omega_{ni} k(s, t_{ni}) f(t_{ni}), \quad n \geq 1, \quad s \in V, \quad f \in C(V). \tag{5.8}$$

We will require the ω_{ni} and t_{ni} to satisfy

$$\text{(i)} \quad \sum_{i=1}^{n} \omega_{ni} f(t_{ni}) \rightarrow \int_0^1 f(t)dt \quad \text{as} \quad n \rightarrow \infty, \quad f \in C(V),$$

and (5.9)

$$\text{(ii)} \quad \sum_{i=1}^{n} |\omega_{ni}| \leq B < +\infty, \quad n \geq 1 .$$

Then the operators $\{K_n\}$ are uniformly bounded,

$$\|K_n\| = \text{Max} \{ \sum_{i=1}^{n} |\omega_{ni}| \cdot |k(s, t_{ni})| : s \in V \} \leq QB. \tag{5.10}$$

We can now write (5.7) as

$$(I - K_n)x_n = y. \tag{5.11}$$

We are now prepared to discuss a method for inverting the operator $(I-K_n)$ and getting particular solutions to (5.7). Anselone and Moore [2] discuss a method of solving for particular solutions of (5.7). They also give a test for the existence of a solution for (5.1) based on existence of a solution for (5.7), and they derive a bound for $\|x-x_n\|$. We will discuss their justification of the method and their results. Later we will show how partial results of this method can be used to create in the computer an operator which approximates $(I-K)^{-1}$.

To solve (5.7), the first step is to solve the linear system

$$x_n(t_{nj}) - \sum_{i=1}^{n} \omega_{ni} k(t_{nj}, t_{ni}) x_n(t_{ni}) = y(t_{nj}), \quad n \geq 1, \quad 1 \leq j \leq n . \tag{5.12}$$

If the solution $\{x_n(t_{ni}), \quad 1 \leq i \leq n\}$ exists, we can then get the solution x_n of (5.7) by

$$x_n(s) = y(s) + \sum_{i=1}^{n} \omega_{ni} k(s, t_{ni}) x_n(t_{ni}), \quad n \geq 1, \quad s \in V. \tag{5.13}$$

This solution x_n may or may not be a good approximation to the solution x of (5.1).

Anselone and Moore [2] find the following results which justify the use of the method just presented. These results will be presented without proof.

Lemma 5.1

Let K, K_n, $n \geq 1$, be the bounded linear operators given in Definitions 5.1 and 5.2. Then

$$\| K_n f - Kf \| \to 0 \quad \text{as} \quad n \to \infty, \qquad f \in C(V),$$

however,

$$\| K_n - K \| \not\to 0 \quad \text{as} \quad n \to \infty, \quad \text{if} \quad k(s,t) \not\equiv 0. \tag{5.14}$$

Theorem 5.2

Again, let K, K_n, $n \geq 1$, be as above. Then

$$\| K_n K - K^2 \| \to 0 \quad \text{as} \quad n \to \infty. \tag{5.15}$$

The next theorem is very important as it provides a test for the existence of a solution for (5.1) given the results of solving (5.7). It is stated here without proof and in less generality than in [2].

Theorem 5.3

Let K, K_n, $n \geq 1$, be as above. Suppose $(I-K_n)^{-1}$ exists and satisfies

$$\| K_n K - K^2 \| < 1 / \| (I-K_n)^{-1} \|. \tag{5.16}$$

Then $(I-K)^{-1}$ exists and

$$\| (I-K)^{-1} \| \leq \{ 1 + \| (I-K_n)^{-1} \| \cdot \| K \| \} / \{ 1 - \| (I-K_n)^{-1} \| \cdot \| K_n K - K^2 \| \}. \tag{5.17}$$

The question now arises, how close is the solution of (5.7) to the solution of (5.1) The next theorem from [2] will help answer that question. Again, it is stated without proof and in less generality than in [2].

Theorem 5.4

Let K, K_n, $n \geq 1$, be as above. Assume that $(I-K)^{-1}$ and $(I-K_n)^{-1}$ exist and that (5.6) and (5.11) hold. Then

$$\|x-x_n\| \leq \|(I-K_n)^{-1}\| \{\|K_n y-Ky\| + \|K_n K-K^2\| \cdot \|x\|\}, \tag{5.18}$$

and if (5.16) holds,

$$\|x-x_n\| \leq \|(I-K_n)^{-1}\| \{\|K_n y-Ky\| + \|K_n K-K^2\| \cdot \|x_n\|\} / \{1-\|K_n K-K^2\| \cdot \|(I-K_n)^{-1}\|\}, \tag{5.19}$$

and

$$\|x-x_n\| \to 0 \quad \text{as} \quad n \to \infty. \tag{5.20}$$

From the results just presented we can, using information from solving (5.7), test for the existence of a solution to (5.1). If such a solution exists, we can also bound the quantity $\|x-x_n\|$.

We now consider a way of creating an operator in the computer which approximates $(I-K)^{-1}$. In the process of solving the linear system (5.12) let us define the matrix $\overline{K_n} = \{m_{ij}\}$, where

$$m_{ij} = \omega_{nj} k(t_{ni}, t_{nj}). \tag{5.21}$$

We can then write the solution of (5.12)

$$\bar{x}_n = (I-\bar{K}_n)^{-1}\bar{y} \tag{5.22}$$

where $\bar{x}_n = (x_n(t_{n1}), x_n(t_{n2}), \cdots, (x_n(t_{nn}))$ and $\bar{y} = (y(t_{n1}), y(t_{n2}), \cdots, y(t_{nn}))$. Now let us give the following definitions from Anselone [1]:

Definition 5.3

Define operator $\psi_n : C(V) \to R^n$ by

$$\psi_n(f) = (f(t_{n1}), f(t_{n2}), \cdots, f(t_{nn})) \tag{5.23}$$

for $f \in C(V)$.

It is easily seen that $\|\psi_n\| = 1$.

Definition 5.4

Define operator $\phi_n : R^n \to C(V)$ by

$$(\phi_n \bar{v}_n)(s) = \sum_{i=1}^{n} \omega_{ni} k(s, t_{ni}) v_n(t_{ni}) \tag{5.24}$$

for $\bar{v}_n = (v_n(t_{n1}), v_n(t_{n2}), \cdots, v_n(t_{nn})) \in R^n$.

It can be shown that $\|\phi_n\| = \|K_n\|$. Also, it can be shown that

$$(I-K_n)^{-1} = I + \phi_n(I-\bar{K}_n)^{-1}\psi_n. \tag{5.25}$$

From this it is obvious that

$$\| (I-K_n)^{-1} \| \leq 1 + \| K_n \| \cdot \| (I-\overline{K}_n)^{-1} \| . \qquad (5.26)$$

Since the operators ϕ_n and ψ_n can be simulated in the computer, and we have $(I-\overline{K}_n)^{-1}$ during the solution of (5.7), we have in the computer the necessary information for forming an operator that approximates $(I-K)^{-1}$.

In the preceding discussion we have ignored the roundoff error in the solution of (5.7). We now consider the problem of bounding it. The groundwork for this has been done in Chapters III and IV. From Chapter IV we get a bound for the roundoff and truncation error in inverting the matrix $(I-\overline{K}_n)$. This, combined with the analysis of the inner product calculation given in Chapter III, will yield a bound on the error in each $\overline{x_n(t_{ni})}$, $1 \leq i \leq n$, denoted by $e(\overline{x_n(t_{ni})})$. As before we denote the machine version of the exact quantity Q by $\overline{Q}$. Assuming that we have bounds for the error in $\overline{y(s)}$, denoted by $e(\overline{y(s)})$, and in $\overline{k(s,t_{ni})}$, denoted by $e(\overline{k(s,t_{ni})})$, $1 \leq i \leq n$, we are ready to complete the calculation of an error bound on the solution to (5.7). Let $T_i = \overline{\overline{k(s,t_{ni})}\,\overline{x_n(t_{ni})}}$ and assume that $\omega_{ni} = 1/n$, for all $i \leq n$. It follows from (3.6) that the error in $\overline{x_n(s)}$, denoted by $e(\overline{x_n(s)})$, satisfies

$$e(\overline{x_n(s)}) \leq e(\overline{y(s)}) + e(\overline{\frac{1}{n}\sum_{i=1}^{n} T_i}) + 2^{-35}\,\mathrm{Max}\{|\overline{y(s)}|, |\overline{\frac{1}{n}\sum_{i=1}^{n} T_i}|\}, \tag{5.27}$$

where

$$e(\overline{\frac{1}{n}\sum_{i=1}^{n} T_i}) \leq e(\overline{\sum_{i=1}^{n} T_i})/n + 2^{-35}\,|\overline{\frac{1}{n}\sum_{i=1}^{n} T_i}|, \tag{5.28}$$

and $e(\overline{\sum_{i=1}^{n} T_i})$ is evaluated as in the inner product analysis of Chapter III. The result (5.27) and the result (5.19) can be combined to give a bound for the total discrepancy $D(s)$ between the computed solution for (5.7) and the solution for (5.1) at s,

$$D(s) \leq \|x-x_n\| + e(\overline{x_n(s)}). \tag{5.29}$$

Thus if we have $\|K_n\|$ and $\|K_nK-K^2\|$, we can get $\|(I-\overline{K}_n)^{-1}\|$ during the solution of (5.7) and test for the existence of a solution for (5.1). If, in addition, we have $\|y\|$ and $\|K_ny-Ky\|$, we can find $e(\overline{x_n(s)})$ and put an interval around $\overline{x_n(s)}$ in which $x(s)$ is guaranteed to lie. It is usually possible to obtain the quantities $\|y\|$, $\|K_n\|$, $\|K_ny-Ky\|$, and $\|K_nK-K^2\|$. In the next chapter we will discuss problems related to obtaining them and implementing the method described.

In the method described above, the operator $(I-K_n)^{-1}$ was used as an approximation for $(I-K)^{-1}$. Anselone and Moore [2] suggest that a more appropriate approximation to $(I-K)^{-1}$ might be $(I + (I-K_n)^{-1}K)$. In the following theorem we find a result analogous to Theorem 5.4.

Theorem 5.5

Let K, K_n, $n \geq 1$, be the bounded linear operators of Definitions 5.1 and 5.2. Assume that $(I-K)^{-1}$ and $(I-K_n)^{-1}$ exist and that $x = (I-K)^{-1}y$ and $x_n = (I+(I-K_n)^{-1}K)y$. Then

$$\|x-x_n\| \leq \|(I-K_n)^{-1}\| \, \|K_nK-K^2\| \, \|x\|, \tag{5.30}$$

from which we get

$$\|(I-K)^{-1} - (I + (I-K_n)^{-1}K)\| \to 0 \quad \text{as} \quad n \to \infty. \tag{5.31}$$

Also, if (5.16) holds,

$$\|x-x_n\| \leq \|(I-K_n)^{-1}\| \, \|K_nK-K^2\| \, \|x_n\| / \{1-\|(I-K_n)^{-1}\| \, \|K_nK-K^2\|\}, \tag{5.32}$$

and

$$\|x-x_n\| \to 0 \quad \text{as} \quad n \to \infty. \tag{5.33}$$

Proof: This proof is quite similar to the proof given in [2] for the

generalization of Theorem 5.4. Observe that

$$(I-K_n+K)(I-K) = I-K_n+K_nK-K^2.$$

Operating on this by $(I-K_n)^{-1}$ on the left and then by $(I-K)^{-1}$ on the right we get

$$(I+(I-K_n)^{-1}K) = (I-K)^{-1} + (I-K_n)^{-1}(K_nK-K^2)(I-K)^{-1}.$$

Hence

$$(I+(I-K_n)^{-1}K)y - (I-K)^{-1}y = (I-K_n)^{-1}(K_nK-K^2)(I-K)^{-1}y,$$

and

$$x_n-x = (I-K_n)^{-1}(K_nK-K^2)x,$$

from which we easily get (5.30). Using

$$\|x\| \leq \|x_n-x\| + \|x_n\|$$

we also get (5.32). (5.31) and (5.33) can be seen from (5.30), (5.32), and (5.15).

If the quantity Ky can be found in some way, we can use the original approach of Anselone and Moore [2] to solve the equation

$$(I-K_n)x_n^* = Ky \tag{5.34}$$

and get x_n by

$$x_n = y + x_n^* . \tag{5.35}$$

In solving (5.34) the roundoff error considerations are exactly the same as given for the original method. Then we need only bound the additional error in (5.35) to get a complete error bound. It has been found experimentally that, if one can find Ky analytically, this alternate method gives better results. For cases where the exact solution was known, the computed solution by this method was closer to the exact solution than the computed solution by the original method. Also, the error bound D(s) in (5.29) was correspondingly smaller for this method.

VI. FREDHOLM INTEGRAL OPERATOR INVERSION - PART II

In the previous chapter we considered, in a theoretical manner, several related methods for solving an approximation of (5.1) for particular solutions and bounding the truncation and rounding error. In addition, we found that we could construct an approximation to the operator $(I-K)^{-1}$ in the computer. Here we consider some of the more practical aspects of converting the methods presented into workable programs. Also included will be some numerical results on two examples.

The ideal program would take the given functions y and k of (5.1) and produce numerical results, including rigorous error bounds, with no additional information. This, however, is very difficult or impossible to do because of problems arising in the bounding of the error. Bounding the roundoff error is easily accomplished using the results of Chapters III and IV. In Chapter V we have bounds on the truncation error in terms of the quantities $\|y\|$, $\|x_n\|$, $\|K_nK-K^2\|$, $\|K_ny-Ky\|$, and $\|(I-K_n)^{-1}\|$. However, the calculation of these quantities require some additional information.

Let us consider the problem of calculating these needed quantities. First, it is easily seen from (5.11), that

$$\|x_n\| \leq \|(I-K_n)^{-1}\| \, \|y\|. \tag{6.1}$$

Using this and (5.26), the list of needed quantities becomes $\|y\|$, $\|K_n\|$, $\|(I-\overline{K}_n)^{-1}\|$, $\|K_n y - Ky\|$, and $\|K_n K - K^2\|$. The results of Chapter IV allow us to bound $\|(I-\overline{K}_n)^{-1}\|$. If we are given L_1, L_2 satisfying

$$|y(s)-y(t)| \leq L_1 |s-t|, \quad 0 \leq s, t \leq 1, \tag{6.2}$$

and

$$|k(u,s)-k(u,t)| \leq L_2 |s-t|, \quad 0 \leq s,t,u \leq 1, \tag{6.3}$$

we can determine

$$\|y\| = \text{Max} \{|y(s)| : 0 \leq s \leq 1\}, \tag{6.4}$$

and

$$\|K_n\| = \text{Max} \left\{ \sum_{i=1}^{n} |\omega_{ni}| \, |k(s,t_{ni})| : 0 \leq s \leq 1 \right\}. \tag{6.5}$$

This is done by evaluating the quantities to be maximized on an equally spaced grid $\{g_i : 1 \leq i \leq m\}$, with spacing δ, and computing

$$\|y\| \leq \text{Max} \{|y(g_i)| : 1 \leq i \leq m\} + L_1 \delta/2, \tag{6.6}$$

and

$$\|K_n\| \leq \text{Max} \left\{ \sum_{i=1}^{n} |\omega_{ni}| \, |k(g_i,t_{ni})| : 1 \leq j \leq m \right\} + L_2 \delta/2. \tag{6.7}$$

The quantities

$$\|K_n y - Ky\| = \text{Max}\{\left|\int_0^1 k(s,t)y(t)dt - \sum_{i=1}^{n} \omega_{ni} k(s,t_{ni})y(t_{ni})\right| : 0 \leq s \leq 1\} \tag{6.8}$$

and

$$\|K_n K - K^2\| = \text{Max}\{\int_0^1 \left| \int_0^1 k(s,t)k(t,u)dt - \sum_{i=1}^{n} \omega_{ni} k(s,t_{ni})k(t_{ni},u)\right| du : 0 \leq s \leq 1\} \tag{6.9}$$

pose a bigger problem. Quantities of the form

$$\left| \int_0^1 g(s,t)dt - \sum_{i=1}^{n} \omega_{ni} g(s,t_{ni})\right|$$

can be bounded using an error term for the quadrature formula. This usually involves bounds on an appropriate derivative of g with respect to t. Although we might increase the bound, we can replace the outside integral in (6.9) with a maximum with respect to $0 \leq u \leq 1$. We have now reduced the problem of finding $\|K_n y - Ky\|$ and $\|K_n K - K^2\|$ to that of finding maxima which was treated above for $\|y\|$ and $\|K_n\|$. We note that with the amount of outside information needed, it may be easier to find the quantities analytically and treat them as input to the program.

From the above discussion it is clear that before rigorous bounds can be obtained there must be considerable prior analysis.

Hence, for our program we require that $\|y\|$, $\|K_n\|$, $\|K_n y - Ky\|$, and $\|K_n K - K^2\|$ be supplied at the outset. If automation is more important than rigor, there is an alternative. While rigorous bounding of the above quantities may be a somewhat formidable task, it is easy to get estimates for them. To do this for $\|y\|$ and $\|K_n\|$ we use the above method with a fine grid and ignore the intervals between grid points. For $\|K_n y - Ky\|$ and $\|K_n K - K^2\|$ we use (6.8) and (6.9), replacing the outside integral in (6.9) by a maximum with respect to $0 \leq u \leq 1$. We approximate the remaining integrals by a high order quadrature formula. Here we can use a much higher order formula because we need not solve an $n \times n$ algebraic system as we do in the original approximation of (5.1). This procedure will yield reasonable estimates to the quantities in question, $\|K_n y - Ky\|$ and $\|K_n K - K^2\|$.

In an attempt to develop a truly useful tool, our program has been designed to give the choice of preparing rigorous bounds for the needed quantities and getting rigorous error bounds, or letting the program automatically estimate these and give reasonable estimates of the error. In addition, choice of either of the two quadrature formulas mentioned in Chapter II is given for use in the solution of the problem.

One final comment should be made. To retain rigor one must consider also errors which may take place in the input of

information into the program. Very little information about such errors has been found. However, in the program an attempt has been made to bound any such errors in order to complete the analysis.

From Tricomi [15] we take the following example,

$$x(s) - .5 \int_0^1 e^{s-t} x(t)dt = 1.0, \tag{6.10}$$

for which we know the solution $x(s) = 1 + e^{x} - e^{x-1}$. Using the methods from the previous discussion or direct computation with the given definitions, we find the following values for the key quantities (using 15 abscissas in $[0,1]$):

Quantity		Repeated Midpoint	Repeated 5 Point Chebychev
$\|K_n K - K^2\|$	$\leq$	0.0	0.0
$\|K\|$	$\leq$	0.8592	0.8592
$\|K_n y - Ky\|$	$\leq$	0.0230	0.0002
$\|y\|$	$\leq$	1.0	1.0 .

Using these values and also estimating them using the methods given above, our program gives us the following error bounds:

$$\|D\| \leq \begin{cases} 7.6 \times 10^{-2} & \text{rigorous} \\ 5.3 \times 10^{-4} & \text{estimated} \end{cases} \qquad \begin{cases} 6.6 \times 10^{-4} & \text{rigorous} \\ 3.1 \times 10^{-8} & \text{estimated} \end{cases} .$$

Repeated Midpoint Repeated Chebychev

The roundoff error, in both cases, was on the order of 2.0×10^{-8}. This is negligible as compared with $\|x-x_n\|$. The largest observed error, based on comparison with the known solution, was about 3.0×10^{-4} using the repeated midpoint rule. For the repeated Chebychev rule, the answers were correctly rounded to the 5 significant digits printed. For the alternate method, using

$$(Ky)(s) = -.05\, e^{s}(e^{-1}-1),$$

$$\|Ky\| \leq 0.8592,$$

and the repeated midpoint rule, the program finds $\|D\| \leq 1.0 \times 10^{-8}$ and the answers were correctly rounded to the 5 significant digits printed.

As a second example, from Kopal [10], we take

$$x(s) - \int_0^1 k(s,t)x(t)dt = \frac{1}{2}\, s(1-s), \qquad (6.11)$$

where

$$k(s,t) = \begin{cases} t(1-s) & 0 \leq t \leq s \\ s(1-t) & s \leq t \leq 1 \end{cases}.$$

The solution is $x(s) = (\tan \frac{1}{2}) \sin s + \cos s - 1$. In this example we find the key quantities to be as follows (again using 15 abscissas in $[0,1]$):

Quantity		Repeated Midpoint	Repeated Chebychev
$\|K_n K - K^2\|$	$\leq$	0.0042	0.00425
$\|K_n\|$	$\leq$	0.25	0.25
$\|K_n y - Ky\|$	$\leq$	0.0105	0.01063
$\|y\|$	$\leq$	0.1250	0.1250

Our program gives the following error bounds:

	Repeated Midpoint		Repeated Chebychev
$\|D\| \leq$	1.45×10^{-2}	rigorous	1.47×10^{-2}
	8.4×10^{-5}	estimated	1.1×10^{-4} .

The roundoff error, in both cases, was on the order of 1.0×10^{-10}. This is negligible as compared with $\|x - x_n\|$. The largest observed error was about 1.5×10^{-4} for the repeated midpoint rule and about 2.0×10^{-4} for the repeated Chebychev rule. For the alternate method, using

$$(Ky)(s) = (s^4 - 2s^3 + s)/24,$$

$$\|Ky\| \leq 0.0131 ,$$

and the repeated midpoint rule, the program finds $\|D\| \leq 7.7 \times 10^{-4}$ and the largest observed error is about 1.0×10^{-5}.

The results from the program using interval arithmetic were

similar to those from the program using the other method given in this paper. This is to be expected since the basic method is quite similar and the roundoff error is usually negligible. To give an idea of the machine time required for the above examples, we give the following times on the first example:

(i) using interval arithmetic - 11 sec.,

(ii) using ordinary arithmetic - 17 sec.

When the program was asked to estimate the key quantities, the ordinary arithmetic solution required 34 sec.

Our examples illustrate the fact, noted in Chapter II, that for functions having only a bounded first derivative or a Lipschitz constant, the repeated midpoint formula is as good as the more sophisticated formulas. Also, we can see that for functions having higher derivatives the repeated midpoint formula is not as satisfactory.

Finally, our examples show that being rigorous in bounding the error does not lead to unnecessarily large bounds. We feel that the bounds found by the methods of this thesis are realistic enough to be useful.

VII. SUMMARY

We have discussed methods for inverting $n \times n$ matrices and linear Fredholm integral operators of the second kind. We have developed techniques which allow us to prove the existence of and find approximations to inverses for the above types of operators using the computer. Also, we were able to bound rigorously the error in the approximations.

The above techniques were implemented in the form of computer programs, and some numerical results were given. It was found that the error bounds resulting from these programs were sufficiently realistic to be of interest and of use.

It was noted that the interval arithmetic technique is much easier to implement than the step by step accumulation of error. Why then do we consider the latter technique? Interval arithmetic became available here only recently and is not widely available. A disadvantage of using interval arithmetic is that higher precision interval arithmetic is not presently available.

In the inversion of Fredholm operators we used quadrature formulas having equal weights at the abscissas because this simplified the error analysis. However, the techniques used in this thesis can be applied to other formulas such as Gauss quadrature which are more precise for smooth functions. In interval arithmetic, the use

of more sophisticated formulas is especially appealing since there is essentially no increase in complexity involved.

We note that the methods and techniques of this thesis might be applied to the solution of nonlinear integral equations using Newton's method. For a theoretical discussion of the use of Newton's method for nonlinear integral equations see [13]. Consider the equation

$$x - K(F(x)) = y , \tag{7.1}$$

where $(F(x))(s) = f(x(s))$, $s \in [0,1]$, and f is a continuous, real valued function of a real variable. Let

$$G(x) = x - K(F(x)) - y . \tag{7.2}$$

By Newton's method we want to solve $G(x) = 0$. Considering

$$x_{i+1} = x_i - (G'(x_i))^{-1} G(x_i), \tag{7.3}$$

using the prime to indicate a Fréchet derivative, we need $(G'(x_i))^{-1}$. Now from (7.2) we see that

$$G'(x) = I - K(F'(x)), \tag{7.4}$$

which is an operator of the form that we discussed in Chapter V. Hence we see that we might use our previous work in getting an approximation to $(G'(x))^{-1}$ and bounding the error in the

approximation. Then using the Newton-Kantorovic Theorem and techniques discussed previously we should be able to prove the existence of a solution to (7. 1) and bound the error in the approximate solution. This method has not been explored in detail or implemented; however, it does illustrate a possible extension of the present work.

Thus we see that rigorous results may be obtained from the computer for many operator equations.

BIBLIOGRAPHY

1. Anselone, P. M. Convergence and error bounds for approximate solutions of integral and operator equations. In: Error in digital computation: Proceedings of an advanced seminar conducted by the Mathematics Research Center, United States Army, at the University of Wisconsin, Madison, October 5-7, 1964, ed. by Louis B. Rall. New York, Wiley, 1965. p. 231-252. (U.S. Army. Mathematics Research Center. Publication no. 15)

2. Ansleone, P. M. and R. H. Moore. Approximate solutions of integral and operator equations. Journal of Mathematical Analysis and Application 9: 268-277. 1964.

3. Berezin, I. S. and N. P. Zhidov. Computing methods, tr. by O. M. Blunn and ed. by A. D. Booth. Vol. 1. Reading, Massachusetts, Addison-Wesley, 1965. 464 p.

4. Control Data Corporation. 3300 computer maintenance training manual. Preliminary ed. Vol. 3. St. Paul, Minnesota, 1965. Various paging. (Publication no. 60158800)

5. _____. 3300 computer system reference manual. Preliminary ed. St. Paul, Minnesota, 1965. Various paging. (Publication no. 60157000)

6. Faddeeva, V. N. Computational methods of linear algebra, tr. by Curtis D. Benster. New York, Dover, 1959. 252 p.

7. Hansen, Eldon. Interval arithmetic in matrix computations. Journal of SIAM, ser. B, 2: 308-320. 1965.

8. Hildebrand, F. B. Introduction to numerical analysis. New York, McGraw-Hill, 1956. 511 p.

9. Kolmogorov, A. N. and S. V. Fomin. Elements of the theory of functions and functional analysis, tr. by Leo F. Boron. Vol. 1. Rochester, New York, Graylock Press, 1957. 129 p.

10. Kopal, Z. Numerical analysis. New York, Wiley, 1961. 594 p.

11. Krylov, V. I. Approximate calculation of integrals, tr. by Arthur H. Stroud. New York, MacMillan, 1962. 357 p.

12. Moore, Ramon E. The automatic analysis and control of error in digital computing based on the use of interval numbers. In: Error in digital computation: Proceedings of an advanced seminar conducted by the Mathematics Research Center, United States Army, at the University of Wisconsin, Madison, October 5-7, 1964. Vol. 1. New York, Wiley, 1965. p. 61-130. (U.S. Army. Mathematics Research Center. Publication no. 14)

13. Moore, R.H. Newton's method and variations. In: Nonlinear integral equations: Proceedings of an advanced seminar conducted by the Mathematics Research Center, United States Army, at the University of Wisconsin, Madison, April 22-24, 1963, ed. by P. M. Anselone. Madison, University of Wisconsin Press, 1964. p. 65-98. (U.S. Army. Mathematics Research Center. Publication no. 11)

14. Oregon State University Computer Center. Interval arithmetic. Corvallis, Oregon, 1967. 8 numb. leaves. (Report no. 67-1)

15. Tricomi, F. G. Integral equations. New York, Interscience, 1957. 238 p.

16. Wilkinson, J. H. Rounding errors in algebraic processes. Englewood Cliffs, New Jersey, Prentice Hall, 1963. 161 p.