NOTICE WARNING CONCERNING COPYRIGHT RESTRICTIONS:
The copyright law of the United States (title 17, U.S. Code) governs the making of photocopies or other reproductions of copyrighted material. Any copying of this document without permission of its author may be prohibited by law.
NUMERICAL STABILITY OF THE CHEBYSHEV METHOD
FOR THE SOLUTION OF LARGE LINEAR SYSTEMS

H. Woźniakowski
Department of Computer Science
Carnegie-Mellon University
(On leave from University of Warsaw)
March 1975

This work was supported in part by the Office of Naval Research under Contract
N0014-67-0314-0010, NR 044-422 and by the National Science Foundation under
Grant CJ32111.
NUMERICAL STABILITY OF THE CHEBYSHEV METHOD FOR THE SOLUTION OF LARGE LINEAR SYSTEMS

H. Woźniakowski
Department of Computer Science
Carnegie-Mellon University
(On leave from University of Warsaw)
March 1975
ABSTRACT

This paper contains the rounding error analysis for the Chebyshev method for the solution of large linear systems \( Ax + g = 0 \) where \( A = A^T \) is positive definite. We prove that the Chebyshev method in floating point arithmetic is numerically stable, which means that the computed sequence \( \{x_k\} \) approximates the solution \( x \) such that \( \lim_{k \to \infty} ||x_k - x|| \) is of order \( \zeta ||A|| ||A^{-1}|| ||x|| \) where \( \zeta \) is the relative computer precision.

We also point out that in general the Chebyshev method is not well-behaved, which means that \( x_k, k \) large, is not the exact solution for a slightly perturbed \( A \) or equivalently that the computed residuals \( r_k = Ax_k + g \) are of order \( \zeta ||A|| ||A^{-1}|| ||x|| \).

1. INTRODUCTION

Direct methods of numerical interest for the solution of linear systems \( Ax + g = 0 \) are numerically stable. This means that they produce an approximation \( y \) of the exact solution \( x \) such that \( ||y - x|| \) is of order \( \zeta ||A|| ||A^{-1}|| ||x|| \) where \( \zeta \) is the relative computer precision.

It might seem that the numerical accuracy of iterations for solving large linear systems can be better or even not depend on the condition number of \( A, k(A) = ||A|| ||A^{-1}|| \). In this paper we consider the Chebyshev method which is one of the most effective iterations for the solution of large linear systems. We show that this method is stable and that the condition number of \( A \) is crucial for this iteration.

Moreover direct methods are also well-behaved which means that the computed \( y \) is the exact solution for a slightly perturbed \( A \), i.e., \( (A + \Delta A)y + g = 0 \) where \( ||\Delta A|| \) is of order \( \zeta ||A|| \). Unfortunately this does not hold for the Chebyshev method. Thus, from the numerical accuracy point of view direct methods seem to be better than Chebyshev.

In Section 2 we briefly recall the main properties of the Chebyshev method \( T[a,b] \) for the solution of large linear systems \( Ax + g = 0 \) where \( A = A^T \) is positive definite, shortly denoted by \( A = A^+ > 0 \). In the classical case, the interval \( [a,b] \) contains all eigenvalues of \( A \). We consider the case where \( b > ||A|| \) and \( a \) is an arbitrary positive number. We also propose an extension of the Chebyshev method for singular matrices \( A = A^* \geq 0 \).

Section 3 deals with a perturbed Chebyshev method which generates a sequence \( \{x_k\} \) such that

\[
(1.1) \quad x_{k+1} = x_k + \left( p_{k-1}(x_k - x_{k-1}) - r_k \right) / q_k + \xi_k, \quad r_k = Ax_k + g,
\]

for suitable \( p_{k-1} \) and \( q_k \). We express the solution of (1.1) in terms of \( \xi_k \) and prove some asymptotic results.

In Section 4 we present an algorithm for the computation of \( p_{k-1} \) and \( q_k \). We prove that this
algorithm in floating point arithmetic computes $p_{k-1}$ and $q_k$ with high relative precision.

Section 5 deals with the proof of numerical stability of the Chebyshev method. We prove that $T[a,b]$ generates $\{x_k\}$ such that $\lim_{k \to \infty} |x_k - o|$ is of order $\frac{C \|A\| \|A^{-1}\| \|p\|}{\|b\|}$ whenever $b/a$ is of order $\|A\| \|A^{-1}\|$. In Section 6 we discuss well-behavior of the Chebyshev method. In general, the residual vectors in the Chebyshev method $r_k = Ax_k + g$ are of order $\frac{C \|A\| \|A^{-1}\| \|p\|}{\|b\|}$ which contradicts well-behavior. However, sometimes $r_k$ can be of order $\frac{C \|A\| \|p\|}{\|b\|}$. Such a case yields well-behavior.

2. CHEBYSHEV METHOD

Let us consider the numerical solution of a large linear system

$$A x + g = 0$$

where $A = A^* > 0$ is a given complex $n \times n$ matrix and $g$ is a given $n \times 1$ complex vector. Suppose $A$ is a sparse matrix of high order. Such systems commonly arise in the numerical solution of partial differential equations. Suppose we can only compute $y = A x$ for any vector $x$. Due to the sparseness of $A$ the vector $y$ can be computed in time and storage proportional to $n$ rather than $n^2$. For sufficiently large $n$, (2.1) can be solved only by iteration. Let $x_0$ be an arbitrary initial approximation of the solution $x = A^{-1} g$ and let

$$x_0 - \alpha = \sum_{j=1}^{m} c_j v_j$$

where $v_j$ are eigenvectors of $A$ associated with eigenvalues $\lambda_j$, $A v_j = \lambda_j v_j$, $(v_k, v_j) = \delta_{kj}$ and without loss of generality we can assume $c_j \neq 0$, for $1 \leq j \leq m$ and $\lambda_1 < \lambda_2 < \ldots < \lambda_m$, for $m \leq n$.

We consider a class of iterative methods which generate the sequences $\{x_k\}$ of the approximation of $\alpha$ such that

$$x_k - \alpha = W_k(A)(x_0 - \alpha)$$

where $W_k$ is a polynomial of degree $\leq k$. Since we only do know $Ax + g = 0$ than to eliminate $\alpha$ from (2.3) we have to assume

$$W_k(0) = 1.$$
where \( W_k \) and \( U_k \) are arbitrary polynomials of degree \( \leq k \). Assume that if \( x_0 = \alpha \) then \( x_k = \alpha \) for any \( \alpha \).

Then \( W_k(x) = U_k(x) + 1 \) and

\[
x_k - \alpha = (I + U_k(A)A)(x_0 - \alpha) = W_k(A)(x_0 - \alpha)
\]

which is equivalent to (2.3) and (2.4).

From (2.3) we get

\[
|x_k - \alpha|_2 \leq |W_k(A)|_2 \frac{|x_0 - \alpha|_2}{|x_0 - \alpha|_2} \leq |W_k|_2 \frac{|x_0 - \alpha|_2}{|x_0 - \alpha|_2}
\]

where

\[
(2.5) \quad |W_k|_2 = \max_{\lambda \in [a,b]} |W_k(\lambda)| \quad \text{and} \quad [\lambda_1, \lambda_m] \subset [a,b].
\]

Let \( P_k(0,1) \) denote a class of polynomials \( P \) of degree \( \leq k \) such that \( P(0) = 1 \).

In the Chebyshev method \( T[a,b] \), the \( W_k \) are defined as the polynomials of the smallest possible norms (2.5), i.e.,

\[
(2.6) \quad |W_k|_2 = \inf_{P \in P_k(0,1)} ||P||,
\]

and the solution of (2.6) is given by

\[
(2.7) \quad W_K(z) = T_k(f(z))/T_k(f(0))
\]

where \( f(z) = \frac{b+a}{b-a} - \frac{z}{b-a} \) and \( T_k \) denotes the Chebyshev polynomial of the first kind of degree \( k \). From (2.7) it follows that in the Chebyshev method \( T[a,b] \) we get

\[
(2.8) \quad |x_k - \alpha|_2 \leq 2 \left( \frac{b - \alpha}{5 + \sqrt{5}} \right)^k |x_0 - \alpha|_2
\]

for all \( k \) whenever \( [\lambda_1, \lambda_m] \subset [a,b] \), and

\[
(2.9) \quad x_{k+1} = x_k + \frac{b-a}{4} \frac{t_{k-1} + k \sqrt{k}}{t_k}, \quad k = 0, 1, \ldots,
\]

where \( t_k = A x_k + g \) and

\[
(2.10) \quad P_{-1} = 0, \quad p_{k-1} = \frac{b-a}{4} \frac{t_{k-1}}{t_k},
\]

\[
(2.11) \quad q_0 = \frac{b-a}{2}, \quad q_k = \frac{b-a}{4} \frac{t_{k-1} + k \sqrt{k}}{t_k}, \quad t_k = T_k(f(0)) \text{, } k \geq 1.
\]

(See, for instance, Stiefel (1958) and Rutishauser and Stiefel (1959).)

Usually, the eigenvalues \( \lambda_1 \) and \( \lambda_m \) from (2.2) are equal to the smallest eigenvalue \( \lambda_{\min} \), and to the
largest eigenvalue, $\lambda_{\text{max}}$, of $A$. Hence, the best convergence in $T[a,b]$ is for $a = \lambda_{\text{min}}$ and $b = \lambda_{\text{max}}$. However, in numerical practice $\lambda_{\text{min}}$ and $\lambda_{\text{max}}$ are known only for a few problems. In many cases we can easily find $b \geq \lambda_{\text{max}}$ (setting for instance $b = ||A||$ where $||\cdot||$ is any matrix norm, see Young (1971), page 32). A much harder problem is to find a suitable approximation of $\lambda_{\text{min}}$. Without knowledge of $\lambda_{\text{min}}$ one can use the Chebyshev method $T[a,b]$ for any values of $a > 0$ and $b \geq \lambda_{\text{max}}$. Then instead of (2.8) we get

$$\|x_k - \sigma\| \leq 2q(\lambda_{\text{min}})^k \|x_0 - \sigma\|_2$$

where

$$q(\lambda) = \frac{\sqrt{b-\lambda} + \sqrt{(a-\lambda)}}{\sqrt{b-\lambda} - \sqrt{(a-\lambda)}}$$

for $(a-\lambda) = a-\lambda$ if $a-\lambda \geq 0$ and zero otherwise.

Note that

(i) if $\lambda \in (0,a)$ then $q(\lambda) < 1$ which means the convergence of $T[a,b]$, however for $\lambda \rightarrow 0^+$, $q(\lambda) \rightarrow 1$,

(ii) if $a \leq \lambda \leq b$ then $q(\lambda) = \frac{\sqrt{b} - \sqrt{a}}{\sqrt{b} + \sqrt{a}}$,

(iii) if $\lambda < 0$ then $q(\lambda) > 1$. This implies that $T[a,b]$ is divergence whenever $\lambda_1$ from (2.2) is negative.

One can also consider the Chebyshev method for a singular matrix $A = A^* \neq 0$. In such case by $\sigma$ we mean the normal solution of $Ax+g = 0$, i.e., the vector of the minimal spectral norm which minimizes the spectral norm of the residual. Let $g = g_1 + g_2$ where $Ag_1 = 0$ and $g_1$ is orthogonal to $g_2$. Note that

$$A\sigma + g_2 = 0.$$ 

It is straightforward to verify that $\{x_k\}$ defined by (2.9) in $T[a,b]$ for $a > 0$ and $b \geq \lambda_{\text{max}}$ satisfies

$$x_k - \sigma = W_k(A)(x_0 - \sigma) + W'_k(0)g_1$$

for $W_k$ from (2.7) and

$$W'_k(0) = \frac{-k}{\sqrt{ab}} \cdot \frac{1-q(a)^{2k}}{1+q(a)^{2k}}.$$ 

Let us rewrite (2.2) as

$$x_0 - \sigma = c_1 v_1 + \sum_{j=1}^{n} c_j v_j$$

where $Av_j = \lambda_j v_j$, $\lambda_1 = 0$, $0 < \lambda_2 < \lambda_3 < \ldots < \lambda_n$, $(v_i,v_j) = 0_{i\neq j}$. Note that the normal solution $\sigma$ is orthogonal to $v_1$. Let us discuss the two cases.

Case 1. Let $g_1 = 0$. This means that $Ax+g = 0$ is solvable. From (2.14) it follows
Thus, if \( c_1 = 0 \) (which holds for instance if \( x_0 = 0 \)) the Chebyshev method is convergent and the best possible speed of convergence is for \( a = \lambda_2 \), i.e.,

\[
||x_k - \alpha||_2 \leq 2q(\lambda_2)^k||x_0 - \alpha||_2.
\]

Case II. Let \( g_1 \neq 0 \). In that case the iterative process is divergent, although \( \lim r_k = g_1 \). This suggests constructing \( y_k = x_k - W_k(0)r_k \). Then

\[
y_k - \alpha = W_k(A)(x_0 - \alpha) - W_k(0)W_k(A)(x_0 - \alpha)
\]

and for \( x_0 = 0 \) we get

\[
||y_k - \alpha|| \leq 2q(\lambda_2)^k||\alpha|| + \frac{2k}{\sqrt{b + \sqrt{b}}} \frac{1 - q(a)\lambda_2}{1 + q(a)\lambda_2} q(\lambda_2)^k||\alpha||,
\]

which once more implies the convergence of the Chebyshev method.

3. PERTURBED CHEBYSHEV METHOD

Recall we consider a large linear system

\[
Ax + g = 0
\]

where \( A = A^* > 0 \). We want to solve it by the Chebyshev method \( T[a,b] \) where it is only assumed that \( b = \lambda_{\max} \) and \( a > 0 \). The Chebyshev method generates a sequence \( \{x_k\} \) defined by (2.9), (2.10) and (2.11). However a sequence computed in floating point arithmetic can at best satisfy a perturbed relation (2.9), i.e.

\[
x_{k+1} = x_k + [p_{k-1}(x_k - x_{k-1}) - x_k]/q_k + \varepsilon_k,
\]

for suitable vectors \( \varepsilon_k \). A form of \( \varepsilon_k \) will be discussed in Section 5. In order to analyze the Chebyshev method in fl arithmetic we start to solve (3.1) for an arbitrary \( \{\varepsilon_k\} \) and find some asymptotical properties of the perturbed sequence \( \{x_k\} \).

Let \( e_k = x_k - \alpha \) be the error of the kth approximant. Then from (3.1) we get

\[
e_{k+1} = e_k + [p_{k-1}(e_k - e_{k-1}) - A\alpha]/q_k + \varepsilon_k.
\]

**Theorem 3.1**

Let \( \{\varepsilon_k\} \) be an arbitrary sequence and let \( \{x_k\} \) be a perturbed sequence generated by \( T[a,b] \) defined by (3.1), (2.10) and (2.11). Then
(3.3) \[ e_{k+1} = W_{k+1}(A)e_0 + \sum_{i=0}^{k} \kappa_{i,i} [(2-\beta_i)W_{k-1}(A) + (\beta_i+1)R_{k-1}(A)] \xi_i \]

where

(3.4) \[ \beta_k = 1 + \frac{P_{k-1}}{q_k} = \frac{q_0}{q_k} = 2 \frac{b+a}{b-a} \frac{t_k}{t_{k+1}}, \quad 1 \leq \beta_k \leq 2, \]

(3.5) \[ \kappa_{i,i} = \prod_{j=1}^{i-1} \frac{\beta_j+1}{\beta_j}, \quad (\kappa_{i,k,k-1} = 1), \quad \frac{1}{2} < \kappa_{i,i} \leq 1, \quad \lim_{k \to \infty} \kappa_{i,i} = \left(\frac{\sqrt{b} + \sqrt{a}}{2(b+a)}\right)^2, \]

(3.6) \[ W_k(z) = \frac{T_k(f(z))}{t_k}, \quad R_k(z) = \frac{U_k(f(z))}{t_k}, \quad f(z) = \frac{b+a}{z} - \frac{2}{b-a} \]

and \( T_k, U_k \) denote the Chebyshev polynomials of the first and second kind of degree \( k \), respectively.  

**Proof**

Induction on \( k \). Let \( k = 0 \). Since \( W_0 = R_0 = 1 \) and \( W_1(z) = 1 - \frac{1}{q_0} z, R_1 = 2W_1 \), then (3.3) is equal to

\[ e_1 = W_1(A)e_0 + \kappa_{0,0} [2-\beta_1 + \beta_1-1] \xi_0 = e_0 - \frac{1}{q_0} \tau_0 + \xi_0 \]

which is equivalent to (3.2).

Assume now that (3.3) holds for all \( i \leq k \). Let \( R_k = \beta_k R_{k-1} - \frac{1}{q_k} A, W_k = W_k(A) \) and \( R_k = R_k(A) \). Note that (3.4) easily follows from (2.10) and (2.11) and it is easy to verify that

(3.7) \[ W_{k+1} = B_k W_k + (1-\beta_k)W_{k-1}, \]

(3.8) \[ B_k = \beta_k W_k. \]

From (3.2), (3.3), (3.4), (3.7) and (3.8) we get

\[ e_{k+1} = [(1 + \frac{P_{k-1}}{q_k})I - \frac{1}{q_k} A] e_k - \frac{P_{k-1}}{q_k} e_{k-1} + e_k = B_k [W_k e_0 + \sum_{i=0}^{k-1} \kappa_{i,i} [(2-\beta_i)W_{k-1-i} + (\beta_i+1)R_{k-1-i}] \xi_i] + \]

\[ + (\beta_{i+1-1}R_{k-1-i} \xi_{i+1}) + (1-\beta_k) [W_{k-1} e_0 + \sum_{i=0}^{k-2} \kappa_{i,i} [(2-\beta_i)W_{k-2-i} + (\beta_i+1)R_{k-2-i}] \xi_i] + \]

\[ + \xi_k = U_k e_0 + e_k + B_k e_{k-1} + \sum_{i=0}^{k-2} [(2-\beta_i+1) [\kappa_{i+1,i} B_k W_{k-1-i} + \kappa_{i+2,i} (1-\beta_k) W_{k-2-i}] + \]

\[ + (\beta_{i+1-1} \kappa_{i+1,i} B_k R_{k-1-i} + \kappa_{i+2,i} (1-\beta_k) R_{k-2-i}] \xi_{i+1} \]

We want to verify that
(3.9) $b_k = \kappa_{k,k-1}[(2-\beta_k)\omega_1 + (\beta_k-1)r_1].$

(3.10) $2\beta_i \{\kappa_{k-1,i} \beta_k \omega_{k-1+i} + \kappa_{k-2,i} (1-\beta_k)\omega_{k-2-i} + (\beta_{i+1}-1)\{\kappa_{k-1,i} \beta_k r_{k-1,i} + \kappa_{k-2,i} (1-\beta_k) r_{k-2,i}\} \}

\quad = \kappa_{k,i} \{(2-\beta_{i+1})\omega_{k+1} + (\beta_{i+1}-1)r_{k-1,i}\}, \quad 0 \leq i \leq k-2.

Since $\kappa_{k,k-1} = 1$ and $r_1 = 2\omega_1$, (3.9) follows from (3.8). To prove (3.10) we use (3.7) which holds for $\omega_{k-1}$ and $r_{k-1}$. By comparing the coefficients at $AW_{k-1-i}$, $W_{k-1-i}$, $W_{k-2-i}$, $AR_{k-1-i}$, $R_{k-1-i}$ and $R_{k-2-i}$ we get three equations on $\kappa_{k,i}$:

(3.11) $\kappa_{k,i}/\omega_{k-i} = \kappa_{k-1,i}/\omega_{k-i}.$

(3.12) $\kappa_{k,i}/\beta_{k-i} = \kappa_{k-1,i}/\beta_{k-i}.$

(3.13) $\kappa_{k,i}(1-\beta_{k-i}) = \kappa_{k-2,i}(1-\beta_k).$

From (3.5) and (3.4) it follows

$$\frac{\kappa_{k,i}}{\kappa_{k-1,i}} = \frac{\beta_k}{\beta_{k-1-i}} = \frac{\omega_{k-i}}{\omega_{k-1-i}} = \frac{\beta_{k-i}}{\beta_{k-2-i}} = \frac{\beta_{k-i}}{q_k} = \frac{\omega_{k-i}}{\omega_{k-2-i}} = \frac{\beta_{k-i}}{\beta_{k-i-1}},$$

which gives (3.11) and (3.12). Next, observe that from (3.4) and (2.10) we get

$$\frac{\kappa_{k,i}}{\kappa_{k-1,i}} = \frac{\beta_k}{\beta_{k-1-i}} = \frac{\omega_{k-i}}{\omega_{k-1-i}} = \frac{\beta_{k-i}}{\beta_{k-2-i}} = \frac{\beta_{k-i}}{q_k} = \frac{\omega_{k-i}}{\omega_{k-2-i}} = \frac{\beta_{k-i}}{\beta_{k-i-1}},$$

which proves (3.13) and completes the inductive proof of (3.3). To prove the limit of $\kappa_{k,i}$ note that

$$\kappa_{k,i} = \frac{\beta_k}{\beta_{k-1-i}} = \frac{\omega_{k-i}}{\omega_{k-1-i}} = \frac{\beta_{k-i}}{\beta_{k-2-i}} = \frac{\beta_{k-i}}{q_k} = \frac{\omega_{k-i}}{\omega_{k-2-i}} = \frac{\beta_{k-i}}{\beta_{k-i-1}},$$

where

$$q(a) = \frac{\sqrt{b} - \sqrt{a}}{\sqrt{b} + \sqrt{a}}.$$

This completes the proof of Theorem 3.1. ■

Using Theorem 3.1 we prove a bound on the perturbed errors. Recall that $\lambda_{\text{min}}$ is the smallest eigenvalue of $A$ and $q(\lambda)$ is defined by (2.13).

**Corollary 3.1**

Let

$$\delta = \frac{\sqrt{b} - \sqrt{a}}{\sqrt{b} - \sqrt{a}} + q, \quad \eta_k = 1 - \frac{2(2k+1)}{1-\delta}, \quad \text{for } k = 0, 1, \ldots, \text{and } q = q(\lambda_{\text{min}}).$$

Then
\[(3.15) \quad |e_{k+1}| \leq 2^{k+1} |e_0| + 2 \sum_{i=0}^{k} \eta_{k-i} |\xi_i|,\]

\[(3.16) \quad e = \limsup_k |e_k| \leq \frac{a \delta (\sqrt{a} + \sqrt{\beta})}{\min(a, \lambda_{\text{min}}^2)} \xi + \frac{b}{\min(a, \lambda_{\text{min}}^2)} \xi.\]

where \(\xi = \limsup_k |\xi_k|\).

**Proof**

First of all observe that

\[1 \leq \eta_k \leq k+1 \text{ and } \eta_k = k+1 \text{ whenever } \lambda_{\text{min}} \geq a,\]

\[|e_k| \leq 2^k \eta_k \text{ and } |e_k| \leq 2^k \eta_k.\]

From (3.3), (3.4) and (3.5) it follows

\[|e_{k+1}| \leq 2^{k+1} |e_0| + \sum_{i=0}^{k} \max(|e_{k-i}||\xi_i| + |\xi_i| |\xi_i|) \leq 2^{k+1} |e_0| + 2 \sum_{i=0}^{k} \eta_{k-i} |\xi_i|\]

which proves (3.15).

Let \(\epsilon\) be any positive number. There exists \(k_0\) such that \(|\xi_k| \leq \xi + \epsilon\) for all \(k > k_0\). From (3.15) we get

\[e \leq 2 \limsup_k \left( \sum_{i=0}^{k_0} \eta_{k-i} |\xi_i| + \sum_{i=k_0+1}^{k} \eta_{k-i} (\xi + \epsilon) \right).\]

Note that \(\eta_{k-i} \to 0\) for \(i = 0, 1, \ldots, k_0\) and

\[\limsup_k \sum_{i=k_0+1}^{k} \eta_{k-i} = \lim_{i \to \infty} \sum_{i=0}^{k_0} \eta_{k-i} = \frac{\frac{1}{1-i-q} - \frac{1}{1-q(a) \delta}}{1-q(a) \delta} = \frac{(\sqrt{a} + \sqrt{\beta})^2 (2-q(a) \delta)}{4 \min(a, \lambda_{\text{min}}^2)} \leq \frac{2b}{\min(a, \lambda_{\text{min}}^2)}\]

which proves (3.16).

**Corollary (3.2)**

(i) If \(\lim_k \xi_k = \xi\) then \(\lim_k e_k = \left(\frac{\sqrt{a} + \sqrt{\beta}}{2}\right) A^{-1} \xi,\)

(ii) If \(\limsup_k |\xi_k - \xi_{k-1}| = \xi\) then \(\limsup_k \left|A_{k+1} - q_k \xi_k\right| \leq x (\epsilon + a)\)

**Proof**

Note that \(\lim_k \eta_k = q^* = \left(\frac{\sqrt{a} + \sqrt{\beta}}{2}\right)^2.\) Let \(z_k = e_k - q^* A^{-1} \xi.\) From (3.2) it follows

\[z_{k+1} = z_k + (p_{k-1}(z_{k-1}^2 - z_k^2) - Az_k) + q_k^* \xi_k - z_k + (1-q_k^*) \xi.\]

Applying Corollary 3.1 and Theorem 3.1 to \(z_k\) we get...
\[ \limsup_{k} \|x_k\| \leq 4 \frac{b}{\min(a, \lambda_{\text{min}})} \limsup_{k} \|x_k - x^{*}(1 - q_k)\xi\| = 0 \]

which proves conclusion (i).

To prove conclusion (ii) we rewrite (3.2) as follows:

\[ \mathbf{A}e_k - q_k \xi = p_{k-1}(e_{k-1} - e_k) - q_k(e_{k+1} - e_k). \]

Since \( x_k - x_{k-1} = e_{k-1} - e_k \) and \( \lim_{k} p_k = p^* = \left( \frac{\sqrt{5} - \sqrt{3}}{2} \right)^2 \) then

\[ \limsup_{k} \|\mathbf{A}e_k - q_k \xi\| \leq (q^* + p^*)k = \kappa(b + a) \]

which completes the proof of Corollary 3.2.

4. ALGORITHM OF \( p_{k-1} \) AND \( q_k \)

In this section we deal with the computation of \( p_{k-1} \) and \( q_k \) which appear in the Chebyshev method \( T[a, b] \) in (2.9). Recall that

(4.1) \( p_1 = 0, p_{k-1} = \frac{b-a}{4} t_{k-1}^{-1}, \lim_{k} p_k = p^* = \left( \frac{\sqrt{5} - \sqrt{3}}{2} \right)^2, \)

(4.2) \( q_0 = \frac{a+b}{2}, q_k = \frac{b-a}{4} t_k^{-1}, \lim_{k} q_k = q^* = \left( \frac{\sqrt{5} + \sqrt{3}}{2} \right)^2, \)

where \( t_k = T_k(\frac{b+a}{b-a}), k \geq 1. \)

Let

(4.3) \( c = \frac{a+b}{2}, d = \left( \frac{b-a}{4} \right)^2 \) and \( \gamma_k = q^* - q_k \) for \( k \geq 0. \)

From the recurrence formula of the Chebyshev polynomials it follows

(4.4) \( q_k = \frac{b-a}{4} \left( \frac{b-a}{b-a} t_k - t_{k-1} \right)/t_k = c - d/q_{k-1}, \quad k \geq 2. \)

From (4.2), (4.3) and (4.4) we get

\[ \gamma_k = q^* - c + d/q_{k-1} = d/q_{k-1} - d/q^* = d \gamma_{k-1}/(q_{k-1}q^*), \quad k \geq 2. \]

Note that (4.1) and (4.2) gives \( p_{k-1}q_{k-1} = d, \quad k \geq 2. \)

This suggests the following algorithm for the computation of \( p_{k-1} \) and \( q_k. \)

Algorithm 4.1

(4.5) \( c = \frac{a+b}{2}, d = \left( \frac{b-a}{4} \right)^2, q = \frac{a+b+2\sqrt{ab}}{4}, \)

(4.6) \( p_1 = 0, q_0 = c. \)
Let us consider the above algorithm in t digit floating point arithmetic, fl, and let \( \text{rd}(x) \) denote the numerical representation of any real number \( x \) and \( \text{fl}(xOy) \) denote the computed result of an arithmetic operation \( \circ \in \{+,-,\times,\div\} \). Then

\[
\text{rd}(x) = x(1+\varepsilon), \quad |\varepsilon| = |\varepsilon(x)| \leq \zeta,
\]

for \( x = \text{rd}(x) \) and \( y = \text{rd}(y) \),

\[
\text{fl}(xOy) = (xOy)(1+\varepsilon), \quad |\varepsilon| = |\varepsilon(x,y,\circ)| \leq \zeta
\]

where \( \zeta = 2^{-t} \).

We also assume that for \( x = \text{rd}(x) \), \( \text{fl}(\sqrt{x}) = \sqrt{x}(1-\varepsilon) \), \( |\varepsilon| \leq \zeta \). (See Wilkinson (1963).) To simplify the further estimations of roundoff errors we shall use the relation \( \prec \), i.e., if \( a(t) \) and \( b(t) \) are bounded functions of \( t \), \( t \geq t_0 > 0 \) then \( a(t) \prec b(t) \) iff there exists \( K \) independent of \( t \) such that

\[
a(t) = b(t)(1+\varepsilon(t)2^{-t}) \quad \text{where} \quad |\varepsilon(t)| \leq K \quad \text{for} \quad t \geq t_0.
\]

Next \( a(t) \prec b(t) \) iff \( a(t) \leq b(t) \) or \( a(t) \sim b(t) \). (For more details see e.g. Wozniakowski (1974).)

Let us denote any computed value \( x \) in Algorithm 4.1 by \( \overline{x} \) and let \( \overline{x} = x(1+\eta_k) \). Thus \( \eta_k \) is the relative error of \( x \).

**Theorem 4.1**

Let \( a = \text{rd}(a) \) and \( b = \text{rd}(b) \). The computed values \( \overline{p}_k \) and \( \overline{q}_k \) are equal to

\[
\begin{align*}
(4.11) \quad \overline{p}_k &= p_k(1+\eta_{p_k}), \quad |\eta_{p_k}| \leq (4+L_k)\zeta, \\
(4.12) \quad \overline{q}_k &= q_k(1+\eta_{q_k}), \quad |\eta_{q_k}| \leq L_k \zeta,
\end{align*}
\]

where \( 0 \leq L_k \leq 15.5 + 64k \) for \( k = \sqrt{a/b}/(1+\sqrt{a/b})^2 \) and \( \lim_{k \to \infty} L_k = 3.5 \).

Theorem 4.1 means that we compute \( \overline{p}_k \) and \( \overline{q}_k \) with high relative precision for all values of \( a \) and \( b \). There are some other algorithms for computing \( p_k \) and \( q_k \) but usually for these algorithms one can prove (4.11) and (4.12) with \( L_k \) which is proportional to \( b/a \). (For instance, an algorithm based on (4.4) and (4.8).)
Proof

We verify (4.11) and (4.12) for \( k = 0 \) and \( k = 1 \). From (4.5), (4.6) and (4.7) we get

\[
\bar{\eta}_0 = \bar{\epsilon} = \frac{a+b}{2} (1+\epsilon_1) = c(1+\eta_0), \quad |\eta_0| \leq \zeta,
\]

\[
\bar{\eta}_1 = \frac{2(1+\eta_0)}{a+b}(1+\epsilon_2) = \frac{2(1+\eta_0)}{a+b} (1+\eta_0), \quad |\eta_0| \leq \frac{3}{2} \zeta,
\]

\[
\bar{\eta}_q = \frac{(a+b)(1+\epsilon_1) + 2 \sqrt{ab} (1+\epsilon_2) (1+\epsilon_3)}{4} (1+\epsilon_6) = q \left( 1 + \frac{e_1 (a+b) + \sqrt{1+e_1} (1+\epsilon_2) - 1) 2 \sqrt{ab}}{a+b + 2 \sqrt{ab}} \right) (1+\epsilon_6) =
\]

\[
= q^{(1+\eta_0)}, \quad |\eta_q| \leq \frac{5}{2} \zeta,
\]

\[
\bar{\eta}_q = \frac{2 \sqrt{ab} (1+\epsilon_2) (1+\epsilon_3)}{a+b} (1+\epsilon_6) (1+\epsilon_6) = q \left( 1 + \frac{e_1 (a+b) + \sqrt{1+e_1} (1+\epsilon_2) - 1) 2 \sqrt{ab}}{a+b + 2 \sqrt{ab}} \right) (1+\epsilon_6) =
\]

\[
\frac{2 \sqrt{ab} (1+\epsilon_2) (1+\epsilon_3)}{a+b} = \gamma_1 (1+\eta_1), \quad |\eta_1| \leq 11 \zeta
\]

where \(|\epsilon_i| = \zeta\) for \( i = 1, 2, \ldots, 12 \).

Hence (4.11) and (4.12) hold for \( k = 0 \) and \( k = 1 \). Let us analyze (4.8), (4.9) and (4.10). We get

\[
\bar{P}_{k-1} = \frac{d(1+\eta_0)}{q_k-1} (1+\epsilon_1) = p_k-1 (1+\eta_0), \quad |\eta_0| < 4 \zeta + |\eta_{k-1}|,
\]

\[
\bar{q}_k = \frac{p_k-1 (1+\eta_0)}{q_k} (1+\epsilon_2) = \gamma_k (1+\eta_0), \quad |\eta_0| < 4 \zeta + |\eta_{k-1}| + |\eta_{q_k-1}|,
\]

\[
q_k = (q(1+\eta_0) - \gamma_k (1+\eta_0)) (1+\epsilon_3) = q_k \left( 1 + \frac{\eta_0 + \gamma_k (1+\eta_0)}{q_k} \right) (1+\epsilon_3) = q_k (1+\eta_0),
\]

\[
|\eta_{q_k}| \leq (1+2.52 \zeta) \zeta + \frac{\gamma_k}{q_k} |\eta_{\bar{q}_k}|.
\]

Substituting (4.15) to (4.14) we get

\[
|\eta_{\bar{q}_k}| \leq \left( 12 + 2.52 \frac{\eta_{k-1}}{q_{k-1}} \right) \zeta \leq \left( 1 + \frac{\gamma_k}{q_k} \right) |\eta_{\bar{q}_k}|.
\]

Note that

\[
\frac{\gamma_k}{q_k} = \frac{d}{q_k} \cdot \frac{1}{q_k}, \quad \frac{\gamma_k-1}{q_k} = \left( \prod_{j=2}^{k} \frac{d}{q_j} \right) \frac{\gamma_1}{q_1} = 4 \kappa \frac{2k}{1+\kappa} \leq 4 \kappa
\]

where \( \kappa = \sqrt{a/b} \), and \( q = (\sqrt{a} - \sqrt{b})/(\sqrt{a} + \sqrt{b}) \).
Thus, (4.16) becomes

\[ |\eta_k| \leq (12 + 10\kappa)\zeta + (1 + 4\kappa)|\eta_{k-1}| \]

Since \(|\eta_k| \leq 11\zeta\), the solution of (4.17) is given by

\[ |\eta_k| \leq (1 + 4\kappa)^{k-1}11\zeta + (12 + 10\kappa)^{1+4\kappa\zeta} \]

Coming back to \(\eta_\kappa\), we have

\[ |\eta_\kappa| \leq 3.5\zeta + 10\kappa\zeta^2 + 44\kappa\zeta^2 + (12 + 10\kappa)^2 \]

Note that

\[ q^2(1 + 4\kappa) = \frac{(\sqrt{b} - \sqrt{a})^2(b + a + 4\sqrt{ab})}{(\sqrt{b} + \sqrt{a})^4} < 1. \]

Thus,

\[ L_\kappa \leq 3.5 + 10\kappa\zeta^2 + (12 + 10\kappa)^2 \leq 15.5 + 64\kappa, \]

and

\[ \lim_{\kappa} L_\kappa = 3.5. \]

Finally, from (4.13) it follows

\[ |\eta_k| \leq 4\zeta + L_\kappa \zeta, \]

which completes the proof of Theorem 4.1.

5. NUMERICAL STABILITY OF THE CHEBYSHEV METHOD

In this section we deal with numerical stability for the Chebyshev method. Let us briefly recall that an iterative method for the solution of the linear equation \(Ax + g = 0\) is numerically stable if it produces a sequence \(\{x_k\}\) such that

\[ \lim \sup_{k} \|x_k - x\| \leq CK\|a\| \|a\|^{-1} \|g\| + o(\zeta^2) \]

where \(K\) can only depend on the size \(n\) (see Wozniakowski (1975)).

We propose the following algorithm of the Chebyshev method (see (2.3) and Rutishauser, Stiefel and others (1959)).

**Algorithm 5.1**

The Chebyshev method \(T[a, b], 0 < a\) and \(\|a\| \leq b\).

- \(x_0\) is a given initial approximation,
- for \(k = 0, 1, \ldots\),
- compute \(q_k\) and \(p_{k-1}\) by Algorithm 4.1,

\[ r_k := Ax_k + g. \]
Proof

We verify (4.11) and (4.12) for \( k = 0 \) and \( k = 1 \). From (4.5), (4.6) and (4.7) we get

\[
q_0 = \tau = \frac{a+b}{2}(1+\epsilon_k) = c(1+\eta^e), \quad |\eta^e| \leq \zeta,
\]

\[
\tau = \left(\frac{(b-a)(1+\epsilon_k)}{4}\right)^2(1+\epsilon_3) = d(1+\eta_d), \quad |\eta_d| \leq 3\zeta,
\]

\[
q^e = \frac{(a+b)(1+\epsilon_k) + 2/ab(1+\epsilon_k)}{4}q_0(1+\epsilon_6) = \eta^e \left(1 + \frac{\epsilon_k(a+b) + [\sqrt{1+\epsilon_k(1+\epsilon_3)} - 1]2/ab}{a+b + 2/ab}\right)(1+\epsilon_6) = q^e(1+\eta^e^e), \quad |\eta^e^e| \leq \frac{3}{2}\zeta,
\]

\[
\bar{p}_0 = \frac{2d}{c}(1+\epsilon_7) = \frac{2d(1+\eta_d)}{c(1+\eta_d)}(1+\epsilon_2) = \bar{P}_0(1+\eta_0), \quad |\eta_0| \leq 5\zeta,
\]

\[
\bar{q}_1 = \left(\frac{(a+b)(1+\epsilon_k)}{4} + \frac{ab(1+\epsilon_k)}{(a+b)(1+\epsilon_1)}(1+\epsilon_8)\right)(1+\epsilon_9) = q_1(1+\eta_1), \quad |\eta_1| \leq 4\zeta,
\]

\[
\bar{q}_1 = \frac{2/ab(1+\epsilon_k)(1+\epsilon_8)d(1+\eta_0)(1+\epsilon_4)}{(a+b)(1+\epsilon_1)q_1(1+\eta^e_1)(1+\epsilon_1)} = \eta_1(1+\eta^e_1), \quad |\eta^e_1| \leq 11\zeta
\]

where \( |\epsilon_i| \leq \zeta \) for \( i = 1, 2, \ldots, 12 \).

Hence (4.11) and (4.12) hold for \( k = 0 \) and \( k = 1 \). Let us analyze (4.8), (4.9) and (4.10). We get

\[
(4.13) \quad \bar{p}_k = \frac{d(1+\eta_d)}{q_{k-1}(1+\eta^e_k)}(1+\epsilon_k), \quad |\eta_{k-1}| \leq 4\zeta + |\eta_{q^e_{k-1}}|,
\]

\[
(4.14) \quad \bar{q}_k = \frac{p_{k-1}(1+\eta^e_{k-1})q_k(1+\eta^e_{k-1})}{q_{k}(1+\eta^e_k)}(1+\epsilon_k) = \gamma_k(1+\eta^e_k), \quad |\eta^e_k| \leq 8.5\zeta + |\eta_{k-1}| + |\eta_{q^e_{k-1}}|,
\]

\[
(4.15) \quad \bar{q}_k = \left(q(1+\eta^e_k) - \gamma_k(1+\eta^e_k)\right)(1+\epsilon_k) = q_k \left(1 + \frac{\gamma_k}{q_k}\right)(1+\epsilon_k) = q_k(1+\eta^e_k),
\]

\[
|\eta^e_k| \leq \left(1+2.5\frac{2k}{\xi_k}\right)\zeta + \frac{q_k}{q_{k}}|\eta^e_k|.
\]

Substituting (4.15) in (4.14) we get

\[
|\eta_k| \leq \left(12 + 2.5\frac{\gamma_k}{q_{k}}\right)\zeta \left(1 + \frac{\gamma_k}{q_k}\right)|\eta_{k-1}|.
\]

Note that

\[
\gamma_k = \frac{d}{q_k^2} \cdot \frac{\gamma_{k-1}}{q_{k-1}} = \left(\prod_{i=2}^{k} \frac{d}{q_i^2}\right) \frac{\gamma_1}{q_1} = 4\kappa \cdot \frac{2k}{1+q^2_{k+2}} \leq 4\kappa
\]

where

\[
\kappa = \sqrt{\frac{1}{2}} \left(1+\sqrt{5}\right)^2 \quad \text{and} \quad q = (\sqrt{5} - \sqrt{3})/\sqrt{5} + \sqrt{3}.
\]
Thus, (4.16) becomes

\[(4.17) \ | \eta_{y_k} | \leq (12 + 10\kappa)\zeta + (1 + 4\kappa)|\eta_{y_{k-1}}| \]

Since \(| \eta_{y_1} | \leq 11\zeta\), the solution of (4.17) is given by

\[| \eta_{y_k} | \leq (1 + 4\kappa)^{k-1}11\zeta + (12 + 10\kappa)\frac{(1+4\kappa)^{k-1} - 1}{4\kappa} \cdot \zeta. \]

Coming back to \(\eta_{q_k}\), we have

\[(4.18) \ | \eta_{q_k} | \leq 3.5\zeta + 10\kappa q^2 + 10\kappa q^2(1 + 4\kappa)^{k-1}\zeta + (12 + 10\kappa)q^2[(1+4\kappa)^{k-1} - 1] = L_k\zeta. \]

Note that

\[q^2(1 + 4\kappa) = \frac{(\sqrt{5} - \sqrt{1})^2(b+c+4\sqrt{ab})}{(\sqrt{5} + \sqrt{1})^2} < 1. \]

Thus,

\[L_k \leq 3.5 + 10\kappa + 44\kappa q^2 + (12 + 10\kappa)q^2 \leq 15.5 + 64\kappa, \]

and

\[\lim_{k \to \infty} L_k = 3.5. \]

Finally, from (4.13) it follows

\[| \eta_{p_k} | \leq 4\zeta + L_k\zeta, \]

which completes the proof of Theorem 4.1.

5. NUMERICAL STABILITY OF THE CHEBYSHEV METHOD

In this section we deal with numerical stability for the Chebyshev method. Let us briefly recall that an iterative method for the solution of the linear equation \(Ax + g = 0\) is numerically stable if it produces a sequence \(\{x_k\}\) such that

\[(5.1) \limsup_{k \to \infty} ||x_k - x|| \leq CK||x|| ||A^{-1}|| ||g|| + O(\zeta^2), \]

where \(K\) can only depend on the size \(n\) (see Wozniakowski (1975)).

We propose the following algorithm of the Chebyshev method (see (2.3) and Rutishauser, Stiefel and others (1959)).

Algorithm 5.1

The Chebyshev method \(T[a,b], 0 < a \text{ and } ||A|| \leq b. \)

\(x_0\) is a given initial approximation,

for \(k = 0,1,\ldots\)

compute \(q_k\) and \(p_{k-1}\) by Algorithm 4.1,

\[(5.2) \ r_k := A x_k + g; \]

12
Theorem 5.1

Let \( \{x_k\} \) be the sequence computed in \( \mathbb{F}_\ell \) arithmetic by Algorithm 5.1. If

\[
\ell_1(Ax_k + g) = (I + \delta I_k)((A + E_k)x_k + g)
\]

where \( \|E_k\| \leq K_1 \|A\| \) and \( \|\delta I_k\| \leq K_2 \|e\| \), \( K_1 = K_1(n) \) for \( i = 1, 2 \),

then for small \( \zeta \),

\[
\limsup_{k} \|x_k - \alpha\| \leq \frac{4(1 + 4K_1)}{\min(a, \lambda \min)} \|\alpha\| \left(1 - \frac{4b(59 + 4K_1 + 4K_2)}{\min(a, \lambda \min)} \zeta \right). \tag{5.5}
\]

Proof

From Theorem 4.1 and (5.5) the computed \( x_{k+1} \) is equal to

\[
x_{k+1} = (I + D_k^1) [x_k + (I + 2L_k^1)(p_{k-1}(1 + \eta p_{k-1}^1)(I + 2D_k^1)(x_k - x_{k-1}) - (I + \delta I_k)(Ax_k + g + E_k x_k))]/q_k (1 + \eta) \tag{5.6}
\]

where \( D_k^1 \) denotes a diagonal matrix and \( \|x_k\| \leq \zeta \), \( i = 1, 2, 3 \). After some transformations, (5.6) becomes

\[
x_{k+1} = x_k + [p_{k-1}(x_k - x_{k-1}) - (Ax_k + g)]/q_k + e_k \tag{5.7}
\]

where

\[
e_k = D_k^1 x_k - q_k x_k + \Theta_k \tag{5.8}
\]

and

\[
\|\Theta_k\| \leq C(1 + I_k + L_{k-1} + (3 + I_k + K_1 + K_2)\|A\|/q_k)\|e_k\| + C(9 + I_k + L_{k-1})\|e_k\| + 2K_1(3 + K_2 + I_k)\|A\|/q_k. \]

Here, as always, \( e_k = x_k - \alpha \) and \( I_k \) is defined in Theorem 4.1. Since \( \lim q_k = q^* = b/4 \geq \|A\|/4 \) and \( \lim L_k = 3.5 \) we get

\[
\limsup_k \|e_k\| \leq C(1/4) + \zeta \limsup_k \|e_k\| (59 + 4K_1 + 4K_2). \]

Finally, applying Corollary 3.1 we get, \( e = \limsup_k \|e_k\| \),

\[
e \leq \frac{4 \times (1/4) C}{\min(a, \lambda \min)} \|\alpha\| + \frac{4b(59 + 4K_1 + 4K_2)}{\min(a, \lambda \min)} C \zeta. \tag{5.9}
\]

Hence, (5.5) follows from the last relation which completes the proof. \( \blacksquare \)
From Theorem 5.1 we can easily get (5.1). Since \( A = A^* > 0 \) then \( \| A \| \| A^{-1} \| = \lambda_{\text{max}}/\lambda_{\text{min}} \). It leads us to the following.

**Corollary 5.1**

If there exists a constant \( \mathcal{L} = \mathcal{L}(n) \) such that for every matrix \( A = A^* > 0 \) we use the Chebyshev method \( T[a,b] \) where

\[
(5.9) \quad \frac{b}{\min(a,\lambda)} \leq \frac{\lambda_{\text{max}}}{\lambda_{\text{min}}}
\]

then the Chebyshev method is numerically stable. Specifically \( T[a,b] \) produces a sequence \( \{ x_k \} \) such that

\[
(5.10) \quad \limsup_k \| x_k - x \| \leq C(1 + K_1) \| A \| \| A^{-1} \| \| x \| + O(\varepsilon^2)
\]

where \( K_1 \) is defined by (5.4).

**Proof**

From Theorem 5.1 and from the definition of the relation \( \leq \) it follows

\[
e \leq 4(1 + 4K_1) \| A \| \| A^{-1} \| \| x \| \varepsilon (1 + O(\varepsilon))
\]

which gives (5.10).

If \( \varepsilon \) is small then one can prove that the constant which appears in the "0" notation in (5.10) only depends on \( (\| A \| \| A^{-1} \|)^2, K_1 \) and \( K_2 \) (see (5.4)). Note that if \( b < 2 \lambda_{\text{max}} \) then for any \( \varepsilon > \lambda_{\text{min}} \) (5.9) holds; however, for increasing \( \varepsilon \) the convergence of \( T[a,b] \) is getting worse, see (2.13) and (3.15).

We want to show that without additional assumption on \( D_k, F_k, E_k \) and \( \Theta_k \) in (5.8), the estimate (5.10) is sharp which means that the condition number of \( A \) is crucial for the accuracy of the solution of linear equation solving by iteration.

Let us assume for simplicity that \( a = \lambda_{\text{min}} \) and \( b = \lambda_{\text{max}} \). From (5.4) and (5.6) we know that \( D_k \) and \( E_k \) are small but arbitrary. Assume theoretically that \( D_k = 0, \Theta_k = 0 \) and \( \lim_k E_k = E \) where

\[
\| A^{-1} E_k \| = \| A^{-1} \| \| E_k \| \| x \| \quad \text{and} \quad \| E_k \| = K_1 \| A \|.
\]

From Corollary (3.2) we get

\[
\lim_k e_k = -A^{-1} E x \quad \text{and} \quad e = C K_1 \| A \| \| A^{-1} \| \| x \|
\]

which is essentially the righthand side of (5.10). Furthermore, if \( E_k = 0, \Theta_k = 0 \) and \( \lim_k D_k = D \) where

\[
\| A^{-1} D \| = \| A^{-1} \| \| D \| \| x \|, \quad \| D \| = \varepsilon, \quad \text{we have}
\]

\[
\lim_k e_k = \frac{\sqrt{\varepsilon} + \sqrt{\varepsilon}}{2} A^{-1} D x, \quad e = C(1 + \sqrt{\varepsilon}/2)^2 \| A \| \| A^{-1} \| \| x \|
\]

This implies that even using the double precision for the evaluation of \( Ax + g \), \( \| k_k \| \leq K_1 \varepsilon^2 \| A \| \), we cannot guarantee the high relative precision of the computed \( \{ x_k \} \).
Thus, (5.10) is sharp. Although, from (5.8) we get

\[ \limsup_{k} \| x_k - x \| \leq 4 \| x \| \| A^{-1} \| \limsup_{k} \| b_k - 1 \frac{E_k}{q_k} \| (1 + o(1)) \]

and if \( \| b_k - 1 \frac{E_k}{q_k} \| \ll \| x \| \) we can expect a better result.

6. WELL-BEHAVIOR OF THE CHEBYSHEV METHOD

Let us briefly recall that a method for the solution of linear systems \( Ax + g = 0 \) is said to be well-behaved if a slightly perturbed computed approximation \( y \) is the exact solution of a slightly perturbed problem, i.e.,

\[ (A + \Delta A)(y + \Delta y) + g + \Delta g = 0 \]

where \( \| \Delta A \| \leq c_1 \| A \| \), \( \| \Delta y \| \leq \zeta c_2 \| y \| \) and \( \| \Delta g \| \leq \zeta c_3 \| g \| \), \( c_4 = c_1(n) \).

Let \( \Delta y \) and \( \Delta g \) be matrices defined by

\[ (I + \Delta y)y = y + \Delta y; \ (I + \Delta g)g = g + \Delta g \]

and

\[ \| \Delta y \| \leq \zeta c_2 \| y \|, \ | \Delta g \| \leq \zeta c_3 \| g \| . \]

Hence, (6.1) becomes

\[ (A + \Delta A)y + g = 0 \]

where \( \| \Delta A \| \leq \zeta c_4 \| A \| \) for \( c_4 = c_1 + c_2 + c_3 \).

Thus, without loss of generality, a method is well-behaved if the computed \( y \) is the exact solution of the problem with a slightly perturbed matrix \( A \).

Let \( r = f(\Delta y + g) \) be the computed residual vector. Assume

\[ r = (I + \Delta I)((A + E)y + g) \]

where \( \| \Delta I \| \leq \zeta c_5 \) and \( \| E \| \leq \zeta c_6 \| A \| \).

It is easy to verify that a method is well-behaved iff \( r \) satisfies

\[ \| r \| \leq \zeta c_7 \| A \| \| y \| \]. \]

Indeed, if (6.2) holds then \( \| r \| \leq \zeta (c_4 + c_6) \| A \| \| y \| \). If (6.4) holds then

\[ \left( A + E - (I + \Delta I)^{-1} \frac{\Sigma y}{\| y \|} \right)y + g = 0. \]

Thus, \( \Delta A = E - (I + \Delta I)^{-1} \frac{\Sigma y}{\| y \|} \) and \( \| \Delta A \| \leq \zeta (c_6 + c_7) \| A \| \).

We wish to consider the well-behavior problem for the Chebyshev method \( T[a,b] \). This means we must
verify if the computed vectors \( r_k^* = f_l(Ax^* + g) \) satisfies condition (6.4) for large \( k \). From (5.4) we get

\[
\|r_k^* - r_{k-1}^*\| \leq K_1 \zeta \|A\| \|b\|
\]

where \( r_k^* = A^2 x_k \).

Thus the Chebyshev method is well-behaved iff \( r_k^* \) satisfies (6.4). Let us assume for simplicity that \( a = \lambda_{\text{min}} \) and \( b = \lambda_{\text{max}} \). Note that \( [r_k^*] \) satisfies similar recurrence formula as \( [x_k] \), see (5.7), i.e.,

\[
r_{k+1}^* = r_k^* + [p_{k-1}(r_k^* - r_{k-1}^*) - A^2 r_{k-1}^*]/q_k + A^2 x_k.
\]

Applying Theorem 3.1 and Corollary 3.1 we have

\[
\limsup_k \|r_k^*\| \leq 4\|A\| \|A^{-1}\| \limsup_k \|A^2 x_k\|.
\]

Unfortunately, \( \limsup_k \|A^2 x_k\| \) is of order \( \zeta \|\|A\| \|b\| \) and

\[
(6.6) \limsup_k \|r_k^*\| \leq 4\zeta(1 + 4K_1)\|A\| \|A^{-1}\| \|b\|.
\]

Numerical tests of Algorithm 5.1 confirm that (6.6) is sharp which means that in general the Chebyshev method is not well-behaved. Note that direct methods for small dense systems such as Gaussian elimination with pivoting, the Householder method and the Gram-Schmidt reorthogonalization method are well-behaved (see Wilkinson (1965) for two first, Kielbasinski (1974), Kielbasinski and Jankowska (1974) for the last). The lack of well-behavior for the Chebyshev method makes the termination of iteration which is based on \( [r_k] \) difficult. For instance, if we want to find \( x_k \) such that \( \|x_k - x^*\| \leq \varepsilon \), then, in general, we can guarantee the existence of such \( x_k \) only if \( \varepsilon \) is of order \( \zeta \|\|A\| \|A^{-1}\| \|b\| \|x^*\| \) and

However, it can happen that (6.4) holds. Let us mention only two examples (rather theoretical).

If \( [r_k^*] \) from (6.5) is convergent to \( \xi \), \( \|\xi\| \) is of order \( \zeta \|\|b\| \), then applying Corollary (3.2) we get

\[
\limsup_k r_k^* = \left( \frac{s + \sqrt{s^2 - 4}}{2} \right) \xi
\]

from which the well-behavior holds.

Next if \( \limsup_k \|x_k - x_{k-1}\| \leq K_3 \zeta \|\|b\| \) then from condition (ii) of Corollary (3.2) we have

\[
\limsup_k \|r_k^*\| \leq \zeta(2K_3 + 4(1 + 4K_1))\|\|A\| \|\|b\| \|x^*\| \).
\]

ACKNOWLEDGMENT

I would like to thank H. T. Kung and J. F. Traub for their comments on this paper.
REFERENCES

A. Kielbasinski (1974)
A. Kielbasinski and J. Jankowska (1974)
H. Rutishauser, E. Stiefel and others (1959)
E. Stiefel (1958)
J. H. Wilkinson (1963)
J. H. Wilkinson (1965)
H. Wozniakowski (1974)
H. Wozniakowski (1975)
D. Young (1971)

"Numerical Analyses of the Gram-Schmidt Orthogonalization Algorithm," Mat. Stosowana 2, 1974, 15-35 (in Polish).

"Fehleranalyse der Schmidtschen und Powellschen Orthonormalisierungsverfahren," ZAMM 54, T223 (1974).

Refined Iterative Methods for Computation of the Solution and the Eigenvalues of Self-Adjoint Boundary Value Problem

"Kernel Polynomials in Linear Algebra and their Numerical Applications," NBS. Appl. Math., Series 49, 1958, 1-22.

Rounding Errors in Algebraic Processes, Prentice-Hall, Englewood Cliffs, N. J., 1963.

The Algebraic Eigenvalue Problem, Clarendon Press, Oxford, 1965.

"Rounding Error Analysis for the Evaluation of a Polynomial and Some of its Derivatives," SIAM J. Numer. Anal., Vol. 11, No. 4, September, 1974.

Numerical Stability for Solving Nonlinear Equations, Department of Computer Science Report, Carnegie-Mellon University, 1975.

Iterative Solution of Large Linear Systems, Academic Press, New York, 1971.
This paper contains the rounding error analysis for the Chebyshev method for the solution of large linear systems $Ax + g = 0$ where $A = A^T$ is positive definite. We prove that the Chebyshev method in floating point arithmetic is numerically stable, which means that the computed sequence $\{x_k\}$ approximates the solution $x$ such that $\lim_{k \to \infty} ||x_k - x||$ is of order $\zeta ||A|| ||A^{-1}|| ||g||$ where $\zeta$ is the relative computer precision.

We also point out that in general the Chebyshev method is not well-behaved.
which means that $x_k$, $k$ large, is not the exact solution for a slightly perturbed $A$ or equivalently that the computer residuals $r_k = Ax_k + g$ are of order $\epsilon \|A\| \|A^{-1}\| \|x\|$. 