

On the convergence rate of the conjugate gradients in presence of rounding errors*

Yvan Notay**

Service de Métrologie Nucléaire, Université Libre de Bruxelles (C.P. 165), 50, Avenue F.D. Roosevelt, B-1050 Brussels, Belgium

Received September 26, 1991/Revised version received October 12, 1992.

Summary. We investigate here rounding error effects on the convergence rate of the conjugate gradients. More precisely, we analyse on both theoretical and experimental basis how finite precision arithmetic affects known bounds on iteration numbers when the spectrum of the system matrix presents small or large isolated eigenvalues.

Mathematics Subject Classification (1991): 65 F10, 65 B99

1. Introduction

There are many excellent works which analyse the convergence rate of the conjugate gradient method for solving the symmetric positive definite system

$$(1.1) \quad Au = b,$$

among which Axelsson [2], Andersson [1], Jennings [8] and Axelsson and Lindskog [4], where explicit bounds on the number of iterations are derived under simple assumptions on the eigenvalue distribution, are particularly relevant for the discussion of preconditioning techniques. All of them however make the assumption of exact arithmetic computation and, since rounding errors imply some loss of orthogonality (see for instance [7]), their validity in the context of finite precision arithmetic remains subject to further examination.

Greenbaum made recently [5] a thorough stability analysis of the conjugate gradient algorithm, mentioning already as corollary some conclusions about rounding error effects on the convergence rate. The convergence behaviour was further investigated by Strakos [10] and by Geenbaum and Strakos [6], who found

* The present work was supported by the "Programme d'impulsion en Technologie de l'Information", financed by Belgian State, under contract No. IT/IF/14

** Supported by the "Fonds National de la Recherche Scientifique", Chargé de recherches

reasonable agreement with the predictions of Greenbaum's theory. Our purpose is to pursue these works by carefully analysing the validity of the bounds on the number of iterations derived when the spectrum presents isolated eigenvalue(s) at (one of) its ends. We also compare our results with the experiment by running the conjugate gradient algorithm in single and double precision for some characteristic eigenvalue distributions.

The paper is organised as follows: standard definitions and notation are given below, together with needed properties of the Chebyshev polynomials; the conjugate gradient algorithm and its basic properties (including Greenbaum stability analysis) are recalled in Sect. 2, and the convergence rate in presence of rounding errors is discussed in Sects. 3 and 4, where we successively consider the case of small and large isolated eigenvalues.

Notation. The matrix A of the linear system (1.1) to be solved is an $n \times n$ symmetric positive definite matrix. Its eigenvalues are denoted in the following way: $v_{\min}^{(i)}$ is the i^{th} eigenvalue, and $v_{\max}^{(i)}$ the i^{th} eigenvalue by decreasing order (i.e. $v_{\max}^{(i)} = v_{\min}^{(n-i+1)}$); we also write sometimes v_{\min} for $v_{\min}^{(1)}$ and v_{\max} for $v_{\max}^{(1)}$.

The scalar product in \mathbb{R}^n is denoted (x, y) . For any $n \times n$ symmetric positive definite matrix B , $\sigma(B)$ denotes the spectrum of B and $\|x\|_B$ the B -norm of x , i.e. $\|x\|_B = \sqrt{(x, Bx)}$.

Chebyshev polynomials. For $k \geq 0$, we define the polynomials

$$(1.2) \quad \mathcal{P}_k(a, b, x) = T_k\left(\frac{a+b-2x}{a-b}\right)$$

where T_k is the Chebyshev polynomial of the first kind, i.e.

$$(1.3) \quad T_k(x) = \begin{cases} \cos(k \arccos x) & \text{if } |x| \leq 1 \\ \frac{1}{2} [(x + \sqrt{x^2 - 1})^k + (x - \sqrt{x^2 - 1})^k] & \text{if } |x| \geq 1 \end{cases}$$

We also use $\mathcal{P}'_k(a, b, x)$ to denote $(\partial/\partial\xi)\mathcal{P}_k(a, b, \xi)|_{\xi=x}$. We quote below needed properties of these polynomials; (1.4) is obvious from (1.2), (1.3); (1.5) and (1.6) are standard results (see for instance [3] for a proof); (1.7), (1.8) and (1.9) follow straightforwardly from (1.2) and (1.3); (1.10), (1.11) and (1.12) are proved in the Appendix; (1.13) obviously follows from (1.11) and (1.12).

- If $a \leq b$,

$$(1.4) \quad |\mathcal{P}_k(a, b, x)| \leq 1 \quad \text{for all } a \leq x \leq b.$$

- If $0 < a < b$,

$$(1.5) \quad \mathcal{P}_k(a, b, 0) = \frac{1}{2} \left(\left(\frac{\sqrt{b/a} + 1}{\sqrt{b/a} - 1} \right)^k + \left(\frac{\sqrt{b/a} - 1}{\sqrt{b/a} + 1} \right)^k \right)$$

and therefore

$$(1.6) \quad k = \text{int} \left[\frac{1}{2} \sqrt{\frac{b}{a}} \ln \frac{2}{\varepsilon} \right] + 1 \Rightarrow \mathcal{P}_k(a, b, 0) > \frac{1}{\varepsilon}.$$

- If $0 < a < b$,

$$(1.7) \quad |\mathcal{P}_k(a, b, x)| < \mathcal{P}_k(a, b, 0) \quad \text{for all } 0 < x < b.$$

- If $b > 0$ and $a = -b \tan^2 \pi/4k$,

$$(1.8) \quad \mathcal{P}_k(a, b, 0) = 0$$

$$(1.9) \quad |\mathcal{P}'_k(a, b, 0)| = \frac{k}{b} \cot \frac{\pi}{4k}$$

$$(1.10) \quad |\mathcal{P}_k(a, b, x)| \leq x |\mathcal{P}'_k(a, b, 0)| \quad \text{for all } 0 \leq x \leq b.$$

- If $c \gg b \gg \delta > 0$ and $|a| < b$,

$$(1.11) \quad |\mathcal{P}_k(a, b, c)| = \frac{1}{2} \left(\frac{4c}{b-a} \right)^k \left(1 + O\left(\frac{b}{c}\right) \right),$$

$$(1.12) \quad \mathcal{P}_k(b - \delta, b + \delta, 0) = \frac{1}{2} \left(\frac{2b}{\delta} \right)^k \left(1 + O\left(\frac{\delta^2}{b^2}\right) \right),$$

and therefore

$$(1.13) \quad \left| \frac{\mathcal{P}_k(b - \delta, b + \delta, c)}{\mathcal{P}_k(b - \delta, b + \delta, 0)} \right| = \left(\frac{c}{b} \right)^k \left(1 + O\left(\frac{b}{c}\right) \right) \left(1 + O\left(\frac{\delta^2}{b^2}\right) \right).$$

2. Algorithm and basic properties

We recall in the following algorithm the most popular implementation of the conjugate gradient method for solving the positive definite linear system $Au = b$.

Algorithm 2.1. Given an initial guess u_0 , set $r_0 = b - Au_0$, $\delta_0 = r_0$ and execute, for $k = 0, 1, \dots$:

$$\begin{aligned} a_k &= \frac{(r_k, r_k)}{(\delta_k, A\delta_k)} \\ r_{k+1} &= r_k - a_k A\delta_k \\ u_{k+1} &= u_k + a_k \delta_k \\ d_{k+1} &= \frac{(r_{k+1}, r_{k+1})}{(r_k, r_k)} \\ \delta_{k+1} &= r_{k+1} + d_{k+1} \delta_k \end{aligned}$$

The basic properties of this algorithm in exact arithmetic are recalled in the following theorem.

Theorem 2.1. Let A be a positive definite matrix, b some given vector and $\hat{u} = A^{-1}b$ the solution to (1.1). Letting u_k be the vector obtained after k iterations of Algorithm 2.1, one has

$$(2.1) \quad \|\hat{u} - u_k\|_A = \min_{P_k \in \Pi_k} \|P_k(A)(\hat{u} - u_0)\|_A$$

and

$$(2.2) \quad \frac{\|\hat{u} - u_k\|_A}{\|\hat{u} - u_0\|_A} \leq \min_{P_k \in \Pi_k^1} \max_{v \in \sigma(A)} |P_k(v)|$$

where Π_k^1 denotes the set of polynomials of order k satisfying $P_k(0) = 1$.

As is well known, rounding errors imply some loss of orthogonality, so that formulas (2.1) and (2.2) are no more valid in finite precision arithmetic. One may however deduce the following results from Greenbaum stability analysis of Algorithm 2.1 [5] (Theorems 1' and 3').

- (1) There exists a matrix Φ and a vector β such that the coefficients $a_k, k \geq 0$ and $d_k, k \geq 1$ generated by a perturbed sequence applied to the system $Au = b$ are equal to those generated by an unperturbed sequence applied to $\Phi\psi = \beta$. Further,

$$(2.3) \quad \sigma(\Theta) \subset \bigcup_{v \in \sigma(A)} [v - \tau, v + \tau]$$

where $\tau \ll \|A\|$ depends on the magnitude of the perturbations.

- (2) Letting $e_k = A^{-1}r_k$ be the error vector at the k^{th} step of the perturbed sequence, and $\bar{e}_k = \Theta^{-1}\bar{r}_k$ the error vector at the k^{th} step of the unperturbed sequence, one has

$$(2.4) \quad \frac{\|e_k\|_A}{\|e_0\|_A} = \left(1 + O\left(\frac{\tau}{\|A\|}\right)\right) \frac{\|\bar{e}_k\|_\Theta}{\|\bar{e}_0\|_\Theta}.$$

Greenbaum [5] gives an upper bound on τ that seems to strongly overestimate perturbation effects. However [6], the actual convergence behaviour in a practical context of finite precision calculation compared with that of the "exact" (i.e. with full reorthogonalization) algorithm applied to a matrix Φ satisfying (2.3) with 11 eigenvalues in each tiny interval shows a remarkable agreement provided that one uses a rather small value for τ , say about $50\eta\|A\|$, where η is the machine precision.

Therefore, depending on the value of τ which is used, the above mentioned results may be seen either as rigorous (large) upper bounds, or only as semi-empirical (close) estimates of rounding error effects. The choice of the point of view will be left open in the following; as one will see, it has anyway little influence on our main conclusions provided that ones discards, as we shall do, the cases where v_{\min} is so small that $(v_{\min} - \tau) / v_{\min}$ is much less than 1 or negative.

Combining (2.3) and (2.4), we obtain

$$(2.5) \quad \frac{\|e_k\|_A}{\|e_0\|_A} \leq \left(1 + O\left(\frac{\tau}{\|A\|}\right)\right) \min_{P_k \in \Pi_k^1} \max_{v \in \sigma(A)} \max_{x \in [v - \tau, v + \tau]} |P_k(x)|$$

which will be the basis of our analysis.

Consider now the standard estimate based on the polynomials

$$P_k(x) = \frac{\mathcal{P}_k(v_{\min}, v_{\max}, x)}{\mathcal{P}_k(v_{\min}, v_{\max}, 0)},$$

i.e

$$(2.6) \quad k_\epsilon \leq \text{int} \left[\frac{1}{2} \sqrt{\kappa \ln \frac{2}{\epsilon}} \right] + 1$$

for the number of iterations k_ε necessary to reduce the relative error in the A -norm by a factor ε . We get as a first consequence, already mentioned in [5], that this bound is valid provided that one defines κ by

$$(2.7) \quad \kappa = \frac{v_{\max} + \tau}{v_{\min} - \tau}$$

rather than by $\kappa = v_{\max}/v_{\min}$, which has nearly no practical effects. The same conclusion holds for any bound on k_ε based on polynomials P_k which guarantee $|P_{k_\varepsilon}(v)| < \varepsilon$ for all $v \in \sigma(A)$ by assuming $\sigma(A)$ included in the union of intervals of nonzero lengths (see [1, 2] for examples of such bounds).

In the presence of isolated eigenvalues, suitable families of polynomials are generally obtained by enforcing $P_k(v) = 0$ at these values. It is then clear from the results above that these polynomials will lead to valid bounds if and only if $P_k(v)$ is not only zero at these values, but also kept close to zero for all v belonging to small intervals around these values. This will be further discussed in the following sections for isolated eigenvalues at the ends of the spectrum.

3. Small isolated eigenvalues

We first derive a bound on the convergence rate in exact arithmetic. For this purpose, let us use the polynomials P_k proposed by Axelsson and Lindskog [4], i.e.

$$(3.1) \quad P_k(v) = \frac{\mathcal{P}_{k-r}(v_{\min}^{(p)}, v_{\max}, v)}{\mathcal{P}_{k-r}(v_{\min}^{(p)}, v_{\max}, 0)} \prod_{i=1}^{p-1} \frac{v^{-1} \mathcal{P}_{r_i}(a_i, v_{\max}, v)}{\mathcal{P}'_{r_i}(a_i, v_{\max}, 0)} \left(1 - \frac{v}{v_{\min}^{(i)}}\right)$$

where $p \geq 1$ and $r_i \geq 1, i = 1, \dots, p-1$ are integers such that $r = \sum_i r_i \leq k$ and $a_i, i = 1, \dots, p-1$ numbers such that $\mathcal{P}_{r_i}(a_i, v_{\max}, 0) = 0$, so that $v^{-1} \mathcal{P}_{r_i}(a_i, v_{\max}, v)$ is effectively a polynomial while $P_k(0) = 1$ follows from l'Hospital rule. More particularly, we set (cf. (1.8))

$$(3.2) \quad a_i = -v_{\max} \tan^2 \frac{\pi}{4r_i},$$

and, since $P_k(v_{\min}^{(i)}) = 0$ for $i = 1, \dots, p-1$, one then obtains, with (1.4), (1.9):

$$(3.3) \quad \max_{v \in \sigma(A)} |P_k(v)| < \frac{1}{\mathcal{P}_{k-r}(v_{\min}^{(p)}, v_{\max}, 0)} \prod_{i=1}^{p-1} \frac{1}{|\mathcal{P}'_{r_i}(a_i, v_{\max}, 0)| v_{\min}^{(i)}}$$

$$(3.4) \quad = \frac{1}{\mathcal{P}_{k-r}(v_{\min}^{(p)}, v_{\max}, 0)} \prod_{i=1}^{p-1} \frac{v_{\max} \tan \frac{\pi}{4r_i}}{v_{\min}^{(i)}}.$$

With (1.6), the latter expression leads to

$$(3.5) \quad k_\varepsilon \leq \text{int} \left[\frac{1}{2} \sqrt{\frac{v_{\max}}{v_{\min}^{(p)}}} \left(\ln \frac{2}{\varepsilon} + \sum_{i=1}^{p-1} \ln \frac{v_{\max} \tan \frac{\pi}{4r_i}}{v_{\min}^{(i)} r_i} \right) \right] + r + 1$$

for the maximal number of iterations k_ε necessary to reduce the A -norm of the relative error by a factor ε . The optimal r_i are difficult to find, but it turns out from the discussion in [4] that a nearly optimal choice is obtained by letting

$$(3.6) \quad r_i = \text{int} \left[\sqrt{\frac{v_{\max}}{v_{\min}^{(p)}}} \right] + 1,$$

which gives, using $\tan(\pi/4r) \leq 1/r$ for $r \geq 1$,

$$(3.7) \quad k_\varepsilon \leq \text{int} \left[\frac{1}{2} \sqrt{\frac{v_{\max}}{v_{\min}^{(p)}}} \left(\ln \frac{2}{\varepsilon} + \sum_{i=1}^{p-1} \ln \frac{v_{\min}^{(p)}}{v_{\min}^{(i)}} \right) \right] + (p-1) \text{int} \left[\sqrt{\frac{v_{\max}}{v_{\min}^{(p)}}} + 1 \right] + 1,$$

where the only remaining parameter is p ; note that $p = 1$ gives the standard estimate (2.6). Note also that, by contrast to similar expressions derived in [4, 8], the bound (3.7) is not an asymptotic estimate, i.e. it is valid whatever the values of v_{\max} , $v_{\min}^{(i)}$, $i = 1, \dots, p$.

We now discuss its validity in the presence of rounding errors. As already mentioned, this means that we have to analyse the behaviour of $P_k(v)$ for $v \in [v_{\min}^{(i)} - \tau, v_{\min}^{(i)} + \tau]$, $i = 1, \dots, p-1$. Using (1.7) and (1.10), one readily obtains

$$(3.8) \quad \max_{v \in [v_{\min}^{(i)} - \tau, v_{\min}^{(i)} + \tau]} |P_k(v)| < \frac{\tau}{v_{\min}^{(i)}} \prod_{\substack{j=1 \\ j \neq i}}^{p-1} \left| \frac{v_{\min}^{(i)}}{v_{\min}^{(j)}} - 1 \right| + O\left(\frac{\tau}{v_{\min}^{(i)}}\right)^2,$$

and the bound derived above is thus still valid as long as

$$(3.9) \quad \varepsilon > \tau \frac{(v_{\min}^{(p-1)})^{p-2}}{\prod_{j=1}^{p-1} v_{\min}^{(j)}}.$$

To better understand the significance of this restriction, we made the following experiment: we ran Algorithm 2.1 and computed

$$\frac{\|e_k\|_A}{\|e_0\|_A} = \frac{\|r_k\|_{A^{-1}}}{\|r_0\|_{A^{-1}}}$$

for the linear system $Au = b$ characterized by

$$A = \text{diag}(\lambda_i), \quad b_i = \sqrt{\lambda_i},$$

where the eigenvalues λ_i , $i = 1, \dots, n$ are:

$$\lambda_1 = 10^{-4}, \quad \lambda_2 = 10^{-2}, \quad \lambda_i = 1 + (i-3) \frac{99}{n-3}, \quad i = 3, \dots, n.$$

(i.e. $v_{\min}^{(1)} = 10^{-4}$, $v_{\min}^{(2)} = 10^{-2}$, $v_{\min}^{(3)} = 1$ and $v_{\max} = 100$).

The results are given in Fig. 1 for single (64 bits) and double (128 bits) precision calculation; we display also in the figure our “exact arithmetic” bound (3.7) (with $p = 2$), where it is for convenience allowed to take real values by removing the first integer truncation.

The main conclusion that one can draw from Fig. 1 is that the bound (3.7), which incidentally appears quite accurate, is valid about as long as

$$\|r_k\|_{A^{-1}} \|r_k - (b - Au_k)\|_{A^{-1}},$$

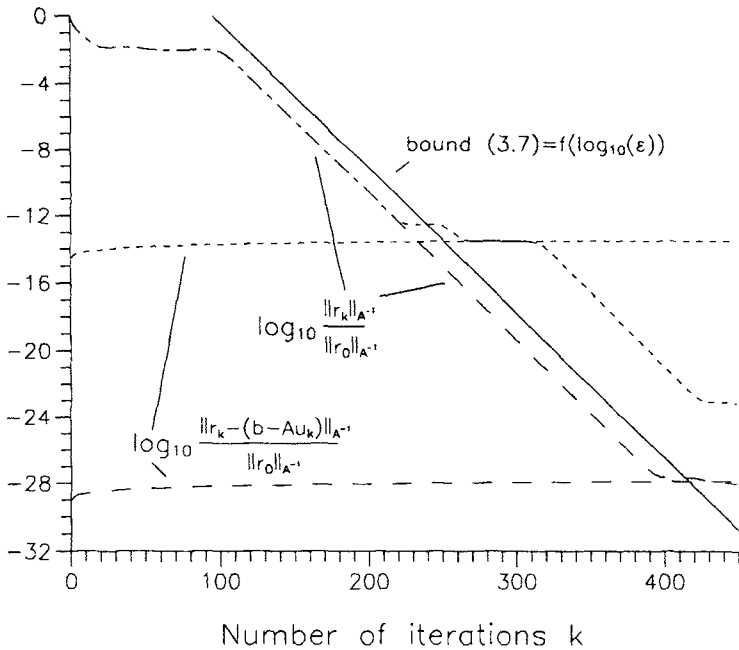


Fig. 1. Numerical results for the test problem of Sect. 3 (with $n = 9900$) run on CDC cyber 855 (OS nos-ve). ---- single precision; — double precision

which means that it has about the same range of validity as Algorithm 2.1, since it makes no sense to pursue the iterations on the updated residual r_k when it becomes smaller than the difference between itself and the true residual $b - Au_k$.¹ Further experiments showed that the choice of b influences only the earliest stages of the convergence process, and that the value of n has almost no influence at all, except when it is relatively small, so that the number of modes between λ_3 and λ_n is not sufficient to avoid other superlinear convergence effects.

To give some further insight, we now corroborate these observations by a theoretical reasoning. Following van der Sluis and van der Vorst [11–13], in presence of small isolated eigenvalues, the conjugate gradient process actually eliminates these eigenvalues by making the first roots of the associate (effectively realised) polynomial equal to these values. This results in a delay in the convergence process, and the ε -independent term in (3.7) may be viewed as an upper bound on this delay. Note also that the time at which this delay occurs depends on the right hand side, which explains the dependence with respect to b observed at the earliest stages of the convergence process. After this, the convergence pursues, and all happens as if the “eliminated” eigenvalues were really removed from the spectrum (for this point, see in particular Theorem 3.1 and its variants in [12]).

Now, what happens in finite precision arithmetic? The iterations which reduce the error corresponding to the uneliminated modes generate, through rounding errors, small perturbation components of the modes which are theoretically

¹ This is also the basis of the stopping criterion used in [14]

eliminated. But it turns out that the process is stable with respect to these perturbations, so that they play almost no role in the later stages, at least as long as the error corresponding to the uneliminated modes remains larger than these perturbations. After this, the process must eliminate these modes again, which causes a second delay. This explains the observed limits of validity of the bound (3.7), and gives us a logical interpretation of the restriction (3.9).

Note that, in this reasoning, we implicitly assume that the number of isolated eigenvalues is small. Otherwise, the elimination of the smallest ones may generate a nonnegligible rounding error component for the modes between them and $v_{\min}^{(p)}$, and one might need several eliminations of the latter. This is attested in restriction (3.9) by the factor multiplying τ , which is an increasing function of the number of considered isolated modes. Now, in such cases, one may use the following generalization of the polynomials (3.1) :

$$(3.10) \quad P_k(v) = \frac{\mathcal{P}_{k-mr}(v_{\min}^{(p)}, v_{\max}, v)}{\mathcal{P}_{k-mr}(v_{\min}^{(p)}, v_{\max}, 0)} \prod_{i=1}^{p-1} \left(\frac{v^{-1} \mathcal{P}_{r_i}(a_i, v_{\max}, v)}{\mathcal{P}'_{r_i}(a_i, v_{\max}, 0)} \left(1 - \frac{v}{v_{\min}^{(i)}} \right) \right)^{m_i}$$

where $m_i, i = 1, \dots, p - 1$ are positive integers. Using still (3.2) and (3.6), one deduces in the same way as (3.7) that, letting $m = \sum_{i=1}^{p-1} m_i$ if $p > 1$ and $m = 0$ otherwise,

$$(3.11) \quad k_\varepsilon \leq \text{int} \left[\frac{1}{2} \sqrt{\frac{v_{\max}}{v_{\min}^{(p)}}} \left(\ln \frac{2}{\varepsilon} + \sum_{i=1}^{p-1} m_i \ln \frac{v_{\min}^{(p)}}{v_{\min}^{(i)}} \right) \right] + m \text{int} \left[\sqrt{\frac{v_{\max}}{v_{\min}^{(p)}}} + 1 \right] + 1$$

is a valid bound on the maximal number of iterations provided that, for all $1 \leq i \leq p - 1$,

$$(3.12) \quad \varepsilon > \tau^{m_i} \left(\frac{(v_{\min}^{(i)})^{i-1}}{\prod_{j=1}^i v_{\min}^{(j)}} \right)^{m_i}.$$

According to the discussion above, it is not surprising to find that (3.11) is obtained from (3.7) by multiplying by m_i the term which represents the cost of one elimination of $v_{\min}^{(i)}$.

4. Large isolated eigenvalues

In exact arithmetic, one cares for a large isolated eigenvalue by letting

$$(4.1) \quad P_k(v) = \frac{\mathcal{P}_{k-1}(v_{\min}, v_{\max}^{(2)}, v)}{\mathcal{P}_{k-1}(v_{\min}, v_{\max}^{(2)}, 0)} \left(1 - \frac{v}{v_{\max}^{(1)}} \right).$$

Assuming k such that $\mathcal{P}_{k-1}(v_{\min}, v_{\max}^{(2)}, 0) \approx 1/\varepsilon$, one then obtains with (1.11) that

$$P_k(v_{\max} + \tau) \approx \frac{\varepsilon \tau}{2v_{\max}^{(1)}} \left(\frac{4 v_{\max}^{(1)}}{v_{\max}^{(2)} - v_{\min}} \right)^{k-1}$$

which may be much greater than ε for sufficiently large k and gap ratio $v_{\max}^{(1)}/v_{\max}^{(2)}$. This is not surprising since it is known from numerical experiments that large isolated eigenvalues may imply strong loss of orthogonality (see e.g. [7, 13]).

According to the discussion of Sect. 2, we can quantify the influence of this loss of orthogonality by exchanging in (4.1) the term $(1 - v/v_{\max})$ for a polynomial

Q_s satisfying $Q_s(0) = 1$, $|Q_s(v)| \leq 1$ for $v \leq v_{\max}^{(2)}$ and such that $|Q_s(v)|$ is sufficiently small for $v \in [v_{\max} - \tau, v_{\max} + \tau]$. These properties are nicely satisfied by the polynomials

$$(4.2) \quad Q_s(v) = \frac{P_s(v_{\max} - \tau, v_{\max} + \tau, v)}{P_s(v_{\max} - \tau, v_{\max} + \tau, 0)}$$

used by Greenbaum [5] in relation with the Lanczos method, which is closely related to our problem since, as observed in [13], the number of conjugate gradient steps needed to deal with large isolated eigenvalues is nothing but about the number of ‘‘copies’’ of these eigenvalues generated by the Lanczos algorithm.

Next allowing for several isolated eigenvalues, we exchange (4.2) for

$$(4.3) \quad Q_s^{(q)}(v) = \prod_{j=1}^{q-1} \frac{\mathcal{P}_{s_j}(v_{\max}^{(j)} - \tau, v_{\max}^{(j)} + \tau, v)}{\mathcal{P}_{s_j}(v_{\max}^{(j)} - \tau, v_{\max}^{(j)} + \tau, 0)},$$

and this leads to the polynomials

$$(4.4) \quad P_k(v) = \frac{\mathcal{P}_{k-mr-s}(v_{\min}^{(p)}, v_{\max}^{(q)}, v)}{\mathcal{P}_{k-mr-s}(v_{\min}^{(p)}, v_{\max}^{(q)}, 0)} \prod_{i=1}^{p-1} \left(\frac{v^{-1} \mathcal{P}_{r_i}(a_i, v_{\max}^{(q)}, v)}{\mathcal{P}'_{r_i}(a_i, v_{\max}^{(q)}, 0)} \left(1 - \frac{v}{v_{\min}^{(i)}} \right) \right)^{m_i} Q_s^{(q)}(v)$$

($p, q, m_i, i = 1, \dots, p - 1$ and $s_j, j = 1, \dots, q - 1$ are positive integers, a_i and $r_i, i = 1, \dots, p - 1$ given respectively by (3.2) and (3.6), $r = \sum_{i=1}^{p-1} r_i$ and $m = \sum_{i=1}^{p-1} m_i$ with $r = m = 0$ if $p = 1, s = \sum_{j=1}^{q-1} s_j$ with $s = 0$ if $q = 1, k - mr - s \geq 0$), where we also take into account the results of the preceding section to cover cases where isolated eigenvalues are present at both extremities of the spectrum.

Note in this respect that by virtue of (1.7) the additional factor $Q_s^{(q)}(v)$, which cares for the large isolated eigenvalues, satisfies $|Q_s^{(q)}(v)| \leq 1$ for all $0 \leq v \leq v_{\max}^{(q)}$, so that the discussion of the preceding section applies to the present context without any change.

We thus obtain for the bound \bar{k}_ε on the maximal number of iterations

$$(4.5) \quad \bar{k}_\varepsilon = \text{int} \left[\frac{1}{2} \sqrt{\frac{v_{\max}^{(q)}}{v_{\min}^{(p)}}} \left(\ln \frac{2}{\varepsilon} + \sum_{i=1}^{p-1} m_i \ln \frac{v_{\min}^{(p)}}{v_{\min}^{(i)}} \right) \right] + m \text{int} \left[\sqrt{\frac{v_{\max}^{(q)}}{v_{\min}^{(p)}}} + 1 \right] + s + 1,$$

which will be reliable in the presence of rounding errors provided that, for $i = 1, \dots, p - 1$, the m_i are sufficiently large to satisfy (3.12) and that, for $j = 1, \dots, q - 1$, the s_j are sufficiently large to ensure $|P_k(v)| < \varepsilon$ for $v \in [v_{\max}^{(j)} - \tau, v_{\max}^{(j)} + \tau]$.

The proof of (4.5) implies in particular (see (3.3))

$$\frac{1}{\mathcal{P}_{\bar{k}_\varepsilon - mr - s}(v_{\min}^{(p)}, v_{\max}^{(q)}, 0)} \prod_{i=1}^{p-1} \left(\frac{1}{|\mathcal{P}'_{r_i}(a_i, v_{\max}^{(q)}, 0)| v_{\min}^{(i)}} \right)^{m_i} < \varepsilon,$$

so that we obtain, with (1.11), (1.12) and (1.13) (using $v_{\max}^{(q)} - a_i > v_{\max}^{(q)} - v_{\min}^{(p)}$)

$$(4.6) \quad \max_{v \in [v_{\max}^{(j)} - \tau, v_{\max}^{(j)} + \tau]} |P_{\bar{k}_\varepsilon}(v)| < \frac{\varepsilon}{2^m} \left(\frac{\tau}{2v_{\max}^{(j)}} \right)^{s_j} \left(\frac{4 v_{\max}^{(j)}}{v_{\max}^{(q)} - v_{\min}^{(p)}} \right)^{\bar{k}_\varepsilon - s} \prod_{i=j+1}^{q-1} \left(\frac{v_{\max}^{(j)}}{v_{\max}^{(i)}} \right)^{s_i}$$

((1.11) and (1.13) require a good separation of the eigenvalues: $v_{\max}^{(q)} \ll v_{\max}^{(q-1)} \ll \dots \ll v_{\max}^{(1)}$, but, see (A.1) they turn to overestimates when this condition is not satisfied).

This allows to bound $s = \sum_{j=1}^{q-1} s_j$ by computing recursively, for $j = q - 1, \dots, 1$,

$$(4.7) \quad s_j = \text{int} \left[\left(\ln \frac{2v_{\max}^{(j)}}{\tau} \right)^{-1} \left(\left(\bar{k}_\varepsilon - s \right) \ln \frac{4 v_{\max}^{(j)}}{v_{\max}^{(q)} - v_{\min}^{(p)}} \right. \right. \\ \left. \left. + \sum_{i=j+1}^{q-1} s_i \ln \frac{v_{\max}^{(j)}}{v_{\max}^{(i)}} + m \ln 2 \right) \right] + 1,$$

where $\bar{k}_\varepsilon - s$ is deduced from (4.5) and τ from Greenbaum theory [5] or from the experiment.

To compare this estimate with the number of iterations actually necessary to deal with large isolated eigenvalues, we use the fact that this number is about the number of "copies" of the concerned eigenvalues generated by the Lanczos algorithm. We may thus deduce it from the experiment by counting the number of Ritz values (i.e. of roots of the (effectively realized) associated polynomial), between $1/2(v_{\max}^{(j+1)} + v_{\max}^{(j)})$ and $1/2(v_{\max}^{(j)} + v_{\max}^{(j-1)})$ for $v_{\max}^{(j)}$, $j = q - 1, \dots, 2$ and greater than $1/2(v_{\max}^{(2)} + v_{\max}^{(1)})$ for $v_{\max}^{(1)}$. In the following, these numbers are denoted \hat{s}_j while $\hat{s} = \sum_{j=1}^{q-1} \hat{s}_j$.

We report in Table 1 the \hat{s}_j resulting from the run of Algorithm 2.1 on the systems $Au = b$ characterized by

$$A = \text{diag}(\lambda_i), \quad b_i = \sqrt{\lambda_i},$$

with $\lambda_1 = v_{\min} = 1$, $\lambda_{n-q+1} = v_{\max}^{(q)} = 100$, and various situations for q and $\lambda_{n-j+1} = v_{\max}^{(j)}$, $j = q - 1, \dots, 1$, the remaining of the spectrum being given by

$$\lambda_i = 1 + (i - 1) \frac{99}{n - q}, \quad i = 1, \dots, n - q + 1.$$

To achieve the comparison, we however need some estimate for τ . Using Greenbaum upper bound would lead to overestimate the perturbation effects while numerical experiment similar to that of [6] would be somewhat cumbersome for large matrices. Therefore, we simply reversed the expression (4.7) and deduced a best fitting value for τ ; more specifically, for three systems as above, with $q = 2$ in each case and respectively $v_{\max}^{(1)} = 10^4, 10^6, 10^{10}$, we computed

$$2 \exp \left(- \frac{\hat{s}_1}{k - \hat{s}_1} \ln \frac{4 v_{\max}^{(1)}}{v_{\max}^{(2)} - v_{\min}} \right)$$

for various k ; we found almost each time a number close to $2 \cdot 10^{-15}$, from which we concluded that the actual interval length was about $2 \cdot 10^{-15} \|A\|$ (note that a sharp estimate is not needed since (4.7) depends only logarithmically on τ). The calculations were made in double precision on CDC 4360 for which the roundoff unit is 2^{-52} , i.e. the same as the one used in the experiment reported by Greenbaum and Strakos [6]. We may therefore compare the value deduced above with that used there, and a remarkable accordance is found since they obtain the best results with $\tau = 5 \cdot 10^{-14} \|A\|$ while no smaller value has been tested.

We are then able to compare the observed values for \hat{s}_j , $j = 1, \dots, q - 1$ against their estimates

$$(4.8) \quad \left(\ln \frac{2v_{\max}^{(j)}}{\tau} \right)^{-1} \left((k - \hat{s}) \ln \frac{4 v_{\max}^{(j)}}{v_{\max}^{(q)} - v_{\min}^{(p)}} + \sum_{i=j+1}^{q-1} \hat{s}_i \ln \frac{v_{\max}^{(j)}}{v_{\max}^{(i)}} \right)$$

where the integer truncation $\text{int}[\cdot] + 1$ of (4.7) (an artifice to get a strict upper bound) has been removed while the observed value $k - \hat{s}$ and \hat{s}_j , $i = j + 1, \dots, q - 1$ have been introduced rather than their estimates (to be discussed below).

The results are reported in Table 1. The five values of k correspond in each case to the stopping criterion $\|r_k\|_{A^{-1}}/\|r_0\|_{A^{-1}} \leq \varepsilon$ with ε successively equal to 10^{-2} , 10^{-4} , 10^{-8} , 10^{-16} and 10^{-32} . The values $k_\varepsilon - s$ deduced from (4.5) are given only for Example 1 because they are necessarily equal for all examples. For $j = 1, \dots, q - 1$, \hat{s}_j is the quantity referred above, "est." its estimate (4.8) and \hat{f}_j the corresponding "frequency"

$$(4.9) \quad \hat{f}_j = \frac{\hat{s}_j}{k - \hat{s}}$$

which our theory predicts to be independent of ε as far as possible for a quotient of integers.

The first fact we would like to point out is that the observed values for $k - \hat{s}$ are identical in each example. This confirms the analysis presenting this number as the number of iterations necessary to deal with the same example from which one has removed the large isolated modes, and thus our interpretation of \hat{s}_j , $j = 1, \dots, q - 1$ as the number of extra iterations needed to eliminate the latter. Further, the numbers of iterations associated with $\nu = 10^4$ are nearly the same in Examples 2, 5 and 7 while those associated with $\nu = 10^6$ are nearly identical in Examples 3 and 6 on the one hand, and in Examples 5 and 7 on the other hand. This leads us to conclude that the number of iterations necessary to deal with some large isolated mode is actually independent of the eigenvalues that are present higher in the spectrum.

Comparing the \hat{s}_j with their estimate (4.8), the latter is found accurate enough for \hat{s}_1 , but systematically an overestimate for the other modes. This explains as follows: while one would like to predict the same number for $\nu = 10^4$ in Examples 5 and 7 as in Example 2, all things are equal in (4.8) except the ratio

$$\frac{\nu_{\max}^{(j)}}{\tau} = \frac{10^4}{2 \cdot 10^{-15} \|A\|}$$

which is in Examples 5 and 7 respectively 100 and 10000 times greater than in Example 2.

This leads us to suggest to use (4.8) with interval lengths proportional to the considered eigenvalue, so that $\nu_{\max}^{(j)}/\tau$ is exchanged for a constant depending only on the machine precision; one then gets for $\nu = 10^4$ the same prediction in all Examples 2, 5 and 7, and so on, so that the estimate is now accurate enough in each case. Of course, this is no proof that the actual interval length is proportional to the considered eigenvalue, only an observation that our semi-empirical model yields better results with this assumption.

It should be mentioned before concluding that an accurate estimation of \hat{s}_j requires an accurate estimate for $k - \hat{s}$, which may be less obvious in more general cases with no regular distribution of the eigenvalues between $\nu_{\min}^{(p)}$ and $\nu_{\max}^{(q)}$. An interesting remark is then that, dividing (4.8) by $k - \hat{s}$, we get an estimate for \hat{f}_j which depends only of the values \hat{f}_i , $i = j + 1, \dots, q - 1$, and the latter may be exchanged without any trouble for their previously computed estimate. In other

Table 1. Numerical results for the test problems of Sect. 4 (with $n = 9900$) run in double precision (-r8 option) on CDC 4360 (OS Unix)

Example 1: $q = 1$

$\frac{\ r_k\ _{A^{-1}}}{\ r_0\ _{A^{-1}}} <$	k	\bar{k}_e
10^{-2}	19	27
10^{-4}	42	50
10^{-8}	88	96
10^{-16}	180	188
10^{-32}	362	372

Example 2: $q = 2, v_{\max}^{(1)} = 10^4$

k	$k - \hat{s}$	\hat{s}_1	est.	\hat{f}_1
23	19	4	3.3	0.211
49	42	7	7.3	0.167
103	88	15	15.3	0.170
212	180	32	31.2	0.178
427	362	65	62.8	0.180

Example 3: $q = 2, v_{\max}^{(1)} = 10^6$

k	$k - \hat{s}$	\hat{s}_1	est.	\hat{f}_1
25	19	6	5.8	0.316
55	42	13	12.9	0.310
114	88	26	27.0	0.295
235	180	55	55.2	0.306
474	362	112	111.1	0.309

Example 4: $q = 2, v_{\max}^{(1)} = 10^{10}$

k	$k - \hat{s}$	\hat{s}_1	est.	\hat{f}_1
30	19	11	10.9	0.579
65	42	23	24.1	0.548
137	88	49	50.5	0.557
284	180	104	103.2	0.578
569	362	207	207.6	0.572

Example 5: $q = 3, v_{\max}^{(2)} = 10^4, v_{\max}^{(1)} = 10^6$

k	$k - \hat{s}$	\hat{s}_2	est.	\hat{f}_2	\hat{s}_1	est.	\hat{f}_1
28	19	3	3.8	0.158	6	6.2	0.316
63	42	7	8.4	0.167	14	13.8	0.333
132	88	15	17.6	0.170	29	29.0	0.330
271	180	31	36.0	0.172	60	59.4	0.333
548	362	63	72.5	0.174	123	119.5	0.340

Example 6: $q = 3, v_{\max}^{(2)} = 10^6, v_{\max}^{(1)} = 10^{10}$

k	$k - \hat{s}$	\hat{s}_2	est.	\hat{f}_2	\hat{s}_1	est.	\hat{f}_1
38	19	6	7.9	0.316	13	12.5	0.684
83	42	13	17.6	0.310	28	27.6	0.667
175	88	27	36.8	0.307	60	57.7	0.682
360	180	57	75.3	0.317	123	118.4	0.683
726	362	115	151.4	0.318	249	238.3	0.688

Example 7: $q = 4, v_{\max}^{(3)} = 10^4, v_{\max}^{(2)} = 10^6, v_{\max}^{(1)} = 10^{10}$

k	$k - \hat{s}$	\hat{s}_3	est.	\hat{f}_3	\hat{s}_2	est.	\hat{f}_2	\hat{s}_1	est.	\hat{f}_1
41	19	3	5.5	0.158	6	8.5	0.316	13	13.7	0.684
92	42	7	12.1	0.167	14	18.8	0.333	29	30.6	0.690
196	88	16	25.4	0.182	29	39.7	0.330	63	64.6	0.716
401	180	32	52.0	0.178	60	81.1	0.333	129	132.0	0.717
813	362	66	104.7	0.182	120	163.4	0.331	265	266.0	0.732

words, taking into account the remark above about the interval lengths, we propose to estimate the “frequencies” \hat{f}_j by the numbers f_j recursively computed according to

$$(4.10) \quad f_j = \left(\ln \frac{2}{\xi} \right)^{-1} \left(\ln \frac{4 v_{\max}^{(j)}}{v_{\max}^{(q)} - v_{\min}^{(p)}} + \sum_{i=j+1}^{q-1} f_i \ln \frac{v_{\max}^{(j)}}{v_{\max}^{(i)}} \right)$$

for $j = q - 1, \dots, 1$.

The result of such a computation with $\xi = 2 \cdot 10^{-15}$ is given in Table 2 for the examples of Table 1. It is found that the values obtained match the observed frequencies within an error less than 5% in nearly all cases.

It should be noted here that this accuracy is unfortunately subject to a good identification of all large isolated modes. Otherwise, the frequencies are underestimate; one gets for instance $f_1 = 0.17$ in Example 5 if one applies (4.10) with $q = 2$ rather than with $q = 3$. This may be troublesome in view of practical application. The quantity

$$(\bar{k}_\epsilon - s) f_j$$

remains however fortunately an upper estimate of δ_j since the decrease of f_j is anyway more than compensated in $\bar{k}_\epsilon - s$ by the overestimation of the effective spectral condition number $v_{\max}^{(q)}/v_{\min}^{(p)}$.

Finally, we also investigated the dependency of ξ upon the roundoff unit ϵ of the machine. For this purpose, we repeated similar experiment in single precision and on Cray Y-MP, and in each case the observed frequencies were in nice agreement with their estimate (4.10) when using

$$(4.11) \quad \xi = 9 \eta$$

which also matches the value used above since η was there 2^{-52} .

5. Isolated eigenvalues at both ends of the spectrum

The theory developed in the preceding section already covers the general case, so that we only pursue here our comparison with the experiment.

To this aim, we repeated (Examples 1’-7’) the experiment of Sect. 4, except that the $n - q + 1$ first eigenvalues are now

$$\lambda_1 = 10^{-4}, \quad \lambda_2 = 10^{-2}, \quad \lambda_i = 1 + (i - 3) \frac{99}{n - q - 2}, \quad i = 3, \dots, n - q + 1.$$

The results are reported in Table 3 (for Example 1’, we give besides \bar{k}_ϵ the value used for $m_1 = m_2$ in (4.5)).

Table 2. Estimate (4.10) of \hat{f}_j for the examples of Table 1

	$q - 1$	f_3	f_2	f_1
Example 2	1			0.174
Example 3	1			0.307
Example 4	1			0.574
Example 5	2		0.174	0.330
Example 6	2		0.307	0.656
Example 7	3	0.174	0.330	0.731

Table 3. Numerical results for the test problems of Sect. 5 (with $n = 9900$) run in double precision (-r8 option) on CDC 4360 (OS Unix)

Example 1': $q = 1$

$\frac{\ r_k\ _{A^{-1}}}{\ r_0\ _{A^{-1}}} <$	k	\bar{k}_ϵ	(m_1, m_2)
10^{-2}	60	118	(1)
10^{-4}	125	141	(1)
10^{-8}	171	187	(1)
10^{-16}	297	370	(2)
10^{-32}	606	646	(3)

Example 2': $q = 2, v_{\max}^{(1)} = 10^4$

k	$k - \hat{s}$	\hat{s}_1	est.	\hat{f}_1
70	60	10	10.4	0.167
147	125	22	21.7	0.176
201	171	30	29.7	0.175
406	344	62	59.7	0.180
716	606	110	105.1	0.182

Example 3': $q = 2, v_{\max}^{(1)} = 10^6$

k	$k - \hat{s}$	\hat{s}_1	est.	\hat{f}_1
79	60	19	18.4	0.317
165	125	40	38.4	0.320
226	171	55	52.5	0.322
454	342	112	104.9	0.327
805	606	199	185.9	0.328

Example 4': $q = 2, v_{\max}^{(1)} = 10^{10}$

k	$k - \hat{s}$	\hat{s}_1	est.	\hat{f}_1
94	60	34	34.4	0.567
195	125	70	71.7	0.560
266	171	95	98.1	0.556
457	292	165	167.5	0.565
955	606	349	347.5	0.576

Example 5': $q = 3, v_{\max}^{(2)} = 10^4, v_{\max}^{(1)} = 10^6$

k	$k - \hat{s}$	\hat{s}_2	est.	\hat{f}_2	\hat{s}_1	est.	\hat{f}_1
90	60	10	12.0	0.167	20	19.7	0.333
189	125	21	25.0	0.168	43	41.2	0.344
260	171	30	34.2	0.175	59	56.5	0.345
524	344	61	68.9	0.177	119	113.7	0.346
926	606	109	121.3	0.180	211	200.5	0.348

Example 6': $q = 3, v_{\max}^{(2)} = 10^6, v_{\max}^{(1)} = 10^{10}$

k	$k - \hat{s}$	\hat{s}_2	est.	\hat{f}_2	\hat{s}_1	est.	\hat{f}_1
117	60	19	25.1	0.317	38	39.5	0.633
244	125	38	52.3	0.304	81	81.8	0.648
275	171	52	71.5	0.304	112	111.9	0.655
683	345	106	144.3	0.307	232	226.1	0.672
1205	606	187	253.5	0.309	412	397.4	0.680

Example 7': $q = 4, v_{\max}^{(3)} = 10^4, v_{\max}^{(2)} = 10^6, v_{\max}^{(1)} = 10^{10}$

k	$k - \hat{s}$	\hat{s}_3	est.	\hat{f}_3	\hat{s}_2	est.	\hat{f}_2	\hat{s}_1	est.	\hat{f}_1
134	60	11	17.3	0.183	20	27.1	0.333	43	44.1	0.717
277	125	22	36.1	0.176	41	56.3	0.328	89	91.4	0.712
379	171	30	49.4	0.175	56	77.0	0.327	122	125.0	0.713
774	344	61	99.5	0.177	114	155.0	0.331	255	252.1	0.741
1373	606	106	175.2	0.175	201	272.8	0.332	460	443.5	0.759

We first observe that the values obtained for $k - \hat{s}$ are here again identical in all examples, which confirms that this part of the number of iterations represents the number of iterations necessary to deal with the same example from which one has removed the large isolated modes, situation for which the analysis of Sect. 3 applies. (One may object the variations of $k - \hat{s}$ observed for $\varepsilon = 10^{-16}$, but the reason is that it is precisely at this time that a second elimination of the small isolated modes becomes necessary, so that very slight perturbations may cause a great variation in the number of iterations.)

Our second remark is that the observed frequencies \hat{f}_j are in all case very similar to those obtained for the corresponding example of Sect. 4 (Table 1), and thus (see Table 2) in agreement with the prediction of our estimate (4.10). Hence, here again we confirm the ability of the developments of the preceding section to deal with the general case.

6. Conclusions

Our results show that, in the presence of isolated eigenvalues at the ends of the spectrum, the number of Conjugate Gradient iterations k_ε necessary to reduce the relative error in the A norm by a factor ε is the sum of three terms

$$k_\varepsilon = k_\varepsilon^{(p,q)} + r_\varepsilon^{(p)} + s_\varepsilon^{(q)}$$

where $p - 1$ and $q - 1$ ($p, q \geq 1$) are respectively the number of small and large isolated eigenvalues and

- $k_\varepsilon^{(p,q)}$ is the number of iterations needed to deal with interior eigenvalues. It is bounded by

$$(6.2) \quad \bar{k}_\varepsilon^{(p,q)} = \text{int} \left[\frac{1}{2} \sqrt{\frac{v_{\max}^{(q)}}{v_{\min}^{(p)}} \ln \frac{2}{\varepsilon}} \right] + 1,$$

and as usual this bound is accurate when there is a large number of modes regularly distributed between $v_{\min}^{(p)}$ and $v_{\max}^{(q)}$.

- $r_\varepsilon^{(p)}$ is the number of iterations needed to eliminate the modes associated with the small isolated eigenvalues. It is bounded by

$$(6.3) \quad \bar{r}_\varepsilon^{(p)} = \text{int} \left[\frac{1}{2} \sqrt{\frac{v_{\max}^{(q)}}{v_{\min}^{(p)}} \left(\sum_{i=1}^{p-1} m_i \ln \frac{v_{\min}^{(p)}}{v_{\min}^{(i)}} \right)} \right] + m \text{int} \left[\sqrt{\frac{v_{\max}^{(q)}}{v_{\min}^{(p)}} + 1} \right] + 1,$$

where $m_i, i = 1, \dots, p - 1$ is the number of needed “eliminations” of the corresponding mode while $m = \sum_{i=1}^{q-1} m_i$. Theoretically, m_i is the smallest integer such that (3.12) holds. In practice, it is observed that only one elimination is necessary provided that p is small and that one does not attempt to make the residual too small with regard to the machine precision.

The bound (6.3) has been found relatively accurate when the small isolated eigenvalues are well separated with a sufficient number of modes between $v_{\min}^{(p)}$ and $v_{\max}^{(q)}$.

- $s_\varepsilon^{(q)}$ represents the number of extra steps necessary to deal with large isolated eigenvalues. Contrarily to both preceding numbers, it is strongly influenced by the rounding errors. A theoretical analysis shows that it is bounded by

$s = \sum_{j=1}^{q-1} s_j$ with s_j computed recursively according to (4.7) (where $\bar{k}_e - s$ is just $\bar{k}_e^{(p,q)} + \bar{s}_e^{(p)}$). In practice, it is observed that

$$(6.4) \quad s_e^{(q)} = \hat{s} = \sum_{j=1}^{q-1} \hat{s}_j$$

where \hat{s}_j is the number of ‘‘copies’’ of the corresponding eigenvalue that are generated by the Lanczos algorithm. Further, the ratios or ‘‘frequencies’’ $\hat{f}_j = \hat{s}_j / (k - \hat{s})$ may be predicted by computing recursively

$$(6.5) \quad f_j = \left(\ln \frac{2}{9\eta} \right)^{-1} \left(\ln \frac{4 v_{\max}^{(j)}}{v_{\max}^{(q)} - v_{\min}^{(p)}} + \sum_{i=j+1}^{q-1} f_i \ln \frac{v_{\max}^{(j)}}{v_{\max}^{(i)}} \right)$$

for $j = q - 1, \dots, 1$, where η is the roundoff unit of the machine. This estimate has been found accurate provided that one has identified all large isolated eigenvalues. Otherwise, the frequencies are underestimated, but

$$(6.6) \quad \bar{s}_e^{(q)} = (\bar{k}_e^{(p,q)} + \bar{s}_e^{(p)}) \left(\sum_{j=1}^{q-1} f_j \right)$$

remains anyway an overestimate of $s_e^{(q)}$. Its derivation is partly empirical, but we believe that it should give a satisfying bound for most practical applications.

Appendix

Proof of (1.10). One easily checks that, under the given assumptions on a, b ,

$$|\mathcal{P}_k(a, b, x)| \leq x |\mathcal{P}'_k(a, b, 0)| \quad \text{for } 0 \leq x \leq b$$

if and only if

$$|T_k(y_0 - y)| \leq y T'_k(y_0) \quad \text{for } 0 \leq y \leq y_0 + 1$$

where $y_0 = \cos \pi / 2k$ and therefore $T'_k(y_0) = k / \sin \pi / 2k$. This relation is obvious in the case $k = 1$. Otherwise, it follows from the fact that $T''_k(y_0) > 0$ (by straightforward calculation); the general properties of the Chebyshev polynomials imply then that $|T_k(y_0 - y)|$ can be greater than $y T'_k(y_0)$ for $y > 0$ only if $y_0 - y < y_1$ where $y_1 = \cos 3\pi / 2k$ is the following root. But this is impossible because

$$(y_0 - y_1) t'_k(y_0) = \frac{k \left(\cos \frac{\pi}{2k} - \cos \frac{3\pi}{2k} \right)}{\sin \frac{\pi}{2k}} = 2k \sin \frac{\pi}{k} > 1$$

while $|T_k(y_0 - y)| \leq 1$ for $0 \leq y \leq y_0 + 1$.

Proof of (1.11). It is straightforward:

$$(A.1) \quad |\mathcal{P}_k(a, b, c)| = \frac{1}{2} \left(\frac{2c}{b-a} \right)^k \left[\left(1 - \frac{a+b}{2c} + \sqrt{1 - \frac{a+b}{2c} + \frac{ab}{c^2}} \right)^k + \left(1 - \frac{a+b}{2c} - \sqrt{1 - \frac{a+b}{2c} + \frac{ab}{c^2}} \right)^k \right] \\ = \frac{1}{2} \left(\frac{4c}{b-a} \right)^k \left[\left(1 - \frac{a+b}{2c} - \frac{(b-a)^2}{16c^2} \right)^k + \left(\frac{(b-a)^2}{16c^2} \right)^k \right] \left(1 + O\left(\frac{b^3}{c^3} \right) \right),$$

whence (1.11).

Proof of (1.12). It is also straightforward:

$$\begin{aligned} |\mathcal{P}_k(b - \delta, b + \delta, 0)| &= \frac{1}{2} \left[\left(\frac{\sqrt{b + \delta} + \sqrt{b - \delta}}{\sqrt{b + \delta} - \sqrt{b - \delta}} \right)^k + \left(\frac{\sqrt{b + \delta} - \sqrt{b - \delta}}{\sqrt{b + \delta} + \sqrt{b - \delta}} \right)^k \right] \\ &= \frac{1}{2} \left[\left(\frac{b + \sqrt{b^2 - \delta^2}}{\delta} \right)^k + \left(\frac{\delta}{b + \sqrt{b^2 - \delta^2}} \right)^k \right] \\ &= \frac{1}{2} \left(\frac{2b}{\delta} \right)^k \left(1 - \frac{\delta^2}{4b^2} \right)^k \left(1 + O\left(\frac{\delta^4}{b^4} \right) \right) + O\left(\frac{\delta^k}{b^k} \right), \end{aligned}$$

whence (1.12).

Acknowledgements. We thank professor R. Beauwens and the referee for useful comments and suggestion.

References

1. Andersson, L. (1976): Ssor preconditioning of toeplitz matrices, Research Report 76.02R, Department of Computer Sciences, Chalmers University of Technology and University of Goeteborg. Goeteborg, Sweden
2. Axelsson, O. (1976): A class of iterative methods for finite element equations. *Comput. Methods Appl. Mech. Eng.* **9**, 123–137
3. Axelsson, O., Barker, V. (1984): *Finite Element Solution of Boundary Value Problems. Theory and Computation*, Academic Press, New York
4. Axelsson, O., Lindskog, G. (1986): On the rate of convergence of the preconditioned conjugate gradient method. *Numer. Math.* **48**, 499–523
5. Greenbaum, A. (1989): Behavior of slightly perturbed Lanczos and conjugate-gradient recurrences. *Linear Algebra Appl.* **113**, 7–63
6. Greenbaum, A., Strakos, Z. (1992): Predicting the behaviour of finite precision Lanczos and Conjugate Gradient computations. *SIAM J. Matrix An. Appl.* **13**, 121–137
7. Hestenes, M., Stiefel, E. (1952): Methods of conjugate gradients for solving linear systems. *J. Res. Natl. Bur. Stand.* **49**, 409–436
8. Jennings, A. (1977): Influence of the eigenvalue spectrum on the convergence rate of the conjugate gradient method. *J. Inst. Maths. Applics.* **20**, 61–72
9. Reid, J. (1971): On the method of conjugate gradients for the solution of large sparse systems of linear equations. In: J. Reid, ed., *Large sparse sets of linear equations*, pp. 231–254 Academic Press, London and New York
10. Strakos, Z. (1991): On the real convergence rate of the Conjugate Gradient method. *Linear Algebra Appl.* **154–156**, 535–549
11. van der Sluis, A. (1992): The convergence behaviour of conjugate gradients and ritz values in various circumstances. In R. Beauwens, P. de Groen, eds., *Iterative methods in linear algebra*, pp. 49–66, North-Holland, Amsterdam London New York Tokyo
12. van der Sluis, A., van der Vorst, H. (1986): The rate of convergence of conjugate gradients. *Numer. Math.* **48**, 543–560
13. van der Vorst, H. (1990): The convergence behaviour of preconditioned CG and CG-S. In: O. Axelsson, L. Kolotilina, eds., *Preconditioned conjugate gradient methods*, Lecture Notes in Mathematics **1457**, pp. 126–136. Springer, Berlin Heidelberg New York
14. Wilkinson, J., Reinsch, C. (1971): *Handbook for automatic computation: linear algebra*. Springer, Berlin Heidelberg New York