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 GJ32111.

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 $A = A^*$ 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 α such that $\overline{\lim_k}\|x_k-\alpha\|$ is of order $\zeta\|A\|\cdot\|A^{-1}\|\cdot\|\alpha\|$ where ζ 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 $\zeta\|A\|\ \|A^{-1}\|\ \|\alpha\|$.

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 α such that $\|y-\alpha\|$ is of order $\zeta\|A\|\ \|A^{-1}\|\ \|x\|$ where ζ 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., $(A+E)y+g = 0$ where $\|E\|$ 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 $A = A^*$ 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 $b \geq \|A\|$ 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

$$(1.1)\quad x_{k+1} = x_k + \{\rho_{k-1}(x_k-x_{k-1}) - r_k\}/q_k + \xi_k,\ r_k = Ax_k+g,$$

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

In Section 4 we present an algorithm for the computation of ρ_{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 $\overline{\lim_k}\|x_k-\alpha\|$ is of order $\zeta\|A\|\ \|A^{-1}\|\ \|\alpha\|$ 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 $\zeta\|A\|\ \|A^{-1}\|\ \|\alpha\|$ which contradicts well-behavior. However, sometimes r_k can be or order $\zeta\|A\|\ \|\alpha\|$. Such a case yields well-behavior.

2. CHEBYSHEV METHOD

Let us consider the numerical solution of a large linear system

$$(2.1)\quad Ax+g = 0$$

where $A = A^* > 0$ is a given complex nxn matrix and g is a given nx1 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 = Ax 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 $\alpha = -A^{-1}g$ and let

$$(2.2)\quad x_0-\alpha = \sum_{j=1}^{m} c_j v_j$$

where v_j are eigenvectors of A associated with eigenvalues λ_j,

$$Av_j = \lambda_j v_j,\ (v_i,v_j) = \delta_{ij}$$

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 α such that

$$(2.3)\quad x_k-\alpha = W_k(A)(x_0-\alpha)$$

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

$$(2.4)\quad W_k(0) = 1.$$

Remark

Another motivation of (2.3) and (2.4) is to consider a class of iterative methods such that

$$x_k = W_k(A)x_0 + U_k(A)g$$

where W_k and U_k are arbitrary polynomials of degree $\le k$. Assume that if $x_0 = \alpha$ then $x_k \equiv \alpha$ for any α. Then $W_k(x) = U_k(x)\cdot x + 1$ 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

$$\|x_k - \alpha\|_2 \le \|W_k(A)\|_2 \, \|x_0 - \alpha\|_2 \le \|W_k\| \, \|x_0 - \alpha\|_2$$

where

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

Let $P_k(0,1)$ denote a class of polynomials P of degree $\le k$ 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.,

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

and the solution of (2.6) is given by

$$(2.7) \quad W_k(z) = T_k(f(z))/T_k(f(0))$$

where $f(z) = \frac{b+a}{b-a} - 2\frac{z}{b-a}$ 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

$$(2.8) \quad \|x_k - \alpha\|_2 \le 2\left(\frac{\sqrt{b} - \sqrt{a}}{\sqrt{b} + \sqrt{a}}\right)^k \|x_0 - \alpha\|_2$$

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

$$(2.9) \quad x_{k+1} = x_k + (p_{k-1}(x_k - x_{k-1}) - r_k)/q_k, \qquad k = 0,1,\ldots,$$

where $r_k = Ax_k + g$ and

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

$$(2.11) \quad q_0 = \frac{b+a}{2}, \; q_k = \frac{b-a}{4}\,\frac{t_{k+1}}{t_k}, \qquad t_k \equiv T_k(f(0)), \quad k \ge 1.$$

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

Usually, the eigenvalues λ_1 and λ_m from (2.2) are equal to the smallest eigenvalue λ_{min}, and to the

largest eigenvalue, λ_{max}, of A. Hence, the best convergence in T[a,b] is for $a = \lambda_{min}$ and $b = \lambda_{max}$. However, in numerical practice λ_{min} and λ_{max} are known only for a few problems. In many cases we can easily find $b \geq \lambda_{max}$ (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 λ_{min}. Without knowledge of λ_{min} one can use the Chebyshev method T[a,b] for any values of $a > 0$ and $b \geq \lambda_{max}$. Then instead of (2.8) we get

$$\|x_k-\alpha\| \leq 2q(\lambda_{min})^k \|x_0-\alpha\|_2 \tag{2.12}$$

where

$$q(\lambda) = \frac{\sqrt{b-\lambda} + \sqrt{(a-\lambda)_+}}{\sqrt{b-\lambda} - \sqrt{(a-\lambda)_+}} \frac{\sqrt{b} - \sqrt{a}}{\sqrt{b} + \sqrt{a}} \tag{2.13}$$

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 $\lambda \to 0^+$, $q(\lambda) \nearrow 1^-$,

(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 λ_1 from (2.2) is negative.

One can also consider the Chebyshev method for a singular matrix $A = A^* \geq 0$. In such case by α 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 $A\alpha+g_2 = 0$. It is straightforward to verify that $\{x_k\}$ defined by (2.9) in T[a,b] for $a > 0$ and $b \geq \lambda_{max}$, satisfies

$$x_k-\alpha = W_k(A)(x_0-\alpha) + W_k'(0)g_1 \tag{2.14}$$

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

$$x_0-\alpha = c_1v_1 + \sum_{j=1}^{n} c_jv_j$$

where $Av_j = \lambda_jv_j$, $\lambda_1 = 0$, $0 < \lambda_2 < \lambda_3 < \ldots < \lambda_m$, $(v_i,v_j) = \delta_{i,j}$. Note that the normal solution α is orthogonal to v_1. Let us discuss the two cases.

Case I. Let $g_1 = 0$. This means that $Ax+g = 0$ is solvable. From (2.14) it follows

$$\|x_k-\alpha\|_2 \le |c_1| + 2\, q(\lambda_2)^k \|x_0-\alpha\|_2.$$

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.,

$$\|x_k-\alpha\|_2 \le 2\left(\frac{\sqrt{b} - \sqrt{\lambda_2}}{\sqrt{b} + \sqrt{\lambda_2}}\right)^k \|x_0-\alpha\|_2.$$

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

$$y_k-\alpha = W_k(A)(x_0-\alpha) - W_k'(0)W_k(A)A(x_0-\alpha)$$

and for $x_0 = 0$ we get

$$\|y_k-\alpha\| \le 2q(\lambda_2)^k\|\alpha\| + \frac{2k}{\sqrt{ab}}\,\frac{1-q(a)^{2k}}{1+q(a)^{2k}}\, q(\lambda_2)^k\|A\alpha\|,$$

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 $b \ge \lambda_{max}$ 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.

$$(3.1)\quad x_{k+1} = x_k + \{p_{k-1}(x_k-x_{k-1}) - r_k\}/q_k + \xi_k,$$

for suitable vectors ξ_k. A form of ξ_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 $\{\xi_k\}$ 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

$$(3.2)\quad e_{k+1} = e_k + \{p_{k-1}(e_k-e_{k-1}) - Ae_k\}/q_k + \xi_k.$$

Theorem 3.1

Let $\{\xi_k\}$ 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

$$(3.3)\quad e_{k+1} = W_{k+1}(A)e_0 + \sum_{i=0}^{k} \varkappa_{k,i}\{(2-\beta_{i+1})W_{k-i}(A) + (\beta_{i+1}-1)R_{k-i}(A)\}\xi_i$$

where

$$(3.4)\quad \beta_k = 1 + \frac{p_{k-1}}{q_k} = \frac{q_0}{q_k} = 2\,\frac{b+a}{b-a}\,\frac{t_k}{t_{k+1}},\quad 1 \le \beta_k \le 2,$$

$$(3.5)\quad \varkappa_{k,i} = \prod_{j=1}^{k-i-1} \frac{\beta_{j+i+1}}{\beta_j} \quad (\varkappa_{k,k} = \varkappa_{k,k-1} = 1),\ \frac{1}{2} < \varkappa_{k,i} \le 1,\ \lim_{k-i\to\infty} \varkappa_{k,i} = \frac{(\sqrt{b}+\sqrt{a})^2}{2(b+a)},$$

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

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

$$e_1 = W_1(A)e_0 + \varkappa_{0,0}\{2-\beta_1 + \beta_1 - 1\}\xi_0 = e_0 - \frac{1}{q_0} r_0 + \xi_0$$

which is equivalent to (3.2).

Assume now that (3.3) holds for all $i \le k$. Let $B_k = \beta_k I - \frac{1}{q_k} A$, $W_k = W_k(A)$ 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

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

$$(3.8)\quad B_k = \beta_k W_1.$$

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

$$e_{k+1} = \{(1 + \frac{p_{k-1}}{q_k})I - \frac{1}{q_k} A\}e_k - \frac{p_{k-1}}{q_k} e_{k-1} + \xi_k = B_k\{W_k e_0 + \sum_{i=0}^{k-1} \varkappa_{k-1,i}\{(2-\beta_{i+1})W_{k-1-i} +$$

$$+ (\beta_{i+1}-1)R_{k-1-i}\}\xi_i\} + (1-\beta_k)\{W_{k-1}e_0 + \sum_{i=0}^{k-2} \varkappa_{k-2,i}\{(2-\beta_{i+1})W_{k-2-i} + (\beta_{i+1}-1)R_{k-2-i}\}\xi_i\} +$$

$$+ \xi_k = W_{k+1}e_0 + \xi_k + B_k\,\xi_{k-1} + \sum_{i=0}^{k-2} \{(2-\beta_{i+1})\{\varkappa_{k-1,i}B_k W_{k-1-i} + \varkappa_{k-2,i}(1-\beta_k)W_{k-2-i}\}$$

$$+ (\beta_{i+1}-1)\{\varkappa_{k-1,i}B_k\,R_{k-1-i} + \varkappa_{k-2,i}(1-\beta_k)R_{k-2-i}\}\}\xi_i.$$

We want to verify that

(3.9) $B_k = \varkappa_{k,k-1}\{(2-\beta_k)W_1 + (\beta_k-1)R_1\}$,

(3.10) $(2-\beta_{i+1})\{\varkappa_{k-1,i}B_k W_{k-1+i} + \varkappa_{k-2,i}(1-\beta_k)W_{k-2-i}\} + (\beta_{i+1}-1)\{\varkappa_{k-1,i}B_k R_{k-1,i} + \varkappa_{k-2,i}(1-\beta_k)R_{k-2-i}\}$

$= \varkappa_{k,i}\{(2-\beta_{i+1})W_{k-i} + (\beta_{i+1}-1)R_{k-i}\}$, $0 \le i \le k-2$.

Since $\varkappa_{k,k-1} = 1$ and $R_1 = 2W_1$, (3.9) follows from (3.8). To prove (3.10) we use (3.7) which holds for W_{k-i} and R_{k-i}. 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 $\varkappa_{k,i}$,

(3.11) $\varkappa_{k,i}/q_{k-i-1} = \varkappa_{k-1,i}/q_k$,

(3.12) $\varkappa_{k,i}\beta_{k-i-1} = \varkappa_{k-1,i}\beta_k$,

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

From (3.5) and (3.4) it follows

$$\frac{\varkappa_{k,i}}{\varkappa_{k-1,i}} = \frac{\beta_k}{\beta_{k-i-1}} = \frac{q_{k-i-1}}{q_k}$$

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

$$\frac{\varkappa_{k,i}}{\varkappa_{k-2,i}} = \frac{\beta_k\beta_{k-1}}{\beta_{k-1-i}\beta_{k-2-i}} = \frac{t_k}{t_{k+1}} \cdot \frac{t_{k-1-i}}{t_{k-2-i}}; \quad \frac{1-\beta_k}{1-\beta_{k-i-1}} = \frac{p_{k-1}}{q_k} \cdot \frac{q_{k-i-1}}{p_{k-i-2}} = \frac{t_{k-1}}{t_{k+1}} \cdot \frac{t_{k-i}}{t_{k-i-2}},$$

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

$$\varkappa_{k,i} = \frac{t_{i+2}}{t_{k+1}}\frac{t_{k-i}}{t_1} = \frac{1}{1+q(a)^2}\frac{(1+q(a)^{2(i+2)})(1+q(a)^{2(k-i)})}{1+q(a)^{2(k+1)}} \searrow \frac{(\sqrt{b}+\sqrt{a})^2}{2(b+a)} \ge \frac{1}{2}.$$

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 λ_{min} is the smallest eigenvalue of A and $q(\lambda)$ is defined by (2.13).

Corollary 3.1

Let

(3.14) $\delta = \dfrac{\sqrt{b-\lambda}-\sqrt{(a-\lambda)_+}}{\sqrt{b-\lambda}+\sqrt{(a-\lambda)_+}}$, $\eta_k = \dfrac{1-\delta^{2(k+1)}}{1-\delta^2}$ for $k = 0,1,\ldots$, and $q = q(\lambda_{min})$,

Then

(3.15) $$\|e_{k+1}\| \le 2q^{k+1}\|e_0\| + 2\sum_{i=0}^{k} q^{k-i}\,\eta_{k-i}\|\xi_i\|,$$

(3.16) $$e \equiv \limsup_k \|e_k\| \le \frac{2-q(a)\delta}{2\min(a,\lambda_{min})}\,(\sqrt{b}+\sqrt{a})^2\,\xi \le 4\frac{b}{\min(a,\lambda_{min})}\,\xi$$

where $\xi = \limsup_k \|\xi_k\|$. ■

Proof

First of all observe that

$$1 \le \eta_k \le k+1 \text{ and } \eta_k = k+1 \text{ whenever } \lambda_{min} \ge a,$$

$$\|W_k\| \le 2q^k \text{ and } \|R_k\| \le 2q^k\,\eta_k.$$

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

$$\|e_{k+1}\| \le 2q^{k+1}\|e_0\| + \sum_{i=0}^{k} \max(\|W_{k-i}\|,\|R_{k-i}\|)\,\|\xi_i\| \le 2q^{k+1}\|e_0\| + 2\sum_{i=0}^{k} q^{k-i}\eta_{k-i}\|\xi_i\|$$

which proves (3.15).

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

$$e \le 2\limsup_k\Big(\sum_{i=0}^{k_0} q^{k-i}\eta_{k-i}\|\xi_i\| + \sum_{i=k_0+1}^{k} q^{k-i}\eta_{k-i}(\xi+\varepsilon)\Big).$$

Note that $q^{k-i}\eta_{k-i} \to 0$ for $i = 0,1,\ldots,k_0$ and

$$\limsup_k \sum_{i=k_0+1}^{k} q^{k-i}\eta_{k-i} = \sum_{i=0}^{\infty} q^i\eta_i = \frac{1}{1-\delta^2}\Big(\frac{1}{1-q} - \frac{\delta^2}{1-q(a)\delta}\Big) = \frac{(\sqrt{b}+\sqrt{a})^2(2-q(a)\delta)}{4\min(a,\lambda_{min})} = \frac{2b}{\min(a,\lambda_{min})}$$

which proves (3.16) ■

Corollary (3.2)

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

(ii) If $\limsup_k \|x_k - x_{k-1}\| = \varkappa$ then $\limsup_k \|Ae_k - q_k\,\xi_k\| \le \varkappa(b+a)$ ■

Proof

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

$$z_{k+1} = z_k + \{p_{k-1}(z_k - z_{k-1}) - Az_k\}/q_k + \xi_k - \xi + (1-q/q_k)\xi.$$

Applying Corollary 3.1 and Theorem 3.1 to z_k we get

$$\limsup_k \|z_k\| \le 4 \frac{b}{\min(a,\lambda_{\min})} \limsup_k \|\xi_k - \xi + (1-q/q_k)\xi\| = 0$$

which proves conclusion (i).

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

$$Ae_k - q_k \xi_k = p_{k-1}(e_k - e_{k-1}) - q_k(e_{k+1} - e_k).$$

Since $x_k - x_{k-1} = e_k - e_{k-1}$ and $\lim_k p_k = p^* = \left(\frac{\sqrt{b} - \sqrt{a}}{2}\right)^2$ then

$$\limsup_k \|Ae_k - q_k \xi_k\| \le (q^* + p^*)\varkappa = \varkappa(b+a)$$

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

$$(4.1)\quad p_{-1} = 0,\ p_{k-1} = \frac{b-a}{4}\frac{t_{k-1}}{t_k},\ \lim_k p_k = p^* = \left(\frac{\sqrt{b} - \sqrt{a}}{2}\right)^2,$$

$$(4.2)\quad q_0 = \frac{a+b}{2},\ q_k = \frac{b-a}{4}\frac{t_{k+1}}{t_k},\ \lim_k q_k = q^* = \left(\frac{\sqrt{b} + \sqrt{a}}{2}\right)^2,$$

where $t_k = T_k\left(\frac{b+a}{b-a}\right)$, $k \ge 1$.

Let

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

From the recurrence formula of the Chebyshev polynomials it follows

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

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

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

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

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

Algorithm 4.1

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

$$(4.6)\quad p_{-1} = 0,\ q_0 = c,$$

(4.7) $p_0 = \frac{2d}{c}$, $q_1 = \frac{a+b}{4} + \frac{ab}{a+b}$, $\gamma_1 = \frac{2\sqrt{ab}}{a+b}\frac{d}{q_1^*}$,

for $k = 2,3,\ldots$

(4.8) $p_{k-1} = d/q_{k-1}$,

(4.9) $\gamma_k = p_{k-1}\gamma_{k-1}/q^*$,

(4.10) $q_k = q^* - \gamma_k$.

Let us consider the above algorithm in t digit floating point arithmetic, fl, and let rd(x) denote the numerical representation of any real number x and fl(x□y) denote the computed result of an arithmetic operation $\square \in \{+,-,/,\circ\}$. Then

$$rd(x) = x(1+\varepsilon), \quad |\varepsilon| = |\varepsilon(x)| \le \zeta,$$

for x = rd(x) and y = rd(x),

$$fl(x\square y) = (x\square y)(1+\varepsilon), \quad |\varepsilon| = |\varepsilon(x,y,\square)| \le \zeta$$

where $\zeta = 2^{-t}$.

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

$$a(t) = b(t)(1+\varepsilon(t)2^{-t}) \quad \text{where } |\varepsilon(t)| \le K \text{ for } t \ge t_0.$$

Next $a(t) \lesssim b(t)$ iff $a(t) \le b(t)$ or $a(t) \simeq b(t)$. (For more details see e.g. Wozniakowski (1974).)

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

Theorem 4.1

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

(4.11) $\tilde{p}_k = p_k(1+\eta_{p_k})$, $|\eta_{p_k}| \lesssim (4+L_k)\zeta$,

(4.12) $\tilde{q}_k = q_k(1+\eta_{q_k})$, $|\eta_{q_k}| \lesssim L_k\,\zeta$,

where $0 \le L_k \le 15.5 + 64\varkappa$ for $\varkappa = \sqrt{a/b}/(1+\sqrt{a/b})^2$ and $\lim_k L_k = 3.5$ ■

Theorem 4.1 means that we compute p_k and 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

$$\tilde{q}_0 = \tilde{c} = \frac{a+b}{2}(1+\varepsilon_1) = c(1+\eta_c), \quad |\eta_c| \le \zeta,$$

$$\tilde{d} = \left(\frac{(b-a)(1+\varepsilon_2)}{4}\right)^2 (1+\varepsilon_3) = d(1+\eta_d), \quad |\eta_d| \le 3\zeta,$$

$$\tilde{q}^* = \frac{(a+b)(1+\varepsilon_1) + 2\sqrt{ab(1+\varepsilon_4)}(1+\varepsilon_5)}{4}(1+\varepsilon_6) = q^*\left(1 + \frac{\varepsilon_1(a+b) + \{\sqrt{1+\varepsilon_4}(1+\varepsilon_5) - 1\}2\sqrt{ab}}{a+b+2\sqrt{ab}}\right)(1+\varepsilon_6) =$$

$$= q^*(1+\eta_{q^*}), \quad |\eta_{q^*}| \le \frac{5}{2}\zeta,$$

$$\tilde{p}_0 = \frac{2\tilde{d}}{\tilde{c}}(1+\varepsilon_7) = \frac{2d}{c}\frac{(1+\eta_d)}{(1+\eta_c)}(1+\varepsilon_7) = p_0(1+\eta_{p_0}), \quad |\eta_{p_0}| \le 5\zeta,$$

$$\tilde{q}_1 = \left(\frac{(a+b)(1+\varepsilon_1)}{4} + \frac{ab(1+\varepsilon_4)}{(a+b)(1+\varepsilon_1)}(1+\varepsilon_8)\right)(1+\varepsilon_9) = q_1(1+\eta_{q_1}), \quad |\eta_{q_1}| \le 4\zeta.$$

$$\tilde{\gamma}_1 = \frac{2\sqrt{ab(1+\varepsilon_4)}(1+\varepsilon_5)d(1+\eta_d)(1+\varepsilon_{10})}{(a+b)(1+\varepsilon_1)\,q^*(1+\eta_{q^*})(1+\varepsilon_{11})}(1+\varepsilon_{12}) = \gamma_1(1+\eta_{\gamma_1}), \quad |\eta_{\gamma_1}| \le 11\zeta$$

where $|\varepsilon_i| \le \zeta$ for $i = 1,2,\ldots,12$.

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

$$(4.13)\quad \tilde{p}_{k-1} = \frac{d(1+\eta_d)}{q_{k-1}(1+\eta_{q_{k-1}})}(1+\varepsilon_{k,1}) = p_{k-1}(1+\eta_{p_{k-1}}), \quad |\eta_{p_{k-1}}| \le 4\zeta + |\eta_{q_{k-1}}|,$$

$$(4.14)\quad \tilde{\gamma}_k = \frac{p_{k-1}(1+\eta_{p_{k-1}})\gamma_{k-1}(1+\eta_{\gamma_{k-1}})}{q^*(1+\eta_{q^*})}(1+\varepsilon_{k,2}) = \gamma_k(1+\eta_{\gamma_k}), \quad |\eta_{\gamma_k}| \le 8.5\zeta + |\eta_{\gamma_{k-1}}| + |\eta_{q_{k-1}}|,$$

$$(4.15)\quad \tilde{q}_k = (q(1+\eta_{q^*}) - \gamma_k(1+\eta_{\gamma_k}))(1+\varepsilon_{k,3}) = q_k\left(1 + \frac{q^*\eta_{q^*} - \gamma_k\eta_{\gamma_k}}{q_k}\right)(1+\varepsilon_{k,3}) = q_k(1+\eta_{q_k}),$$

$$|\eta_{q_k}| \le (1+2.5\frac{q^*}{q_k})\zeta + \frac{\gamma_k}{q_k}|\eta_{\gamma_k}|.$$

Substituting (4.15) to (4.14) we get

$$(4.16)\quad |\eta_{\gamma_k}| \le \left(12 + 2.5\frac{\gamma_{k-1}}{q_{k-1}}\right)\zeta + \left(1 + \frac{\gamma_k}{q_k}\right)|\eta_{\gamma_{k-1}}|.$$

Note that

$$\frac{\gamma_k}{q_k} = \frac{d}{q_k q^*} \cdot \frac{\gamma_{k-1}}{q_{k-1}} = \left(\prod_{i=2}^{k} \frac{d}{q_i q^*}\right)\frac{\gamma_1}{q_1} = 4\kappa\,\frac{q^{2k}}{1+q^{2k+2}} \le 4\kappa$$

where $\kappa = \sqrt{a/b}\,(1+\sqrt{a/b})^2$ and $q = (\sqrt{b} - \sqrt{a})/(\sqrt{b} + \sqrt{a})$.

Thus, (4.16) becomes

$$(4.17)\quad |\eta_{\gamma_k}| \leq (12 + 10\varkappa)\zeta + (1 + 4\varkappa)|\eta_{\gamma_{k-1}}|$$

Since $|\eta_{\gamma_1}| \leq 11\zeta$, the solution of (4.17) is given by

$$|\eta_{\gamma_k}| \leq (1 + 4\varkappa)^{k-1}11\zeta + (12 + 10\varkappa)\frac{(1+4\varkappa)^{k-1}-1}{4\varkappa}\zeta.$$

Coming back to η_{q_k} we have

$$(4.18)\quad |\eta_{q_k}| \leq 3.5\zeta + 10\varkappa q^{2k}\zeta + 44\varkappa q^{2k}(1 + 4\varkappa)^{k-1}\zeta + (12 + 10\varkappa)q^{2k}[(1+4\varkappa)^{k-1}-1]\zeta = L_k\zeta.$$

Note that

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

Thus,

$$L_k \leq 3.5 + 10\varkappa + 44\varkappa q^2 + (12 + 10\varkappa)q^2 \leq 15.5 + 64\varkappa,$$

and

$$\lim_k L_k = 3.5.$$

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

$$(5.1)\quad \limsup_k \|x_k-\alpha\| \leq \zeta K\|A\|\,\|A^{-1}\|\,\|\alpha\| + O(\zeta^2)$$

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,

$$(5.2)\quad r_k := Ax_k+g;$$

Proof

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

$$\tilde{q}_0 = \tilde{c} = \frac{a+b}{2}(1+\varepsilon_1) = c(1+\eta_c), \quad |\eta_c| \le \zeta,$$

$$\tilde{d} = \left(\frac{(b-a)(1+\varepsilon_2)}{4}\right)^2 (1+\varepsilon_3) = d(1+\eta_d), \quad |\eta_d| \le 3\zeta,$$

$$\tilde{q}^* = \frac{(a+b)(1+\varepsilon_1) + 2\sqrt{ab(1+\varepsilon_4)}\,(1+\varepsilon_5)}{4}(1+\varepsilon_6) = q^*\left(1 + \frac{\varepsilon_1(a+b) + \{\sqrt{1+\varepsilon_4}(1+\varepsilon_5) - 1\}2\sqrt{ab}}{a+b+2\sqrt{ab}}\right)(1+\varepsilon_6) =$$

$$= q^*(1+\eta_{q^*}), \quad |\eta_{q^*}| \le \frac{5}{2}\zeta,$$

$$\tilde{p}_0 = \frac{2\tilde{d}}{\tilde{c}}(1+\varepsilon_7) = \frac{2d}{c}\frac{(1+\eta_d)}{(1+\eta_c)}(1+\varepsilon_7) = p_0(1+\eta_{p_0}), \quad |\eta_{p_0}| \le 5\zeta,$$

$$\tilde{q}_1 = \left(\frac{(a+b)(1+\varepsilon_1)}{4} + \frac{ab(1+\varepsilon_4)}{(a+b)(1+\varepsilon_1)}(1+\varepsilon_8)\right)(1+\varepsilon_9) = q_1(1+\eta_{q_1}), \quad |\eta_{q_1}| \le 4\zeta.$$

$$\tilde{\gamma}_1 = \frac{2\sqrt{ab(1+\varepsilon_4)}(1+\varepsilon_5)d(1+\eta_d)(1+\varepsilon_{10})}{(a+b)(1+\varepsilon_1)\,q^*(1+\eta_{q^*})(1+\varepsilon_{11})}(1+\varepsilon_{12}) = \gamma_1(1+\eta_{\gamma_1}), \quad |\eta_{\gamma_1}| \le 11\zeta$$

where $|\varepsilon_i| \le \zeta$ for $i = 1,2,\ldots,12$.

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

$$(4.13)\quad \tilde{p}_{k-1} = \frac{d(1+\eta_d)}{q_{k-1}(1+\eta_{q_{k-1}})}(1+\varepsilon_{k,1}) = p_{k-1}(1+\eta_{p_{k-1}}), \quad |\eta_{p_{k-1}}| \le 4\zeta + |\eta_{q_{k-1}}|,$$

$$(4.14)\quad \tilde{\gamma}_k = \frac{p_{k-1}(1+\eta_{p_{k-1}})\gamma_{k-1}(1+\eta_{\gamma_{k-1}})}{q^*(1+\eta_{q^*})}(1+\varepsilon_{k,2}) = \gamma_k(1+\eta_{\gamma_k}), \quad |\eta_{\gamma_k}| \le 8.5\zeta + |\eta_{\gamma_{k-1}}| + |\eta_{q_{k-1}}|,$$

$$(4.15)\quad \tilde{q}_k = (q(1+\eta_{q^*}) - \gamma_k(1+\eta_{\gamma_k}))(1+\varepsilon_{k,3}) = q_k\left(1 + \frac{q^*\eta_{q^*} - \gamma_k\eta_{\gamma_k}}{q_k}\right)(1+\varepsilon_{k,3}) = q_k(1+\eta_{q_k}),$$

$$|\eta_{q_k}| \le (1+2.5\frac{q^*}{q_k})\zeta + \frac{\gamma_k}{q_k}|\eta_{\gamma_k}|.$$

Substituting (4.15) to (4.14) we get

$$(4.16)\quad |\eta_{\gamma_k}| \le \left(12 + 2.5\frac{\gamma_{k-1}}{q_{k-1}}\right)\zeta + \left(1 + \frac{\gamma_k}{q_k}\right)|\eta_{\gamma_{k-1}}|.$$

Note that

$$\frac{\gamma_k}{q_k} = \frac{d}{q_k q^*} \cdot \frac{\gamma_{k-1}}{q_{k-1}} = \left(\prod_{i=2}^{k} \frac{d}{q_i q^*}\right)\frac{\gamma_1}{q_1} = 4\varkappa\,\frac{q^{2k}}{1+q^{2k+2}} \le 4\varkappa$$

where $\varkappa = \sqrt{a/b}\,(1+\sqrt{a/b})^2$ and $q = (\sqrt{b} - \sqrt{a})/(\sqrt{b} + \sqrt{a})$.

Thus, (4.16) becomes

$$(4.17)\quad |\eta_{\gamma_{ik}}| \leq (12 + 10\varkappa)\zeta + (1 + 4\varkappa)|\eta_{\gamma_{k-1}}|$$

Since $|\eta_{\gamma_1}| \leq 11\zeta$, the solution of (4.17) is given by

$$|\eta_{\gamma_k}| \leq (1 + 4\varkappa)^{k-1}11\zeta + (12 + 10\varkappa)\frac{(1+4\varkappa)^{k-1}-1}{4\varkappa}\zeta.$$

Coming back to η_{q_k} we have

$$(4.18)\quad |\eta_{q_k}| \leq 3.5\zeta + 10\varkappa q^{2k}\zeta + 44\varkappa q^{2k}(1 + 4\varkappa)^{k-1}\zeta + (12 + 10\varkappa)q^{2k}\{(1+4\varkappa)^{k-1}-1\}\zeta = L_k\zeta.$$

Note that

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

Thus,

$$L_k \leq 3.5 + 10\varkappa + 44\varkappa q^2 + (12 + 10\varkappa)q^2 \leq 15.5 + 64\varkappa,$$

and

$$\lim_k L_k = 3.5.$$

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

$$(5.1)\quad \limsup_k \|x_k-\alpha\| \leq \zeta K\|A\|\,\|A^{-1}\|\,\|\alpha\| + O(\zeta^2)$$

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,...

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

$$(5.2)\quad r_k := Ax_k+g;$$

(5.3) $x_{k+1} := x_k + (p_{k-1}(x_k - x_{k-1}) - r_k)/q_k.$

Theorem 5.1

Let $\{x_k\}$ be the sequence computed in fl arithmetic by Algorithm 5.1. If

(5.4) $fl(Ax_k+g) = (I+\delta L_k)((A+E_k)x_k + g)$

where $\|E_k\| \leq K_1\zeta\|A\|$ and $\|\delta L_k\| \leq K_2\zeta$, $K_i = K_i(n)$ for $i = 1,2$,

then for small ζ,

(5.5) $$\limsup_k \|x_k-\alpha\| \leq 4(1 + 4K_1)\zeta \frac{b}{\min(a,\lambda_{min})} \|\alpha\| \Big/ \left(1 - \frac{4b(59+4K_1+4K_2)}{\min(a,\lambda_{min})}\zeta\right).$$ ■

Note that assumption (5.4) holds for the standard algorithm for the computation of Ax_k+g and $K_1 \leq n\sqrt{n}$, $K_2 = 1$ for any matrix A and any vector g (see Wilkinson (1963), p. 83). Due to sparseness of A the constant K_1 usually depends on the maximal number of nonzero elements in rows of A.

Proof

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

(5.6) $$x_{k+1} = (I + D_k^1)\{x_k + (I + 2D_k^2)[p_{k-1}(1 + \eta_{p_{k-1}})(I + 2D_k^3)(x_k - x_{k-1}) - (I + \delta L_k)(Ax_k+g+E_kx_k)]/ /q_k(1 + \eta_{q_k})\}.$$

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

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

where

(5.8) $$\xi_k = D_k^1\alpha - \frac{1}{q_k}E_k\alpha + \Theta_k$$

and $$\|\Theta_k\| \leq \zeta(10 + L_k + L_{k-1} + (3 + L_k + K_1 + K_2)\|A\|/q_k)\|e_k\| + \zeta(9 + L_k + L_{k-1})\|e_{k-1}\| + \zeta^2K_1(3 + K_2 + L_k)\|A\|\,\|\alpha\|/q_k.$$

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

$$\limsup_k \|\xi_k\| \leq \zeta\|\alpha\|(1 + 4K_1) + \zeta \limsup_k \|e_k\|(59 + 4(K_1+K_2)).$$

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

$$e \leq \frac{4(1+4K_1)\zeta b}{\min(a,\lambda_{min})}\|\alpha\| + \frac{4b(59+4K_1+4K_2)}{\min(a,\lambda_{min})}\zeta e.$$

Hence, (5.5) follows from the last relation which completes the proof. ■

From Theorem 5.1 we can easily get (5.1). Since $A = A^* > 0$ then $\|A\| \, \|A^{-1}\| = \lambda_{max}/\lambda_{min}$. It leads us to the following

Corollary 5.1

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

$$(5.9) \quad \frac{b}{\min(a,\lambda_{min})} \leq L \frac{\lambda_{max}}{\lambda_{min}}$$

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

$$(5.10) \quad \limsup_k \|x_k - \alpha\| \leq \zeta \, 4(1 + K_1)L \, \|A\| \, \|A^{-1}\| \, \|\alpha\| + O(\zeta^2)$$

where K_1 is defined by (5.4). ■

Proof

From Theorem 5.1 and from the definition of the relation $\lesssim$ it follows

$$e \leq 4(1 + 4K_1)L \, \|A\| \, \|A^{-1}\| \, \|\alpha\| \, \zeta(1 + O(\zeta))$$

which gives (5.10). ■

If ζ is small then one can prove that the constant which appears in the "O" notation in (5.10) only depends on $(\|A\| \, \|A^{-1}\|)^2$, K_1 and K_2 (see (5.4). Note that if $b \leq L^* \lambda_{max}$ then for any $a \geq \lambda_{min}$, (5.9) holds; however, for increasing a 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^1, E_k and $\ominus_k$ 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 $a = \lambda_{min}$ and $b = \lambda_{max}$. From (5.4) and (5.6) we know that D_k^1 and E_k are small but arbitrary. Assume theoretically that $D_k^1 = 0$, $\ominus_k = 0$ and $\lim_k E_k = E$ where $\|A^{-1}E\alpha\| = \|A^{-1}\| \, \|E\| \, \|\alpha\|$ and $\|E\| = K_1 \zeta \|A\|$. From Corollary (3.2) we get

$$\lim_k e_k = -A^{-1}E\alpha \quad \text{and} \quad e = \zeta K_1 \|A\| \, \|A^{-1}\| \, \|\alpha\|$$

which is essentially the righthand side of (5.10). Furthermore, if $E_k = 0$, $\ominus_k = 0$ and $\lim_k D_k^1 = D$ where $\|A^{-1}D\alpha\| = \|A^{-1}\| \, \|D\| \, \|\alpha\|$, $\|D\| = \zeta$, we have

$$\lim_k e_k = \frac{\sqrt{b} + \sqrt{a}}{2}^2 A^{-1}D\alpha, \quad e = \zeta((1 + \sqrt{a/b}/2)^2 \|A\| \, \|A^{-1}\| \, \|\alpha\|.$$

This implies that even using the double precision for the evaluation of $Ax+g$, $\|E_k\| \lesssim K_1 \zeta^2 \|A\|$, we cannot guarantee the high relative precision of the computed $\{x_k\}$.

Thus, (5.10) is sharp. Although, from (5.8) we get

$$\limsup_k \|x_k - \alpha\| \le 4\|A\|\,\|A^{-1}\| \limsup_k \|D_k^1\alpha - \frac{1}{q_k} E_k\alpha\|\,(1 + O(\zeta))$$

and if $\|D_k^1\alpha - \frac{1}{q_k} E_k\alpha\| \ll \zeta\|\alpha\|$ 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.,

(6.1) $(A + \delta A)(y + \delta y) + g + \delta g = 0$

where $\|\delta A\| \lesssim \delta\, c_1\|A\|$, $\|\delta y\| \lesssim \zeta\, c_2\|y\|$ and $\|\delta g\| \lesssim \zeta\, c_3\|g\|$, $c_i = c_i(n)$.

Let Δy and Δg be matrices defined by

$$(I + \Delta y)y = y + \delta y;\quad (I + \Delta g)g = g + \delta g$$

and

$$\|\Delta y\| \lesssim \zeta\, c_2,\quad \|\Delta g\| \lesssim \zeta\, c_3.$$

Hence, (6.1) becomes

(6.2) $(A + \Delta A)y + g = 0$

where $\|\Delta A\| \lesssim \zeta\, c_4\, \|A\|$ 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 $r = fl(Ay + g)$ be the computed residual vector. Assume

(6.3) $r = (I + \Delta I)((A+E)y + g)$

where $\|\Delta I\| \lesssim \zeta\, c_5$ and $\|E\| \lesssim \zeta\, c_6\, \|A\|$.

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

(6.4) $\|r\| \lesssim \zeta\, c_7\, \|A\|\,\|y\|$.

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

$$\left(A + E - (I + \Delta I)^{-1}\frac{ry^*}{\|y\|^2}\right)y + g = 0.$$

Thus, $\Delta A = E - (I + \Delta I)^{-1}\frac{ry^*}{\|y\|^2}$ and $\|\Delta A\| \lesssim \zeta(c_6 + c_7)\,\|A\|$.

We wish to consider the well-behavior problem for the Chebyshev method T[a,b]. This means we must

verify if the computed vectors $r_k = fl(Ax_k+g)$ satisfies condition (6.4) for large k. From (5.4) we get

$$\|r_k - r_k^*\| \lesssim K_1 \zeta \, \|A\| \, \|x_k\|$$

where $r_k^* = Ae_k$.

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

$$(6.5) \quad r_{k+1}^* = r_k^* + \{p_{k-1}(r_k^* - r_{k-1}^*) - Ar_k^*\}/q_k + A\xi_k.$$

Applying Theorem 3.1 and Corollary 3.1 we have

$$\limsup_k \|r_k^*\| \lesssim 4\|A\| \, \|A^{-1}\| \limsup_k \|A\xi_k\|.$$

Unfortunately, $\limsup_k \|A\xi_k\|$ is of order $\zeta\|A\| \, \|\alpha\|$ and

$$(6.6) \quad \limsup_k \|r_k^*\| \lesssim 4\zeta(1 + 4K_1)\|A\| \, \|A^{-1}\| \, \|\alpha\|.$$

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 $\|r_k\| \le \varepsilon\|r_0\|$ then, in general, we can guarantee the exitence of such x_k only if ε is of order $\zeta\|A\| \, \|A^{-1}\| \, \|\alpha\|/\|r_0\|$.

However, it can happen that (6.4) holds. Let us mention only two examples (rather theoretical). If $\{\xi_k\}$ from (6.5) is convergent to ξ, $\|\xi\|$ is of order $\zeta\|\alpha\|$, then applying Corollary (3.2) we get

$$\lim_k r_k^* = \left(\frac{\sqrt{b} + \sqrt{a}}{2}\right)^2 \xi$$

from which the well-behavior holds.

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

$$\limsup_k \|r_k^*\| \lesssim \zeta(2K_3 + 4(1 + 4K_1))\|A\| \, \|\alpha\|.$$

ACKNOWLEDGMENT

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

REFERENCES

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

A. Kielbasinski and J. Jankowska (1974) "Fehleranalyse der Schmidtschen und Powellschen Orthonormalisierungsverfahren," ZAMM 54, T223 (1974).

H. Rutishauser, E. Stiefel and others (1959) Refined Iterative Methods for Computation of the Solution and the Eigenvalues of Self-Adjoint Boundary Value Problem

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

J. H. Wilkinson (1963) Rounding Errors in Algebraic Processes, Prentice-Hall, Englewood Cliffs, N. J., 1963.

J. H. Wilkinson (1965) The Algebraic Eigenvalue Problem, Clarendon Press, Oxford, 1965.

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

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

D. Young (1971) Iterative Solution of Large Linear Systems, Academic Press, New York, 1971.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE *(and Subtitle)* NUMERICAL STABILITY OF THE CHEBYSHEV METHOD FOR THE SOLUTION OF LARGE LINEAR SYSTEMS		5. TYPE OF REPORT & PERIOD COVERED Interim
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) H. Wozniakowski		8. CONTRACT OR GRANT NUMBER(s) N0014-67-A-0314-0010, NR 044-422
9. PERFORMING ORGANIZATION NAME AND ADDRESS Carnegie-Mellon University Dept. of Computer Science Pittsburgh, PA 15213		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Arlington, VA 22217		12. REPORT DATE March 1975
		13. NUMBER OF PAGES 20
14. MONITORING AGENCY NAME & ADDRESS*(if different from Controlling Office)*		15. SECURITY CLASS. *(of this report)* UNCLASSIFIED
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for Public Release; Distribution unlimited.

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)* This paper contains the rounding error analysis for the Chebyshev method for the solution of large linear systems $Ax+g = 0$ where $A = A^*$ 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 α such that $\overline{\lim_k} \|x_k - \alpha\|$ is of order $\zeta \|A\| \cdot \|A^{-1}\| \cdot \|\alpha\|$ where ζ is the relative computer precision. We also point out that in general the Chebyshev method is not well-behave,

CONTINUED

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

20. abstract CONTINUED

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 $\epsilon \|A\| \|A^{-1}\| \|x\|$.