Journal of Mathematical Sciences, Vol. 101, No. 4, 2000

THE INFLUENCE OF ISOLATED LARGEST EIGENVALUES ON THE NUMERICAL CONVERGENCE OF THE CG METHOD

A. Yu. Yeremin and I. E. Kaporin

UDC 512.643.5

This paper considers the dependence of the convergence history of the CG method on the largest eigenvalues of a symmetric positive-definite matrix. It is demonstrated that, in solving ill-conditioned linear systems, the reproduction of largest eigenvalues can be so intensive that they cannot be treated as isolated. On the other hand, from the moment the smallest isolated eigenvalues start to govern the numerical convergence of the CG method, the convergence is mainly influenced by the smallest Ritz values. Bibliography: 2 *titles.*

1. Statement of the problem

Consider the application of the CG method to a very large (preconditioned) system

$$Mx = b, \tag{1.1}$$

where the matrix M is symmetric and positive definite (SPD), i.e.,

$$M = M^T > 0.$$

Theoretically, the CG method takes advantage of the occurrence of isolated eigenvalues on both sides of the spectrum of M. Therefore, in exact arithmetic, it needs considerably fewer iterations to converge than predicted by the following standard estimate of the convergence rate in terms of the smallest and largest eigenvalues of M:

$$\frac{\|r_k\|_{M^{-1}}}{\|r_0\|_{M^{-1}}} \le \frac{1}{T_k(\frac{\lambda_n+\lambda_1}{\lambda_n-\lambda_1})}. \tag{1.2}$$

Here, n is the order of the matrix M, and

$$0 < \lambda_1 \le \cdots \le \lambda_n$$

are the nondecreasingly ordered eigenvalues of M.

However, in floating-point arithmetic, the isolated *largest* eigenvalues of M are typically responsible for a considerable increase in the number of CG iterations (as compared to the "exact" counterpart of the CG method). This increase can be so dramatic that the fact that the largest eigenvalues of M are isolated must be considered a drawback of the preconditioning used. The present paper provides some numerical evidence to support this claim.

2. The underlying Lanczos process and Ritz values

A comparison of the exact eigenvalues of M with the sequence of its Ritz values (which are the eigenvalues of the tridiagonal matrix T_k obtained at the kth iteration of the CG method) leads to the following conclusions.

The eigenvalues of M can typically be divided into the following three groups:

(a) isolated smallest eigenvalues;

(b) clustered eigenvalues;

(c) isolated largest eigenvalues.

The negative effect of the presence of isolated largest eigenvalues in the spectrum of M when solving very large problems by the PCG method can be explained as follows.

Translated from *Zapiski Nauchnykh Seminarov POMI*, Vol. 248, 1998, pp. 5–16. Original article submitted April 24, 1998.

At the kth CG iteration, the corresponding Ritz values, i.e., the k eigenvalues of the matrix T_k, can similarly be divided into the following three groups:

(a) the Ritz values that accurately approximate the smallest eigenvalues of M;

(b) the Ritz values that lie within the interval of the clustered eigenvalues of M but approximate no eigenvalue of M sufficiently accurately;

(c) the Ritz values that accurately approximate the isolated largest eigenvalues of M but, possibly, occur with redundant multiplicities.

Usually, the larger an eigenvalue in the group (c), the greater its multiplicity. Also, this multiplicity grows linearly with the number of iterations. Furthermore, this multiplicity also grows as the precision of the computer arithmetic decreases (e.g., when passing from the double-precision version of the routine to the single-precision one).

Finally, the number of largest eigenvalues that can be regarded as "isolated" grows with the iteration number k because the ith eigenvalue must be treated as isolated whenever

$$\frac{|\lambda_i - \lambda_{i+1}|}{\lambda_i} = O(k^{-2}), \quad \frac{|\lambda_i - \lambda_{i-1}|}{\lambda_i} = O(k^{-2}). \tag{2.1}$$

The total number of all of the redundant multiplicities of the largest Ritz values corresponds rather exactly to the number of *extra* CG iterations performed in the presence of floating-point errors (cf. [2]).

Therefore, if the preconditioning used is "unstable" (i.e., M has many isolated largest eigenvalues, which are responsible for the loss of orthogonality in the PCG recurrence relations), then many (50% or more) of the required iterations are actually performed as a consequence of floating-point errors.

3. Main results

In the present paper, we consider some numerical examples that confirm the following claims.

(a) In solving large ill-conditioned problems, the copying of the *largest* eigenvalues can be so intensive that the gaps between them become almost negligible.

(b) As long as the copying of the *smallest* isolated eigenvalues does not affect the convergence of the CG method, the actual convergence rate is properly described by the following estimate, which takes into account only the smallest eigenvalues of the preconditioned matrix (cf. [1, 2]):

$$\frac{\|r_k\|_{M^{-1}}}{\|r_0\|_{M^{-1}}} \le \min_{0 \le j \ll k} \frac{\prod_{i=1}^{j} (\frac{e\lambda_j}{\lambda_i})}{T_{k-j}(\frac{\lambda_n + \lambda_1}{\lambda_n - \lambda_1})}, \quad e = \exp(1). \tag{3.1}$$

(c) As soon as the isolated *smallest* eigenvalues start to reproduce themselves and this reproduction starts to affect convergence, the actual convergence rate is properly described by the estimate

$$\frac{|r_k\|_{M^{-1}}}{|r_0\|_{M^{-1}}} \le \min_{0 \le j \ll k} \frac{\prod_{i=1}^{j} (\frac{e\mu_j}{\mu_i})}{T_{k-j}(\frac{\mu_n + \mu_1}{\mu_n - \mu_1})}, \quad e = \exp(1), \tag{3.2}$$

which involves the smallest Ritz values

$$0 < \mu_1 \le \ldots \le \mu_k$$

of the underlying Lanczos process with account of their multiplicities. In (3.1) and (3.2), the symbol "$\ll$" means that one takes the first local minimum found for $j = 0, 1, \ldots$.

4. The description of test matrices

In our numerical tests, we used a family of matrices parametrized by six parameters that specify three segments containing the spectrum of the matrix and the numbers of eigenvalues belonging to each of them. Within every segment, the eigenvalues are uniformly distributed, i.e.,

$$\lambda_i = \gamma_0 + (\gamma_1 - \gamma_0)\frac{i-1}{m_L}, \qquad i = 1, \ldots, m_L;$$

$$\lambda_i = \gamma_1 + (\gamma_2 - \gamma_1)\frac{i - m_L - 1}{n - m_L - m_R - 1}, \qquad i = m_L + 1, \ldots, n - m_R - 1;$$

$$\lambda_i = \gamma_2 + (\gamma_3 - \gamma_2)\frac{i - n + m_R}{m_R}, \qquad i = n - m_R + 1, \ldots, n,$$

where we always set $\gamma_2 = 1$. Thus,

$$\lambda_1 = \gamma_0, \quad \lambda_n = \gamma_3,$$

whereas m_L and m_R are the numbers of the isolated smallest and largest eigenvalues, respectively. We introduce the diagonal matrix

$$\Lambda = \mathrm{Diag}(\lambda_1, \ldots, \lambda_n)$$

and consider test matrices of the form

$$M = \left(I - 2\frac{vv^T}{v^T v}\right)\Lambda\left(I - 2\frac{vv^T}{v^T v}\right),$$

where, in general, v is an arbitrary nonzero vector. However, in our experiments, we always used the simplest choice

$$v = [1, 1, \ldots, 1]^T.$$

Obviously, the eigenvalues of M are just $\lambda_1, \ldots, \lambda_n$ because, for any v, the transformation matrix is an elementary reflection matrix. The latter matrix can be stored using $O(n)$ memory locations and multiplied by a vector in $O(n)$ operations. Thus, using this family of matrices, one can run nontrivial numerical experiments with the CG method, which require nearly minimum computational resources and, at the same time, involve matrices with completely predetermined distributions of eigenvalues.

In the test problems of the form (1.1) described below, we used the right-hand sides

$$b = Mx_*,$$

where the same exact solution x_* was chosen as follows:

$$x_* = [n, n/2, n/3, \ldots, 1]^T.$$

In the CG algorithm, the zero initial guess was used.

5. How many extra iterations can be required due to the occurrence of isolated largest eigenvalues?

In order to destroy the orthogonality properties of the "exact" CG recurrence relations, it is unnecessary to take, say, $\lambda_n = 10^9$, $\lambda_{n-1} = 10^6$, $\lambda_{n-1} = 10^3$, and $\lambda_i \in (0, 1)$, $i = 1, \ldots, n-3$, as was done in [2] and other papers. In accordance with (2.1), in solving ill-conditioned systems (which may require thousands of iterations to converge) it is desirable that the eigenvalues of the matrix be distributed quite densely to the left of λ_n. Otherwise, after several tens (or hundreds) of iterations, orthogonality will be lost as a consequence of the reproduction of the largest Ritz values.

Consider the following test problems.

Problem 1.	**Problem 2.**	**Problem 3.**
$n = 100000$,	$n = 100000$,	$n = 100000$,
$m_L = 200$,	$m_L = 200$,	$m_L = 200$,
$m_R = 20$,	$m_R = 50$,	$m_R = n - 200$,
$\gamma_0 = 0.001$,	$\gamma_0 = 0.001$,	$\gamma_0 = 0.001$,
$\gamma_1 = 0.5$,	$\gamma_1 = 0.5$,	$\gamma_1 = 1.0$,
$\gamma_2 = 1.0$,	$\gamma_2 = 1.0$,	$\gamma_2 = 1.0$,
$\gamma_3 = 10.0$.	$\gamma_3 = 10.0$.	$\gamma_3 = 10.0$.

The gaps between neighboring largest eigenvalues in Problems 1 and 2 are only 5% and 2% of λ_n, respectively. These problems are compared with Problem 3, which has the same condition number 10^4, but its eigenvalues are distributed very densely to the left of λ_n.

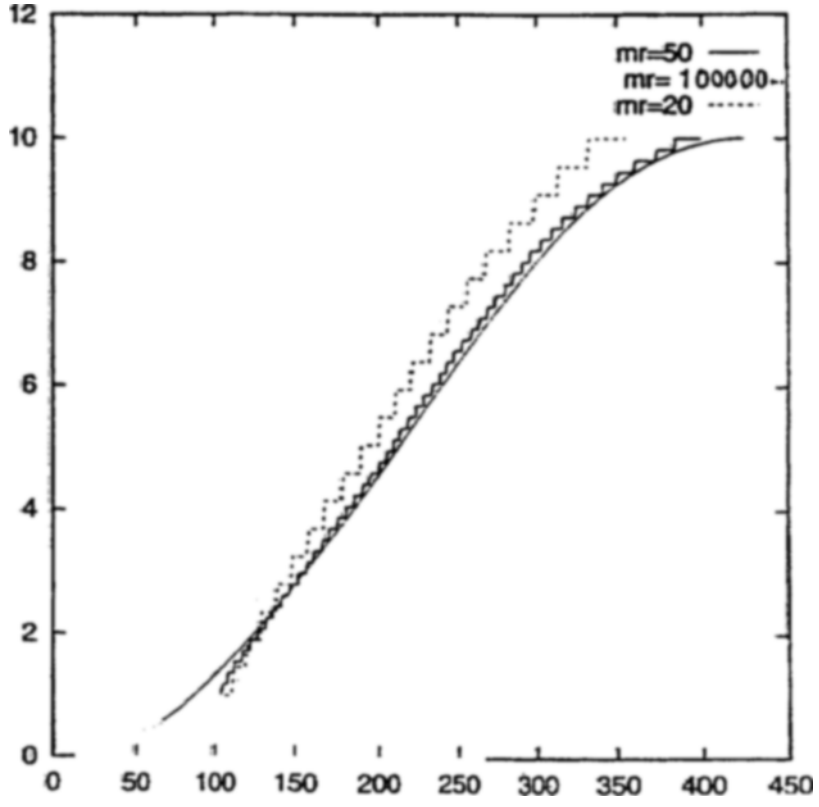

FIG. 1. Magnitudes of Ritz values vs. their numbers for Problems 1-3.

Upon convergence with relative accuracy $\varepsilon = 10^{-7}$ (since we used single-precision floating-point arithmetic, there was no sense in taking smaller ε), we computed all of the eigenvalues of the tridiagonal matrix obtained at the last iteration. From the results of [1] and [2], it follows that the "theoretical" number of iterations can be estimated as the number of (numerically) distinct Ritz values. For this reason, the total number of extra Ritz copies for all of the eigenvalues was accepted as the number of "redundant" iterations, which were caused by floating-point errors.

In Fig. 1, for the three test problems described above, the plots of the magnitudes of the Ritz values μ_i versus their numbers i are presented. For the first two problems, the reproduction of the largest Ritz values obviously results in "staircase" shapes of the curves. Each "footstep" corresponds to an eigenvalue of M, while its length corresponds to the number of its Ritz copies. For Problem 1, convergence was achieved in 355 iterations, 219 of which were redundant. For Problem 2, the corresponding numbers were 401 and 241, respectively. On the other hand, for Problem 3, the CG method converged in 425 iterations, no Ritz value was copied, and the actual number of iterations was close to the theoretically predicted one. The corresponding curve in Fig. 1 is smooth and very similar to a properly scaled and shifted plot of the function $1-\cos(x)$. (The latter function arises when the Ritz values are the roots of a translated Chebyshev polynomial.)

These observations imply the following conclusions. First, in order to improve the standard estimate (1.2) of the convergence rate of the CG method, one should take into account in some nontrivial way (see, e.g., (3.1)) only the smallest isolated eigenvalues but not the largest ones. Second, in considering large ill-conditioned problems, the effect of a good preconditioning must be twofold. On the one hand, the smallest eigenvalues must be "sparsified," and, on the other hand, the largest eigenvalues must be made more "densely" distributed.

6. Numerical testing of some CG-convergence estimates accounting for the isolation of the smallest eigenvalues

As follows from the above discussion, any practical CG-convergence estimate must ignore the fact that

the largest eigenvalues are well separated. Indeed, the "theoretical" gain of taking into account the fact that the largest eigenvalues are isolated almost vanishes as a consequence of the intensive copying of the corresponding Ritz values in the presence of floating-point errors. This conclusion leads us to the estimate (3.1), which stems from the results of [1] and [2]. However, as conjectured in [2] and confirmed by extensive numerical testing, this estimate is only valid until the smallest Ritz values start to be reproduced. Otherwise, only the estimate (3.2) proposed in the present paper describes the actual behavior of the error measure that the CG method attempts to minimize. Furthermore, this estimate involves the quantities $\mu_1, \mu_2, \ldots$, which are readily available at each of the CG iterations (in contrast to the exact eigenvalues, which are unknown until they finally converge). Thus, (3.2) is a realistic upper bound for the quantity $\|r_k\|_{M^{-1}}$, which cannot be measured directly in real-life computations. Nevertheless, the estimate (3.1) is useful in developing practical criteria of preconditioning quality.

In order to demonstrate the relevance of the estimates (1.2), (3.1), and (3.2) to the actual convergence history of the CG iteration, we consider the following test problem.

Problem 4.

$$n = 100000, \quad m_L = 200, \quad m_R = 100, \quad \gamma_0 = 0.001, \quad \gamma_1 = 0.5, \quad \gamma_2 = 1.0, \quad \gamma_3 = 10.0.$$

The right-hand side and initial guess for this problem were chosen as described above. Once again this is a problem whose largest eigenvalues are well separated. However, despite disregarding the potential convergence acceleration due to the isolation of these eigenvalues, both estimates (3.1) (partially) and (3.2) (completely) predict the CG convergence behavior correctly.

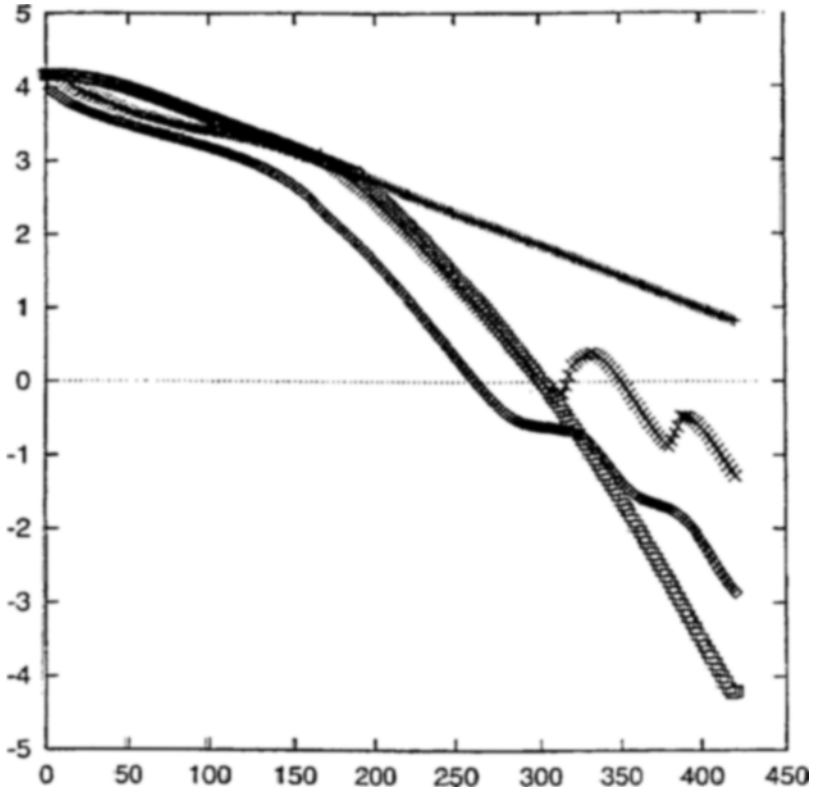

FIG. 2. CG error estimates for Problem 4; $\diamond$: actual error; $+$: estimate in terms of the condition number; $\square$: estimate involving the smallest eigenvalues; $\times$: estimate involving the Ritz values.

In Fig. 2, we present the plots of the decimal logarithm of $\|r_k\|_{M^{-1}}$ and of the right-hand sides of inequalities (1.2), (3.1), and (3.2). Up to the 130th iteration, estimates (1.2) and (3.1) simply coincide and considerably overestimate the actual error. At the same time, the new estimate (3.2) is much more accurate than the previous ones. This can easily be explained by the fact that, at the beginning of the CG iteration, the smallest Ritz values are much larger than the corresponding exact smallest eigenvalues of M.

At iterations between the 130th and 180th, the three estimates give practically the same bounds because the smallest Ritz values nearly coincide with the smallest eigenvalues of M, but the number of iterations is not large enough to reveal the superiority of estimates (3.1) and (3.2).

At iterations between the 180th and 310th, estimates (3.1) and (3.2) still almost coincide (though (3.2) is slightly more accurate) and follow the actual convergence curve quite closely. However, the standard estimate proves to be an overestimate, and its plot is much more flat. In this range, both estimates (3.1) and (3.2) take full advantage of accounting for the isolation of the smallest eigenvalues.

However, starting from the 310–320th iterations, the situation changes. In this stage, the smallest Ritz values start to be reproduced, which slows down the convergence rate considerably. Therefore, the estimate (3.1) becomes too optimistic and actually fails. On the contrary, the estimate (3.2) is still valid and reveals all the details of the convergence history. At this stage, the estimate (3.2) is still much more accurate than the standard estimate (1.2), but the (average) slopes of these curves become more similar to each other. Such behavior is typical for the cases where the required relative accuracy ε approaches the machine precision (which is approximately equal to 10^{-7} in our case).

7. Concluding remarks

The results presented in this paper demonstrate that, in solving very large ill-conditioned systems of linear equations by the CG method, the preconditioners used must result in the largest eigenvalues being distributed sufficiently densely. An example of a good preconditioning strategy satisfying this requirement is provided by the block SSOR (or a similar) method.

Formally, if the matrix M is of the form $M = L^{-1}AL^{-T}$, i.e., it results from preconditioning the original matrix A with a matrix $B = LL^T$ such that

$$B \approx A, \qquad \operatorname{trace}(B^{-1}A) \approx n,$$

then B must be a "stable" approximation of A, i.e., the inequality

$$B^{-1} \leq (1+\gamma)A^{-1}$$

must be satisfied with a sufficiently small $\gamma > 0$.

This work was supported in part by the Netherlands Organization for Scientific Research (NWO) under grant 047 003 017 and by INTAS and RFBR under grant INTAS–RFBR 95-0098.

Translated by A. Yu. Yeremin.

REFERENCES

1. O. Axelsson and G. Lindskog, "On the rate of convergence of the preconditioned conjugate gradient methods," *Numer. Math.*, **48**, 499–523 (1986).
2. Y. Notay, "On the convergence rate of the conjugate gradients in presence of rounding errors," *Numer. Math.*, **65**, 301–317 (1993).