Acta Numerica (2006), pp. 471–542
doi: 10.1017/S096249290626001X

The Lanczos and conjugate gradient algorithms in finite precision arithmetic

Gérard Meurant
CEA/DIF,
BP 12,
91680, Bruyères le Chatel, France
E-mail: gerard.meurant@cea.fr

Zdeněk Strakoš*
Institute of Computer Science,
Academy of Sciences of the Czech Republic,
Pod Vodárenskou věži 2,
182 07 Praha 8, Czech Republic
E-mail: strakos@cs.cas.cz

Dedicated to Chris Paige for his fundamental contributions to the rounding error analysis of the Lanczos algorithm

The Lanczos and conjugate gradient algorithms were introduced more than five decades ago as tools for numerical computation of dominant eigenvalues of symmetric matrices and for solving linear algebraic systems with symmetric positive definite matrices, respectively. Because of their fundamental relationship with the theory of orthogonal polynomials and Gauss quadrature of the Riemann–Stieltjes integral, the Lanczos and conjugate gradient algorithms represent very interesting general mathematical objects, with highly nonlinear properties which can be conveniently translated from algebraic language into the language of mathematical analysis, and *vice versa*. The algorithms are also very interesting numerically, since their numerical behaviour can be explained by an elegant mathematical theory, and the interplay between analysis and algebra is useful there too.

Motivated by this view, the present contribution wishes to pay a tribute to those who have made an understanding of the Lanczos and conjugate gradient algorithms possible through their pioneering work, and to review recent solutions of several open problems that have also contributed to knowledge of the subject.

* Supported by the National Program of Research ‘Information Society’ under project 1ET400300415 and by the Institutional Research Plan AVOZ10300504.

CONTENTS

1. Introduction

The Lanczos algorithm is one of the most frequently used tools for computing a few dominant eigenvalues (and eventually eigenvectors) of a large sparse symmetric n by n matrix A. More specifically, if for instance the extreme eigenvalues of A are well separated, the Lanczos algorithm obtains good approximations to these eigenvalues in only a few iterations. Moreover, the matrix A need not be explicitly available. The Lanczos algorithm only needs a procedure performing the matrix-vector product Av for a given vector v. Hence, it can even be used in some applications for which the matrix cannot be stored as long as one is able to produce the result of the operation matrix times a given vector. Another interesting property is that when one just needs the eigenvalues, the Lanczos algorithm only requires a very small storage of a few vectors (besides storing the matrix where applicable), since a new basis vector is computed using only the two previous ones.

The Lanczos algorithm constructs a basis of Krylov subspaces which are defined for a square matrix A of order n and a vector v by

$$\mathcal{K}_k(v, A) = \operatorname{span}\{v, Av, \dots, A^{k-1}v\}, \quad k = 1, 2, \dots .$$

Since the natural basis $v, Av, \dots, A^{k-1}v$ is badly conditioned, the algorithm constructs an orthonormal basis of $\mathcal{K}_k(v, A)$. The vectors in the natural basis can even become numerically dependent (within the accuracy of the floating point calculations) for a small value of k. In fact, computing successively $A^k v$ for a given vector v is, with a proper normalization, the basis of the power method. Unlike the power method, which focuses at the kth step only on the local information present in $A^{k-1}v$, and aims to converge to the eigenvector corresponding to the eigenvalue of largest modulus, the Lanczos algorithm exploits simultaneously all vector information accumulated in previous steps. Building an orthonormal basis of $\mathcal{K}_k(v, A)$ can therefore be seen as an effective numerical tool for extracting information from the sequence $v, Av, \dots, A^{k-1}v$ while preventing any possible loss which could be caused by effects of existing dominance.

The orthonormal basis vectors v^j, $j = 1, \ldots, k$ are constructed recursively one at a time and can be considered columns of a matrix $V_k = (v^1, \ldots, v^k)$. The method also constructs at iteration k an unreduced symmetric tridiagonal k by k matrix T_k (which is obtained from T_{k-1} by adding one row and one column) having positive subdiagonal entries, whose eigenvalues are approximations to the eigenvalues of A: see, for instance, Lanczos (1950), Wilkinson (1965) and Parlett (1980). Moreover, in exact arithmetic $AV_m = V_m T_m$ for some $m \leq n$, n being the dimension of the problem. It means that the columns of V_m span an invariant subspace of the operator represented by A, and the eigenvalues of T_m are also eigenvalues of A.

All these properties are quite nice. However, it has been known since the introduction of the method by Cornelius Lanczos (1950) that, when used in finite precision arithmetic, this algorithm does not fulfil its theoretical properties. In particular, the computed basis vectors lose their orthogonality as the iteration number k increases. Moreover, as a consequence of the loss of orthogonality, in finite precision computations multiple approximations of the original eigenvalues appear within the set of computed approximate eigenvalues if we do a sufficiently large number of iterations. This phenomenon leads to a delay in the computation of some other eigenvalues. Sometimes it is also difficult to determine whether some computed approximations are additional copies caused by rounding error effects and the loss of orthogonality, or genuine close eigenvalues.

The finite precision behaviour of the Lanczos algorithm was analysed in great depth by Chris Paige in his pioneering PhD thesis, Paige (1971); see also Paige (1972, 1976, 1980). With no exaggeration, Paige's work was revolutionary. He showed that the effects of rounding errors in the Lanczos algorithm can be described by a rigorous and elegant mathematical theory. In the spirit of Wilkinson, the theory built by Paige reveals the mechanics of the finite precision Lanczos algorithm behaviour. It starts with bounds on the elementary round-off errors at each iteration, and ends up with elegant mathematical theorems which link convergence of the computed eigenvalue approximations to the loss of orthogonality. Following Paige, the theory was further developed and applied by Parlett and Scott (1979), Scott (1979), Parlett (1980) and Simon (1982, 1984*a*, 1984*b*). A forward error analysis was attempted by Grcar (1981).

Another fundamental step forward, similar in significance to that of Paige, was made by Anne Greenbaum (1989). On the foundations laid by Paige she developed a backward-like analysis of the Lanczos algorithm (and also of the closely related conjugate gradient algorithm). Her ideas, combined with thoughts of several other authors, stimulated further developments: see, *e.g.*, Druskin and Knizhnerman (1991), Strakoš (1991), Greenbaum and Strakoš (1992), Strakoš and Greenbaum (1992), Knizhnerman (1995*a*) and Druskin, Greenbaum and Knizhnerman (1998). Recently, new analysis of the open problems formulated in the literature has led to work by Zemke (2003), Wülling (2005, 2006) and Meurant (2006).

The Lanczos algorithm has been implemented and applied in two ways. In the first way it is applied without any additional measures designed to limit round-off. The number of iterations is not limited by strict rules and the 'good' eigenvalue approximations are identified by some convergence tests. In particular, this was advocated by Cullum and Willoughby (1985). The second way limits the unwanted effects of rounding errors by some form of reorthogonalization, of varied sophistication. Proposals in this direction were made by Parlett and Scott (1979), Grcar (1981), Simon (1982) and Parlett (1992). Here the theory developed by Paige almost immediately led to successful software implementations. The way the Lanczos algorithm is used in a particular application depends on a particular goal.

After a period of intensive discussions, particularly concentrated at the Institute of Numerical Analysis at UCLA (see Hestenes and Todd (1991) and Golub and O'Leary (1989)), the conjugate gradient (CG) algorithm, independently introduced by Magnus Hestenes and Eduard Stiefel, was thoroughly described in their seminal paper, Hestenes and Stiefel (1952). Intended for solving symmetric positive definite linear systems, it is closely linked to the Lanczos algorithm. Lanczos used his algorithm to solve linear systems in Lanczos (1952) but it was already clear in Lanczos (1950) that it can be used for that purpose. In fact, even though it was not introduced in this way, one can obtain the Hestenes–Stiefel CG from the Lanczos algorithm by doing an LU factorization (with L lower triangular and U upper triangular) of the positive definite matrix T_k given by the Lanczos coefficients (by introducing some intermediate variables). In exact arithmetic the CG residual vectors are proportional to the Lanczos vectors. In finite precision, the residual vectors lose their orthogonality just as the Lanczos vectors do.

The Lanczos algorithm, respectively CG, builds up (in exact arithmetic) orthogonal bases of Krylov subspaces $\mathcal{K}_k(v, A)$, $k = 1, 2, \ldots.$ and the basis vectors can be expressed in terms of polynomials in the matrix A applied to the initial vector v. Using the spectral decomposition of the symmetric (respectively the symmetric positive definite) matrix A, it is easy to see that the corresponding polynomials are orthogonal with respect to a Riemann–Stieltjes integral. Its piecewise constant distribution function is defined by the points of increase equal to the eigenvalues of A and by the sizes of the discontinuities equal to the squared components of v in the corresponding invariant subspaces. In this way, the Lanczos algorithm and CG are intimately related to orthogonal polynomials: see Hestenes and Stiefel (1952) and Fischer (1996). This fact has been emphasized for decades in the work of Gene Golub, who substantially contributed to the whole field by his deep understanding of the interconnections between different mathematical areas and by sharing his ideas with many collaborators: see, *e.g.*, Gautschi (2002). The Lanczos algorithm can be viewed as a matrix formulation of the discretized Stieltjes procedure (see, *e.g.*, Gautschi (1982)), and its roots can therefore be linked to the works of Stieltjes (1884), Christoffel (1877) and Darboux (1878). Such interconnections are fundamental to the understanding of the Lanczos algorithm and CG behaviour in both exact and finite precision arithmetic. In particular, in exact arithmetic the A-norm of the CG error can be written using the Gauss quadrature formula, and this clearly shows that the convergence rate depends, in a rather complicated way, on how well the eigenvalues of A are approximated by the eigenvalues of T_k. This also indicates possible differences in the effect of eigenvalues from different parts of the spectrum of A on the convergence behaviour. In finite precision arithmetic the Gauss quadrature formula is also verified, up to small terms involving the machine precision. However, the appearance of multiple approximations of the original eigenvalues leads to a delay in CG convergence.

The concept of delay is essential to analysis of the CG finite precision behaviour. In short, delay of convergence in a CG finite precision computation is determined by the rank-deficiencies of the computed Krylov subspaces. This understanding emerged from the work of Greenbaum (Greenbaum 1989, Greenbaum and Strakoš 1992) and Notay (1993), and it was strongly advocated by Paige and Strakoš (1999). Analysis and discussion of the Gauss quadrature relationship in finite precision arithmetic can be found in Golub and Strakoš (1994), Strakoš and Tichý (2002) and Meurant (2006).

A finite precision computation does not give the approximate solution with an arbitrarily small error. The error is not reduced below some level, called the maximal attainable accuracy. This is not so important for the Lanczos algorithm, as Paige (1971) shows, but it can become important in solving highly ill-conditioned linear systems and, in particular, in some inner iterations within nonlinear optimization algorithms. Maximal attainable accuracy of CG has been studied for a long time. The early results (see, *e.g.*, Wozniakowski (1978, 1980) and Bollen (1984), with a thorough survey given in Chapter 17 of Higham (2002)) were, however, not applicable to practical implementations. These were analysed more recently by Greenbaum (1997*a*, 1994), Sleijpen, van der Vorst and Fokkema (1994), Sleijpen, van der Vorst and Modersitzki (2001), Björck, Elfving and Strakoš (1998) and Gutknecht and Strakoš (2000). It turns out that a deterioration of the maximal attainable accuracy can be caused at a very early stage of the computation and that CG is unable to correct such a situation in later iterations.

The authors have previously published some surveys of the Lanczos and CG algorithms in exact and finite precision arithmetic as parts of more widely based publications: see Meurant (1999*b*), Strakoš (1998) and Strakoš and Liesen (2005). Following these works, this paper first recalls in Sections 2 and 3 the basic facts on the Lanczos and CG algorithms in exact arithmetic. Then we turn to our main goal – to review the main results on the behaviour of the Lanczos and CG algorithms in finite precision arithmetic, and to present some recent developments related in particular to the appearance of multiple computed approximations of simple original eigenvalues. Section 4 is devoted to the Lanczos algorithm and Section 5 to CG.

For simplicity of exposition we adopt in this paper several restrictions. We will consider real symmetric resp. symmetric positive definite problems. Restriction to real problems is not substantial; we use it for convenience of notation. We will not consider nonsymmetric problems, since this extension would necessarily bring into consideration fundamental issues not present in the symmetric case, some of them still not fully understood. This would, in our opinion, distract from the focus of this paper. In particular, for CG we will assume that the symmetric positive definite matrix A is not close to being singular. Solving near-singular problems (as in singular problems) needs specific approaches. Their presentation and the analysis of their behaviour in finite precision arithmetic is beyond the scope of this paper. We will consider problems with single right-hand sides only. In particular, we will not include the block Lanczos algorithm since that would require significant additional space. Although we understand that preconditioning represents an unavoidable and fundamental part of practical computations, we concentrate here on analysis of basic unpreconditioned algorithms. Most of the results can be extended to preconditioned algorithms: see Strakoš and Tichý (2005) and Meurant (2006).

Unless we need to relate the exact arithmetic quantities to the corresponding results of finite precision computations, we do not use any specific notation for the latter; the meaning will be clear from the context. When helpful, we will emphasize the distinction by using the word ‘ideally’ to refer to a result using exact arithmetic, and ‘computationally’ or ‘numerically’ to refer to a result of a finite precision computation.

2. The Lanczos algorithm

This section briefly describes the Lanczos algorithm in exact arithmetic and presents bounds for the convergence of the eigenvalue approximations. For an extensive and thorough description we refer to Parlett (1980).

Strictly speaking, we should not use the term ‘convergence’ since (with a proper initial vector) the algorithm ideally finds all distinct eigenvalues of A in less than (or equal to) n iterations. Similarly, the term ‘convergence of CG’ used throughout the paper must be understood differently from the classical asymptotic approach: see, *e.g.*, Hackbusch (1994, p. 270), Beckermann and Kuijlaars (2002) and Kuijlaars (2006). Here we must analyse the behaviour from the start since there is no transient phase which can be skipped, just as there is no asymptotic phase which eventually describes convergence.

2.1. Basic properties of the Lanczos algorithm

Let A be a real n by n nonsingular symmetric matrix and v be a given n-dimensional vector of Euclidean norm 1. The kth Krylov subspace is defined by

$$\mathcal{K}_k(v, A) = \mathrm{span}\{v, Av, \dots, A^{k-1}v\}.$$

Ideally, as long as k is less than or equal to the order of the minimal polynomial of v with respect to A (see Chapter VII, §1 and §2 in Gantmacher (1959)), the subspace $\mathcal{K}_k(v, A)$ is of dimension k and the vectors $A^j v, j = 0, \dots, k-1$ are linearly independent. Clearly, for any v the degree of the minimal polynomial of v with respect to A is always less than or equal to the degree of the minimal polynomial of A; there always exists a vector v such that the latter is reached.

Our goal is to construct an orthonormal basis of the Krylov subspace. Although historically things did not proceed in this way, let us consider what is now called the Arnoldi algorithm (Arnoldi 1951). This is a variant of the Gram–Schmidt orthogonalization process applied to the Krylov basis without assuming A to be symmetric. Starting from $v^1 = v$, the algorithm for computing the $(j+1)$st vector of the basis using the previous ones is

$$\begin{aligned} h_{i,j} &= (Av^j, v^i), \quad i = 1, \dots, j, \\ \hat{v}^j &= Av^j - \sum_{i=1}^{j} h_{i,j} v^i, \\ h_{j+1,j} &= \|\hat{v}^j\|, \quad \text{if } \ h_{j+1,j} = 0 \ \text{ then stop}, \\ v^{j+1} &= \frac{\hat{v}^j}{h_{j+1,j}}. \end{aligned}$$

It is easy to verify that the vectors v^j span the Krylov subspace and that they are orthonormal. Collecting the basis vectors up to iteration k in an n by k matrix V_k, the relations defining the vector v^{k+1} can be written in a matrix form as

$$AV_k = V_k H_k + h_{k+1,k} v^{k+1} (e^k)^T,$$

where H_k is an unreduced upper Hessenberg matrix with elements $h_{i,j}$, which means that its elements are nonzero in the upper triangle and on the first subdiagonal, and zero below this. The vector e^k is the kth column of the k by k identity matrix (throughout this paper, e^j denotes the jth column of an identity matrix of the size determined by the context). From the orthogonality of the basis vectors,

$$V_k^T A V_k = H_k.$$

If we suppose that the matrix A is symmetric, then because of the last relation, H_k is also symmetric, and therefore tridiagonal. Consequently $\hat{v}^k$ and hence v^{k+1} can be computed using only the two previous vectors v^k and v^{k-1}, and this gives the elegant Lanczos algorithm. Starting from a vector $v^1 = v$, $\|v\| = 1$, $v^0 = 0$, $\eta_1 = 0$, the iterations are:

for $k = 1, 2, \ldots$

$$\begin{aligned} \alpha_k &= (Av^k, v^k) = (v^k)^T A v^k, \\ \hat{v}^{k+1} &= Av^k - \alpha_k v^k - \eta_k v^{k-1}, \\ \eta_{k+1} &= \|\hat{v}^{k+1}\|, \quad \text{if } \ \eta_{k+1} = 0 \ \text{ then stop}, \\ v^{k+1} &= \frac{\hat{v}^{k+1}}{\eta_{k+1}}. \end{aligned}$$

We point out that the orthogonalization of the newly computed Av^k against the previously computed vectors in the Arnoldi algorithm and in the Lanczos algorithm described above corresponds to the classical version of the Gram–Schmidt orthogonalization. Here the individual orthogonalization coefficients are computed independently of each other. If a mathematically equivalent modified Gram–Schmidt orthogonalization is used, then the orthogonalization coefficients are computed and the orthogonalization is performed recursively, which in the case of the Lanczos algorithm gives the following implementation. Starting from $v^1 = v$, $\|v\| = 1$, $v^0 = 0$, $\eta_1 = 0$:

for $k = 1, 2, \ldots$

$$\begin{aligned} u^k &= Av^k - \eta_k v^{k-1}, \\ \alpha_k &= (u^k, v^k), \\ \hat{v}^{k+1} &= u^k - \alpha_k v^k, \\ \eta_{k+1} &= \|\hat{v}^{k+1}\|, \quad \text{if } \ \eta_{k+1} = 0 \ \text{ then stop}, \\ v^{k+1} &= \frac{\hat{v}^{k+1}}{\eta_{k+1}}. \end{aligned} \tag{2.1}$$

Clearly, this version can be implemented by storing two vectors instead of three. Although mathematically equivalent to the previous version, the last one advocated by Paige (1976, 1980) and Lewis (1977) can, because of the relationship between classical and modified Gram–Schmidt orthogonalization, be expected to be slightly numerically superior.

In matrix notation the Lanczos algorithm can be expressed as follows:

$$AV_k = V_k T_k + \eta_{k+1} v^{k+1} (e^k)^T,$$

where

$$T_k = \begin{pmatrix} \alpha_1 & \eta_2 & & & \\ \eta_2 & \alpha_2 & \eta_3 & & \\ & \ddots & \ddots & \ddots & \\ & & \eta_{k-1} & \alpha_{k-1} & \eta_k \\ & & & \eta_k & \alpha_k \end{pmatrix}$$

is an unreduced symmetric tridiagonal matrix with positive subdiagonal entries storing coefficients of the Lanczos recurrence.

We note that since $\|v^k\| = 1$, α_k is a so-called Rayleigh quotient. This implies that

$$\lambda_{\min}(A) \le \alpha_k \le \lambda_{\max}(A).$$

We denote the eigenvalues of A (which are real) by

$$\lambda_{\min}(A) = \lambda_1 \le \lambda_2 \le \cdots \le \lambda_n = \lambda_{\max}(A),$$

and the corresponding orthonormal eigenvectors $q^1, \ldots, q^n$, $Q \equiv (q^1, \ldots, q^n)$.

If $\eta_j \neq 0$ for $j = 2, \ldots, n$, then $AV_n = V_n T_n$ ($\hat{v}^{n+1}$ must be orthogonal to a set of n orthonormal vectors in a space of dimension n and must therefore vanish). Otherwise there exists an $m+1 < n$ for which $\eta_{m+1} = 0$, $AV_m = V_m T_m$, and we have found an invariant subspace of A, the eigenvalues of T_m being a subset of the eigenvalues of A. When the Lanczos algorithm does not stop before $m = n$, the eigenvalues of A are simple since A is similar to the unreduced symmetric tridiagonal matrix T_n. On the other hand, if A has some multiple eigenvalues, then $\eta_{m+1} = 0$ for some $m+1 < n$. Ideally, the Lanczos algorithm cannot detect the multiplicity of the individual eigenvalues. In exact arithmetic an eigenvalue of A is found as an eigenvalue of T_m only once.

Let

$$\theta_1^{(k)} < \theta_2^{(k)} < \cdots < \theta_k^{(k)}$$

be the eigenvalues of T_k with the corresponding normalized eigenvectors $z_{(k)}^j \equiv (\zeta_{1,j}^{(k)}, \ldots, \zeta_{k,j}^{(k)})^T$, $j = 1, \ldots, k$, $Z_k \equiv (z_{(k)}^1, \ldots, z_{(k)}^k)$. Since the Lanczos algorithm can be considered as a Rayleigh–Ritz procedure, the eigenvalues $\theta_j^{(k)}$ are called Ritz values and the associated vectors $x_{(k)}^j = V_k z_{(k)}^j$ are known as the Ritz vectors. They are the approximations to the eigenvectors of A given by the algorithm. The residual associated with an eigenpair $(\theta_j^{(k)}, x_{(k)}^j)$

obtained from T_k is

$$\begin{aligned} r^j_{(k)} &= Ax^j_{(k)} - \theta^{(k)}_j x^j_{(k)} = (AV_k - V_kT_k)z^j_{(k)} \\ &= \eta_{k+1}(e^k)^T z^j_{(k)}\, v^{k+1} \\ &= \eta_{k+1}\zeta^{(k)}_{k,j}\, v^{k+1}. \end{aligned}$$

Therefore

$$\|r^j_{(k)}\| = \eta_{k+1}|\zeta^{(k)}_{k,j}|.$$

We see that for a given k all residual vectors are proportional to v^{k+1}. When the product of the coefficient η_{k+1} with the absolute value of the bottom element of $z^j_{(k)}$ is small, we have a small residual norm. Moreover, using the spectral decomposition of A and the fact that (in exact arithmetic!) $\|x^j_{(k)}\| = 1$,

$$\min_i |\lambda_i - \theta^{(k)}_j| \le \|r^j_{(k)}\| = \eta_{k+1}|\zeta^{(k)}_{k,j}|.$$

Consequently a small residual norm $\|r^j_{(k)}\|$ means convergence of $\theta^{(k)}_j$ to some eigenvalue of A.

2.2. Relationship to orthogonal polynomials

By using the three-term recurrence, the Lanczos basis vectors v^2, $v^3, \ldots$ can be expressed in terms of polynomials in the matrix A acting on the initial vector v^1. From (2.1) we see that

$$v^{k+1} = p_{k+1}(A)v^1, \quad k = 0, 1, \ldots, \tag{2.2}$$

where the polynomials p_k satisfy the three-term recurrence (with $p_0 \equiv 0$)

$$p_1(\lambda) = 1; \quad \eta_{k+1}p_{k+1}(\lambda) = (\lambda - \alpha_k)p_k(\lambda) - \eta_k p_{k-1}(\lambda), \quad k = 1, 2, \ldots. \tag{2.3}$$

Let $\chi_{1,k}(\lambda)$ (or, where appropriate, simply $\chi_k(\lambda)$) be the characteristic polynomial of T_k (determinant of $T_k - \lambda I$), so that $\chi_0(\lambda) = 1$, $\chi_1(\lambda) = (\alpha_1 - \lambda)$, $\chi_k(\lambda) = (\alpha_k - \lambda)\chi_{k-1}(\lambda) - \eta_k^2\chi_{k-2}(\lambda)$; then for the degree k polynomial

$$p_{k+1}(\lambda) = (-1)^k \frac{\chi_{1,k}(\lambda)}{\eta_2 \cdots \eta_{k+1}}.$$

Using the orthogonality of the vectors $v^1, v^2, \ldots$ and the spectral decomposition of A, the normalized Lanczos polynomials $p_1(\lambda) = 1$, $p_2(\lambda)$, $p_3(\lambda), \ldots$ are orthonormal polynomials with respect to a scalar product defined by the Riemann–Stieltjes integral

$$(p, q) = \int_{\lambda_1}^{\lambda_n} p(\lambda)q(\lambda)\,\mathrm{d}\omega(\lambda) = \sum_{l=1}^{n} \omega_l\, p(\lambda_l)q(\lambda_l), \tag{2.4}$$

where the distribution function ω is a non-decreasing piecewise constant function with at most n points of increase $\lambda_1, \dots, \lambda_n$. For simplicity of exposition, suppose that

$$\lambda_1 < \lambda_2 < \cdots < \lambda_n,$$

i.e., all eigenvalues of A are distinct. Then

$$\omega(\lambda) = \begin{cases} 0 & \text{if } \lambda < \lambda_1, \\ \sum_{l=1}^{i} \omega_l & \text{if } \lambda_i \leq \lambda < \lambda_{i+1}, \\ \sum_{l=1}^{n} \omega_l = 1 & \text{if } \lambda_n \leq \lambda, \end{cases}$$

where $\omega_l = |(v^1, q^l)|^2$ is the squared component of the starting vector v^1 in the direction of the lth invariant subspace of A.

Writing $P_k(\lambda) = (p_1(\lambda), \dots, p_k(\lambda))^T$, the recurrence for the orthonormal polynomials can be written in the matrix form

$$\lambda P_k(\lambda) = T_k P_k(\lambda) + \eta_{k+1} p_{k+1}(\lambda) e^k.$$

Since p_{k+1} is proportional to the characteristic polynomial of T_k, its roots are the eigenvalues of T_k, that is, the Ritz values $\theta_j^{(k)}$, $j = 1, \dots, k$.

Since $\chi_{1,k}(\lambda)$ is (apart from multiplication by $(-1)^k$) a monic polynomial orthogonal with respect to the inner product defined by (2.4) to any polynomial of degree $k-1$ or less, it must resolve the following minimization problem:

$$(-1)^k \chi_{1,k}(\lambda) = \arg \min_{\psi \in \mathcal{M}_k} \int_{\lambda_1}^{\lambda_n} \psi^2(\lambda) \, \mathrm{d}\omega(\lambda), \quad k = 1, 2, \dots, n,$$

where $\mathcal{M}_k$ denotes the set of all monic polynomials of degree less than or equal to k.

Consider the unreduced symmetric tridiagonal matrix T_k defined above. It stores the coefficients of the first k steps of the Lanczos algorithm applied to A with an initial vector v^1. The same T_k can be seen as a result of the Lanczos algorithm applied to T_k with the (k-dimensional) initial vector e^1. Consequently the polynomials $p_1 = 1$, $p_2, \dots, p_{k+1}$ form a set of orthonormal polynomials with respect to a scalar product defined by the Riemann–Stieltjes integral

$$(p, q)_k = \int_{\lambda_1}^{\lambda_n} p(\lambda) q(\lambda) \, \mathrm{d}\omega^{(k)}(\lambda) = \sum_{l=1}^{k} \omega_l^{(k)} p(\theta_l^{(k)}) q(\theta_l^{(k)}), \tag{2.5}$$

where the distribution function $\omega^{(k)}$ is a non-decreasing piecewise constant function with k points of increase $\theta_1^{(k)}, \dots, \theta_k^{(k)}$,

$$\omega^{(k)}(\lambda) = \begin{cases} 0 & \text{if } \lambda < \theta_1^{(k)}, \\ \sum_{l=1}^{i} \omega_l^{(k)} & \text{if } \theta_i^{(k)} \le \lambda < \theta_{i+1}^{(k)}, \\ \sum_{l=1}^{k} \omega_l^{(k)} = 1 & \text{if } \theta_k^{(k)} \le \lambda, \end{cases}$$

and $\omega_l^{(k)} = |(z_{(k)}^l, e^1)|^2$. We see that the first components of the normalized eigenvectors of T_k determine the weights in the Riemann–Stieltjes integral (2.5). Here p_{k+1} represents the $(k+1)$st orthogonal polynomial in the sequence defined by (2.4), and, at the same time, the final polynomial with roots $\theta_1^{(k)}, \dots, \theta_k^{(k)}$ in the same sequence of orthonormal polynomials defined by (2.5). This fact has the following fundamental consequence, formulated as a theorem.

Theorem 2.1. Using the previous notation, (2.5) represents the kth Gauss quadrature approximation to the Riemann–Stieltjes integral (2.4).

Proof. Consider a polynomial $\Phi(\lambda)$ of degree at most $2k-1$. Then we can write

$$\Phi(\lambda) = p_{k+1}(\lambda)\Phi_1(\lambda) + \Phi_2(\lambda) = p_{k+1}(\lambda)\Phi_1(\lambda) + \sum_{l=2}^{k} \nu_l p_l(\lambda) + \nu_1,$$

where $\Phi_1(\lambda)$, $\Phi_2(\lambda)$ are of degree at most $k-1$ and $\nu_1, \dots, \nu_k$ are some scalar coefficients. From the orthogonality of 1, $p_2(\lambda), \dots, p_k(\lambda)$ with respect to both (2.4) and (2.5) it immediately follows that

$$\int_{\lambda_1}^{\lambda_n} \Phi(\lambda)\, \mathrm{d}\omega(\lambda) = \int_{\lambda_1}^{\lambda_n} \nu_1 \, \mathrm{d}\omega(\lambda) = \nu_1 = \int_{\lambda_1}^{\lambda_n} \nu_1 \, \mathrm{d}\omega^{(k)}(\lambda) = \int_{\lambda_1}^{\lambda_n} \Phi(\lambda)\, \mathrm{d}\omega^{(k)}(\lambda).$$

□

Since $\chi_{k-1}(\lambda) = -\chi_k(\lambda)/(\lambda - \theta_l^{(k)})$ + a polynomial of degree at most k–2,

$$\begin{aligned} \int_{\lambda_1}^{\lambda_n} \chi_{k-1}^2(\lambda)\, \mathrm{d}\omega(\lambda) &= -\int_{\lambda_1}^{\lambda_n} \chi_{k-1}(\lambda) \frac{\chi_k(\lambda)}{(\lambda - \theta_l^{(k)})} \, \mathrm{d}\omega(\lambda) \\ &= -\int_{\lambda_1}^{\lambda_n} \chi_{k-1}(\lambda) \frac{\chi_k(\lambda)}{(\lambda - \theta_l^{(k)})} \, \mathrm{d}\omega^{(k)}(\lambda) \\ &= -\sum_{i=1}^{k} \omega_i^{(k)} \left[\chi_{k-1}(\lambda) \frac{\chi_k(\lambda)}{(\lambda - \theta_l^{(k)})} \right]_{\lambda = \theta_i^{(k)}} \\ &= -\omega_l^{(k)} \chi_{k-1}(\theta_l^{(k)}) \chi_k'(\theta_l^{(k)}). \end{aligned}$$

Consequently

$$\omega_l^{(k)} = |(z_{(k)}^l, e^1)|^2 = -\frac{\int_{\lambda_1}^{\lambda_n} \chi_{k-1}^2(\lambda)\,\mathrm{d}\omega(\lambda)}{\chi_{k-1}(\theta_l^{(k)})\chi_k'(\theta_l^{(k)})} = -\frac{\eta_2^2\eta_3^2\dots\eta_k^2}{\chi_{k-1}(\theta_l^{(k)})\chi_k'(\theta_l^{(k)})} \tag{2.6}$$

gives for $l = 1, \dots, k$ the weights $\omega_l^{(k)}$ of the kth Gauss quadrature applied to (2.4). It is worth noticing that this identity gives squares of the first elements of eigenvectors of any unreduced symmetric tridiagonal matrix T_k in terms of the values of the derivative $\chi_k'(\lambda)$ of its characteristic polynomial and of the values of the characteristic polynomial $\chi_{k-1}(\lambda)$ of the reduced matrix with the last row and column omitted. Here (and in several other places below) we do not use the positiveness of the subdiagonal entries of the coefficient matrices in the Lanczos algorithm, since in the theory of unreduced symmetric tridiagonal matrices the positiveness of the subdiagonal entries is insignificant: see Parlett (1980, Lemma 7.2.1).

Clearly we can consider the Lanczos algorithm applied to T_k with the initial vector e^k, leading to the Riemann–Stieltjes integral analogous to (2.5) but with the weights $|(z_{(k)}^l, e^k)|^2$. Then, analogously to (2.6),

$$|(z_{(k)}^l, e^k)|^2 = -\frac{\eta_2^2\eta_3^2\dots\eta_k^2}{\chi_{2,k}(\theta_l^{(k)})\chi_k'(\theta_l^{(k)})}, \tag{2.7}$$

where $\chi_{2,k}(\lambda)$ is the characteristic polynomial of the reduced matrix with the first row and column omitted. It is useful to exploit the knowledge about unreduced symmetric tridiagonal matrices: see, *e.g.*, Wilkinson (1965), Thompson and McEnteggert (1968), Golub (1973), Paige (1971, 1980), Parlett (1980), Elhay, Gladwell, Golub and Ram (1999) and also Strakoš and Greenbaum (1992). For other equivalent expressions for the components of the eigenvectors: see Meurant (2006). In particular,

$$\chi_{2,k}(\theta_l^{(k)})\chi_{k-1}(\theta_l^{(k)}) = \eta_2^2\eta_3^2\dots\eta_k^2,$$

and

$$|(z_{(k)}^l, e^1)|^2 = -\frac{\chi_{2,k}(\theta_l^{(k)})}{\chi_k'(\theta_l^{(k)})}, \qquad |(z_{(k)}^l, e^k)|^2 = -\frac{\chi_{k-1}(\theta_l^{(k)})}{\chi_k'(\theta_l^{(k)})}. \tag{2.8}$$

One of the most beautiful and most powerful features of mathematics is the translation of a given problem into appropriate language where the problem can easily be resolved. The Lanczos algorithm and related mathematical structures offer an excellent example.

- Given A and v^1, the Lanczos algorithm is usually formulated in n-dimensional vector space, and computes the orthonormal basis vectors $v^1, v^2, \dots$ of the Krylov subspaces $\mathcal{K}_k(v^1, A)$, $k = 1, 2, \dots$.

- The Lanczos algorithm can be formulated in terms of the unreduced symmetric tridiagonal matrices T_k, $k = 1, 2, \ldots$, with positive next to diagonal elements, where T_k is appended by a row and a column at each Lanczos step.
- The Lanczos algorithm can be formulated as a Stieltjes procedure in terms of polynomials $p_1(\lambda) = 1$, $p_2(\lambda)$, $p_3(\lambda), \ldots$ orthonormal with respect to the Riemann–Stieltjes integral (2.4).
- The Lanczos algorithm can be formulated in terms of Gauss quadrature approximations (2.5) to the original Riemann–Stieltjes integral (2.4).

Thus the purely algebraic formulation of the problem can be translated to a problem in the classical theory of orthogonal polynomials, and *vice versa*. Similarly, classical tools such as moments, continued fractions and interpolatory quadratures can be directly related to the algebraic tools, developed a century, or many decades, later. These connections are fundamental. They were promoted in modern numerical linear algebra by many distinguished mathematicians. Of these, particular recognition should be given to Gene Golub: see, *e.g.*, Gautschi (2002). For a comprehensive text on generating orthogonal polynomials, Stieltjes procedure and its computational aspects we refer to Gautschi (1982) and the book Gautschi (2004). Other useful information can be found, *e.g.*, in Strakoš and Tichý (2002) and Fischer (1996).

2.3. Approximation from subspaces and the persistence theorem

Approximation results for eigenvalues can be obtained by using the general theory of Rayleigh–Ritz approximations. Good expositions of the theory are given by Stewart (2001) or Parlett (1980). Here is an example of such a result for an eigenpair (λ_i, q^i) of A that we quote from Stewart (2001, p. 285).

Theorem 2.2. Let U be an orthonormal matrix, let $B = U^T AU$ be the matrix Rayleigh quotient, and let θ be the angle between the eigenvector q_i we want to approximate and the range of U, where $Aq^i = \lambda_i q^i$. Then there exists a matrix E satisfying

$$\|E\| \leq \frac{\sin\theta}{\sqrt{1-\sin^2\theta}}\,\|A\|$$

such that λ_i is an eigenvalue of $B + E$.

Then one can apply a general theorem on eigenvalues of perturbed matrices: see Stewart (2001, pp. 285–286).

Corollary 2.3. With the notation of Theorem 2.2, there exists an eigenvalue μ of B such that

$$|\mu - \lambda_i| \leq \|E\|.$$

This shows that if, as a result of an iterative algorithm, we get a small angle θ between the wanted eigenvector q^i and U, then we get an approximate eigenvalue of $B = U^T AU$ converging towards the eigenvalue λ_i of A.

In the case of the Lanczos algorithm we build up a sequence $U = V_k$ and $B = T_k$, $k = 1, 2, \ldots$. We will not further describe the *a priori* error bounds which can be found elsewhere. We will rather concentrate on *a posteriori* bounds and properties important for analysis of the finite precision behaviour.

Let $A = Q\Lambda Q^T$ and $T_k = Z_k \Theta_k Z_k^T$ be the spectral decompositions of A and T_k respectively, $\Lambda = \mathrm{diag}\,(\lambda_i)$, $\Theta_k = \mathrm{diag}\,(\theta_j^{(k)})$. Denote by $\bar{V}_k \equiv Q^T V_k$ the matrix whose columns are composed of the projections of the Lanczos vectors on the eigenvectors of A. Since $T_k = V_k^T A V_k = \bar{V}_k^T \Lambda \bar{V}_k$, we have the following relationship between the Ritz values and the eigenvalues of A.

Proposition 2.4. Let $W_k = (w_{(k)}^1, \ldots, w_{(k)}^k) \equiv Q^T V_k Z_k = \bar{V}_k Z_k$, $w_{(k)}^j \equiv (\xi_{1,j}^{(k)}, \ldots, \xi_{n,j}^{(k)})^T$. Then,

$$\Theta_k = W_k^T \Lambda W_k,$$

$$\theta_j^{(k)} = \sum_{l=1}^{n} (\xi_{l,j}^{(k)})^2 \, \lambda_l \quad \text{and} \quad \sum_{l=1}^{n} (\xi_{l,j}^{(k)})^2 = 1.$$

Proof. The result follows from $T_k = V_k^T A V_k$ and the eigendecompositions of A and T_k, $W_k^T W_k = Z_k^T V_k^T Q Q^T V_k Z_k = I$. □

Clearly the Ritz values are convex combinations of the eigenvalues. We have seen above that a small residual norm $\|r_{(k)}^j\| = \eta_{k+1} |\zeta_{k,j}^{(k)}|$ means that $\theta_j^{(k)}$ is close to some eigenvalue λ_i of A. The following fundamental result proved by Paige, which we formulate for its importance as a theorem, shows that once an eigenvalue λ_i of A is at step t approximated by some Ritz value $\theta_s^{(t)}$ with a small residual norm, it must be approximated to a comparable accuracy by some Ritz value at all subsequent Lanczos steps.

Theorem 2.5. (Persistence Theorem) Let $t < k$. Then,

$$\min_j |\theta_s^{(t)} - \theta_j^{(k)}| \leq \eta_{t+1} |\zeta_{t,s}^{(t)}|.$$

Proof. A proof was given by Paige (1971) using the result in Wilkinson (1965, p. 171); see (3.9) on p. 241 of Paige (1980). □

Theorem 2.5 implies that for every $k > t$ and for any unreduced symmetric tridiagonal extension T_k of T_t there is an eigenvalue $\theta_j^{(k)}$ of T_k within $\eta_{t+1}|\zeta_{t,s}^{(t)}|$ of $\theta_s^{(t)}$. The situation deserves a formal definition. In order to avoid possible subtle ambiguities in the exposition, we slightly modify Definition 1 of Paige (1980).

Definition 2.6. We call an eigenvalue $\theta_s^{(t)}$ of the t by t unreduced symmetric tridiagonal matrix T_t *stabilized* to within $\delta \equiv \eta_{t+1}|\zeta_{t,s}^{(t)}|$. In short, if $\eta_{t+1}|\zeta_{t,s}^{(t)}|$ is small, we call $\theta_s^{(t)}$ stabilized to within small δ.

We will see in Section 4 that, for some Ritz value $\theta_s^{(t)}$ at step t of the Lanczos algorithm, it can happen that for any unreduced symmetric tridiagonal extension T_k of T_t there is an eigenvalue $\theta_j^{(k)}$ very close to $\theta_s^{(t)}$, even though $\delta = \eta_{t+1}|\zeta_{t,s}^{(t)}|$ is not small. However the subsequent Theorem 2.7 and results in Section 4 will also show that in such a case $\theta_s^{(t)}$ must be a close approximation to some eigenvalue of A.

Another useful result in Paige (1971) relates the difference $\theta_s^{(t)} - \theta_j^{(k)}$ to the scalar products of the corresponding eigenvectors.

Theorem 2.7. Using the same notation as in Theorem 2.5,

$$(\theta_s^{(t)} - \theta_j^{(k)})(z_{(k)}^j)^T \begin{bmatrix} z_{(t)}^s \\ 0 \end{bmatrix} = \eta_{t+1}\zeta_{t,s}^{(t)}\zeta_{t+1,j}^{(k)}.$$

Proof. See Paige (1980, p. 241). □

Using this theorem, it is interesting to compare Ritz values on successive steps of the Lanczos algorithm, *i.e.*, take $k = t+1$. Then, because of the interlacing property of Ritz values it is enough to consider $j = s$ or $j = s+1$,

$$(\theta_s^{(t)} - \theta_j^{(t+1)}) \sum_{l=1}^{t} \zeta_{l,s}^{(t)}\zeta_{l,j}^{(t+1)} = \eta_{t+1}\zeta_{t,s}^{(t)}\zeta_{t+1,j}^{(t+1)}.$$

In particular this leads to

$$\eta_{t+1}|\zeta_{t,s}^{(t)}\zeta_{t+1,j}^{(t+1)}| \leq |\theta_s^{(t)} - \theta_j^{(t+1)}|,$$

for $j = s$ or $j = s+1$. Assuming that η_{t+1} is not small (a small η_{t+1} would mean the lucky event indicating closeness to an invariant subspace and convergence of all Ritz values), this shows that if the difference between the Ritz values $|\theta_s^{(t)} - \theta_j^{(t+1)}|$ from two successive steps is small, then the product of the last elements of the corresponding eigenvectors is small. This suggests that Ritz values in two successive steps which are close to each other indicate convergence to some eigenvalue of A. This question has been further investigated by Wülling (2005) following some earlier thoughts in Strakoš and Greenbaum (1992). We will discuss the related results in more detail in Section 4.

3. The conjugate gradient algorithm

The conjugate gradient (CG) algorithm was developed independently by Magnus Hestenes in the US and by Eduard Stiefel in Switzerland, at the beginning of the 1950s. Then they met during a conference in 1951 and wrote a famous joint paper, Hestenes and Stiefel (1952). The algorithm was derived using conjugacy and minimization of functionals. However, it turns out that it is very closely related to the Lanczos algorithm, which can easily be applied for solving linear algebraic systems (Lanczos 1950, 1952).

Consider a symmetric positive definite matrix A, right-hand side b, and the problem $Ax = b$. With an initial vector x^0 and the corresponding residual $r^0 = b - Ax^0$, we can seek an approximate solution to the given linear system in the form $x^k = x^0 + V_k y^k$, where V_k is the matrix of the orthonormal basis vectors of the Krylov subspace $\mathcal{K}_k(v^1, A)$ generated by the Lanczos algorithm with $v^1 = r^0/\|r^0\|$. If we ensure that the residual $r^k = b - Ax^k$ is orthogonal to V_k, then $r^{n+1} = 0$. The resulting method will give (in exact arithmetic) the exact solution in at most n steps and therefore will represent a direct method. Since $r^k = r^0 - AV_k y^k$, this will give

$$0 = V_k^T r^k = V_k^T r^0 - T_k y^k,$$

implying that the coordinates of the approximate solution in V_k are given by the solution of the k by k system with matrix T_k. With the background of the Lanczos algorithm, the whole method can be formulated as

$$T_k y^k = \|r^0\| e^1, \quad x^k = x^0 + V_k y^k. \tag{3.1}$$

The residual r^k is proportional to v^{k+1}, since from the matrix form of (2.1)

$$\begin{aligned} r^k &= r^0 - AV_k y^k = r^0 - (V_k T_k + \eta_{k+1} v^{k+1} (e^k)^T) y^k \\ &= -\eta_{k+1} (y^k, e^k)\, v^{k+1} = (-1)^k v^{k+1} \|r^0\| \eta_2 \cdots \eta_{k+1} / \det(T_k), \end{aligned}$$

using the adjugate of T_k. We next show that ideally (3.1) is equivalent to the CG algorithm of Hestenes and Stiefel (1952).

3.1. Relationship between the formulation of the CG and Lanczos algorithms

In our notation, the Hestenes–Stiefel formulation of the CG algorithm for solving $Ax = b$ with a symmetric positive definite matrix A given in Hestenes and Stiefel (1952) is as follows. Given x^0, $r^0 = b - Ax^0$, $p^0 = r^0$, the subsequent approximate solutions x^k and the corresponding residual vectors $r^k = b - Ax^k$ are computed by:

for $k = 1, 2, \ldots$

$$\begin{aligned} \gamma_{k-1} &= \frac{\|r^{k-1}\|^2}{(p^{k-1}, Ap^{k-1})}, \\ x^k &= x^{k-1} + \gamma_{k-1} p^{k-1}, \\ r^k &= r^{k-1} - \gamma_{k-1} A p^{k-1}, \\ \beta_k &= \frac{\|r^k\|^2}{\|r^{k-1}\|^2}, \\ p^k &= r^k + \beta_k p^{k-1}. \end{aligned} \tag{3.2}$$

With $v^1 = r^0/\|r^0\|$ it can be seen, for example by induction, that

$$\mathcal{K}_k(v^1, A) = \operatorname{span}\{r^0, \ldots, r^{k-1}\} = \operatorname{span}\{p^0, \ldots, p^{k-1}\}.$$

Another straightforward induction (see Hestenes and Stiefel (1952)) gives

$$(r^i, r^j) = 0 \quad \text{and} \quad (p^i, Ap^j) = 0 \quad \text{for } i \neq j.$$

This immediately implies $r^k \perp \mathcal{K}_k(v^1, A)$ and therefore proves the equivalence (up to signs) with the Lanczos algorithm-based formulation described above. Eliminating p^{k-1} from the recurrence for the CG residual, we get, after a simple manipulation,

$$-\frac{1}{\gamma_{k-1}}\, r^k = Ar^{k-1} - \left(\frac{1}{\gamma_{k-1}} + \frac{\beta_{k-1}}{\gamma_{k-2}}\right) r^{k-1} + \frac{\beta_{k-1}}{\gamma_{k-2}}\, r^{k-2}. \tag{3.3}$$

Comparing (3.2) with the Lanczos recurrence (2.1), or more easily with the 3-term recurrence (3.3) for r^k, shows that

$$v^{k+1} = (-1)^k \frac{r^k}{\|r^k\|}. \tag{3.4}$$

If $\hat{v}^{m+1} = 0$ in (2.1), *i.e.*, $\eta_{m+1} = 0$ and the Lanczos algorithm stops, then $r^0, \ldots, A^{m-1} r^0$ are linearly independent while $r^0 \in A\mathcal{K}_m(v^1, A)$, *i.e.*, $r^m = b - Ax^m = r^0 - Au^m = 0$, $u^m \in \mathcal{K}_m(v^1, A)$. Consequently, termination of the Lanczos algorithm implies convergence of CG to the exact solution.

The Lanczos coefficients α_k, η_{k+1} can be determined from the CG coefficients γ_{k-1}, β_k in the following way. Using (3.4) in (3.3) and $\beta_k = \|r^k\|^2/\|r^{k-1}\|^2$, we obtain

$$\frac{\sqrt{\beta_k}}{\gamma_{k-1}}\, v^{k+1} = Av^k - \left(\frac{1}{\gamma_{k-1}} + \frac{\beta_{k-1}}{\gamma_{k-2}}\right) v^k - \frac{\sqrt{\beta_{k-1}}}{\gamma_{k-2}}\, v^{k-1},$$

and therefore we have for $k = 1, 2, \ldots$ the following relations between the coefficients:

$$\alpha_k = \frac{1}{\gamma_{k-1}} + \frac{\beta_{k-1}}{\gamma_{k-2}}, \qquad \beta_0 = 0, \qquad \gamma_{-1} = 1,$$
$$\eta_{k+1} = \frac{\sqrt{\beta_k}}{\gamma_{k-1}}.$$

On the other hand, the CG algorithm (3.2) can be derived from the Lanczos algorithm by the LDL^T decomposition (a variant of the Cholesky decomposition where L is a lower triangular, here lower bidiagonal, factor with ones on the diagonal, and D is a diagonal matrix) of the matrix T_k. This idea, presented in Section 5.7 of Householder (1964), and thoroughly exploited by Paige and Saunders (1975) (see also Stoer (1983)), offers a very insightful explanation of the behaviour of CG when A is indefinite. Since the CG approximate solution satisfies (see (3.1))

$$x^k = x^0 + \|r^0\|\, V_k T_k^{-1} e^1, \quad k = 1, 2, \ldots,$$

it does not exist whenever T_k is singular. When the Lanczos algorithm terminates, the matrix T_m has all its eigenvalues equal to some eigenvalues of the (symmetric and nonsingular) matrix A. Clearly T_m must also be nonsingular. An easy exercise shows that, whenever T_k and T_{k+1} for any $1 \leq k < m - 1$ are simultaneously singular, then $T_{k+2}, \ldots, T_m$ must also be singular, a contradiction. Consequently, at least every second T_k in the sequence $T_1, \ldots, T_m$ must be nonsingular, which means that the CG approximation exists at least at every second step. It cannot, in general, be computed via the formulas (3.2), since the Cholesky decomposition of the singular T_k does not exist and the implementation (3.2) in such a case breaks down. If T_k is close to singular, the Cholesky decomposition is poorly determined numerically for all $j > k$, and so is the recurrence (3.2).

Paige and Saunders showed in a very instructive way how to compute the CG approximation x^k when it exists, and how to avoid numerical instabilities when T_k is close to singular. Their approach is based on exploiting the Lanczos algorithm, but it does not require storing the Lanczos basis V_k. The CG approximations x^k are computed recursively with the help of auxiliary approximations to the solution which exist at every step and which define the method called SYMMLQ. They also suggested an effective implementation of the Krylov subspace method MINRES, which minimizes residual norms and is used for symmetric indefinite problems. Paige and Saunders (1975) resolved open problems that had arisen from the earlier work by Fridman (1963) and Luenberger (1969, 1970). The relationship between different implementations was further studied by Fletcher (1976), and a numerically stable variant of the OD algorithm of Fridman (1963) called STOD was suggested by Stoer and Freund (1982); see the overview in Stoer (1983).

3.2. Orthogonality and optimality properties

In some important applications leading to systems with symmetric positive definite matrices it is natural to measure the error in the A-norm,

$$\|x-u\|_A = (x-u, A(x-u))^{\frac{1}{2}},$$

since the A-norm can be interpreted as the discretized measure of energy which is to be minimized: see, *e.g.*, Arioli (2004) and Arioli, Noulard and Russo (2001). The CG algorithm is, from this point of view, best suited to solving such problems, since it minimizes the A-norm of the error among all possible approximations from the same Krylov subspaces,

$$\|x-x^k\|_A = \min_{u \in x^0 + \mathcal{K}_k(v^1,A)} \|x-u\|_A. \tag{3.5}$$

Indeed, in order to reach the minimum (3.5), $x - x^k$ must be orthogonal with respect to the inner product defined by the matrix A to the Krylov subspace $\mathcal{K}_k(v^1, A)$, *i.e.*,

$$0 = (r^j, A(x-x^k)) = (r^j, r^k) = (v^{j+1}, r^k) \quad \text{for } j = 0, \dots, k-1,$$

which uniquely determines the approximate solution x^k generated by the CG algorithm described above.

Using the A-orthogonality of the direction vectors p^0, p^1, $p^2, \dots$, it can be seen from (3.2) that the kth error, assuming that CG terminates at step m with $x^m = x$, can conveniently be written

$$\begin{aligned}
x - x^0 &= \sum_{l=1}^{k} \gamma_{l-1} p^{l-1} + x - x^k = \sum_{l=1}^{m} \gamma_{l-1} p^{l-1}, \\
x - x^k &= \sum_{l=k+1}^{m} \gamma_{l-1} p^{l-1}, \\
\|x-x^0\|_A^2 &= \sum_{l=1}^{m} \gamma_{l-1}^2 (p^{l-1}, Ap^{l-1}) = \sum_{l=1}^{m} \gamma_{l-1} \|r^{l-1}\|^2, \\
\|x-x^k\|_A^2 &= \sum_{l=k+1}^{m} \gamma_{l-1} \|r^{l-1}\|^2,
\end{aligned}$$

and, finally,

$$\|x - x^0\|_A^2 = \sum_{l=1}^{k} \gamma_{l-1} \|r^{l-1}\|^2 + \|x - x^k\|_A^2, \quad k = 1, 2, \ldots, m. \tag{3.6}$$

The last identity reflects the mathematical elegance of the CG algorithm, but it also demonstrates complications which have to be dealt with in finite precision arithmetic computations. The derivation of (3.6) presented above relies upon the global A-orthogonality of all vectors $p^0, \ldots, p^{m-1}$,

$$(p^i, Ap^j) = 0 \quad \text{for } i \neq j,$$

which holds ideally, but which is not preserved numerically. Unless (3.6) is supported by arguments that also hold numerically, it cannot be used for the results of finite precision computations. This point is of crucial importance. A patient reader will, however, see in Section 5 that (3.6) indeed holds, up to a small insignificant inaccuracy, also numerically: see Strakoš and Tichý (2002).

As the relationship of CG with the Lanczos algorithm suggests, there is of course a three-term recurrence formulation ideally equivalent to (3.2): see, *e.g.*, Rutishauser (1959) and Hageman and Young (1981). The three-term recurrence is reputed to have some disadvantages concerning the maximal attainable accuracy: see Section 5 and Gutknecht and Strakoš (2000). However, it is of some interest for parallel computation.

Convergence bounds for CG are typically derived from its polynomial formulation, which follows from (3.2) (see, *e.g.*, the 3-term recurrence (3.3) for r^k):

$$\begin{aligned} r^k &= \varphi_k(A) r^0, \qquad \varphi_k(0) = 1, \\ x - x^k &= \varphi_k(A)(x - x^0), \end{aligned}$$

where $\varphi_k(0) = 1$ follows, *e.g.*, from induction on the 3-term recurrence for r^k, while the $x - x^k$ expression follows since A is nonsingular. The 3-term recurrences and the definition of the Lanczos polynomials p_k in (2.3) lead to

$$\varphi_k(\lambda) = \frac{p_{k+1}(\lambda)}{p_{k+1}(0)}.$$

Here the assumption that A is symmetric positive definite guarantees that all roots of p_k are no less than $\lambda_1 > 0$, and therefore $p_k(0) \neq 0$. With the spectral decomposition of A we can easily obtain the following theorem.

Theorem 3.1.

$$\|r^k\|^2 = \|r^0\|^2 \sum_{i=1}^{n} \prod_{l=1}^{k} \left(1 - \frac{\lambda_i}{\theta_l^{(k)}}\right)^2 \omega_i,$$

$$\|x - x^k\|^2 = \|r^0\|^2 \sum_{i=1}^{n} \prod_{l=1}^{k} \left(\frac{1}{\lambda_i} - \frac{1}{\theta_l^{(k)}}\right)^2 \omega_i,$$

$$\|x - x^k\|_A^2 = \|r^0\|^2 \sum_{i=1}^{n} \prod_{l=1}^{k} \left(\frac{1}{\sqrt{\lambda_i}} - \frac{\sqrt{\lambda_i}}{\theta_l^{(k)}}\right)^2 \omega_i,$$

where, as in (2.4), $\omega_i = |(v^1, q^i)|^2, v^1 = r^0/\|r^0\|$.

Proof. We remark that $x - x^0 = A^{-1}r^0$ and

$$\frac{p_{k+1}(\lambda_i)^2}{p_{k+1}(0)^2} = \prod_{l=1}^{k} \left(1 - \frac{\lambda_i}{\theta_l^{(k)}}\right)^2.$$

By using this, the proofs become straightforward. □

Since $\|x - x^k\|_A \leq \|\varphi_k(A)\| \, \|x - x^0\|_A$, we have the bound

$$\|x - x^k\|_A \leq \min_{\varphi \in \Pi_k} \max_i |\varphi(\lambda_i)| \, \|x - x^0\|_A, \tag{3.7}$$

where Π_k denotes the set of all polynomials of degree at most k with the constant term equal to one (value one at zero). Any bound which is based on (3.7) holds for *any* initial error (initial residual) and therefore represents a worst case bound. Therefore, even analytic knowledge of the value

$$\min_{\varphi \in \Pi_k} \max_i |\varphi(\lambda_i)| = \left| \sum_{l=1}^{k+1} (-1)^{l-1} \prod_{j=1, j \neq l}^{k+1} \frac{\mu_j}{\mu_j - \mu_l} \right|^{-1}, \tag{3.8}$$

where $\{\mu_1, \ldots, \mu_{k+1}\}$ is some properly chosen subset of the distinct eigenvalues of A (on which the kth minimax polynomial assumes its maximum absolute value – see Greenbaum (1979) and Liesen and Tichý (2005)) does not help in describing possible differences in the behaviour of CG for different initial residuals (right-hand sides): *cf.* Beckermann and Kuijlaars (2002) and Strakoš and Tichý (2005). The error bound (3.7) with (3.8) is sharp, *i.e.*, at any given step k it can be attained with a certain initial vector (which depends on k).

The generally known bound is derived from using the kth degree Chebyshev polynomial on the spectral interval $[\lambda_1, \lambda_n]$, which gives

$$\frac{\|x - x^k\|_A}{\|x - x^0\|_A} \leq 2\left[\left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}\right)^k + \left(\frac{\sqrt{\kappa} + 1}{\sqrt{\kappa} - 1}\right)^k\right]^{-1} \leq 2\left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}\right)^k, \tag{3.9}$$

where $\kappa \equiv \kappa(A) \equiv \lambda_n/\lambda_1$ denotes the condition number of A. This bound is frequently attributed to Kaniel (1966) or Daniel (1967), but it appeared even earlier in the paper by Meinardus (1963); see Li (2005). Though it is useful in the analysis of many model problems, it cannot be identified, except for some specific cases, with convergence behaviour of CG. The bound (3.9) describes linear convergence; it shows that the closer the condition number is to 1, the faster is the convergence of CG when measured in the A-norm. However, we have seen in Theorem 3.1, and we are going to see again in the next section, that CG convergence depends on the distribution of all the eigenvalues of A and not just on the condition number. If the distribution of eigenvalues is favourable, then the convergence of CG significantly accelerates as k increases. For an early investigation of convergence behaviour in relation to the spectrum see, *e.g.*, Axelsson and Linskog (1986) and van der Vorst (1982).

3.3. Estimating quadratic forms and identities for the error norms in CG

We have seen that given A and v^1 respectively r^0, $v^1 = r^0/\|r^0\|$, the Lanczos algorithm and CG can be formulated in terms of the orthogonal polynomials $1, p_1(\lambda), p_2(\lambda), \ldots,$ and therefore in terms of the Gauss quadrature of the Riemann–Stieltjes integral determined by A, v^1. In this way, the Lanczos algorithm and CG can be viewed as matrix representations of Gauss quadrature. That explains the subtle character of problems related to the Lanczos and CG convergence behaviour. Here we will go a step forward to show how the A-norm of the error and the Euclidean norm of the error in CG can be computed using Gauss quadrature and how they can be bounded using some of its modifications.

Computing the A-norm of the error $\epsilon^k \equiv x - x^k$ is closely related to approximating quadratic forms. This has been studied extensively by Gene Golub with many collaborators during the last thirty-five years. The relationship to Gauss quadrature was summarized in Golub and Meurant (1994); see also Golub and Meurant (1997), Golub and Strakoš (1994), Fischer (1996), Golub and von Matt (1991) and Calvetti, Morigi, Reichel and Sgallari (2000). With A symmetric positive definite, the problem considered by Golub and Meurant (1994) was to find upper and lower bounds (or approximations) for the entries of a function of a matrix. This problem leads to the quadratic form

$$u^T f(A) u,$$

where u is a given vector and f is a smooth (possibly C^∞) function on a given interval of the real line. The more general case $u^T f(A) v$ can easily be converted into the previous one using the well-known identity

$$u^T f(A) v = \frac{1}{2}\big(u^T f(A) u + v^T f(A) v - (u-v)^T f(A)(u-v)\big).$$

This problem is of great importance in computational sciences such as computational quantum chemistry or solid state physics. The example we are interested in for CG is $f(\lambda) = 1/\lambda$. This is related to the problem of computing the A-norm of the error, since the error ϵ^k is related to the residual r^k by the equation $A\epsilon^k = r^k$. Therefore,

$$\|\epsilon^k\|_A^2 = (A\epsilon^k, \epsilon^k) = (A^{-1}r^k, r^k) = (r^k)^T A^{-1} r^k.$$

Using the spectral decomposition of A (as above, for simplicity we assume that the eigenvalues of A are distinct and ordered, $\lambda_1 < \lambda_2 < \cdots < \lambda_n$)

$$f(A) = Qf(\Lambda)Q^T.$$

Therefore,

$$\begin{aligned} u^T f(A)u &= u^T Q f(\Lambda) Q^T u \\ &= y^T f(\Lambda) y, \\ &= \sum_{i=1}^{n} \omega_i f(\lambda_i),\ y \equiv Q^T u,\ \omega_i \equiv |(u, q^i)|^2. \end{aligned}$$

We assume, without loss of generality, that $\|u\| = 1$. Clearly, as in Section 2, the last sum is a Riemann–Stieltjes integral, namely

$$I[f] = u^T f(A) u = \int_{\lambda_1}^{\lambda_n} f(\lambda)\, \mathrm{d}\omega(\lambda),$$

where, as above, the distribution function ω is the non-decreasing piecewise constant function, with points of increase at the eigenvalues of A, and discontinuities of sizes $\omega_1, \ldots, \omega_n$.

We are looking for upper and lower bounds $L[f]$ and $U[f]$ for $I[f]$,

$$L[f] \le I[f] \le U[f].$$

They can be obtained, among other techniques, by using Gauss, Gauss–Radau and Gauss–Lobatto quadrature formulas; for the pioneering work see, in particular, Dahlquist, Eisenstat and Golub (1972) and Dahlquist, Golub and Nash (1978). We shall use the general formula

$$\int_{\lambda_1}^{\lambda_n} f(\lambda)\, \mathrm{d}\omega(\lambda) = \sum_{j=1}^{k} \omega_j^{(k)} f(\tau_j^{(k)}) + \sum_{l=1}^{M} \vartheta_l^{(M)} f(\sigma_l^{(M)}) + R^{k,M}[f],$$

where the weights $\omega_j^{(k)}, j = 1, \ldots, k$, $\vartheta_l^{(M)}, l = 1, \ldots, M$, and the nodes $\tau_j^{(k)}, j = 1, \ldots, k$ are to be determined, while the nodes $\sigma_l^{(M)}, l = 1, \ldots, M$ are prescribed; see Davis and Rabinowitz (1984), Gautschi (1968, 1985) and Golub and Welsch (1969). It is well known (see the excellent survey by Gautschi (1981)) that

$$R^{k,M}[f] = \frac{f^{(2k+M)}(\eta)}{(2k+M)!} \int_{\lambda_1}^{\lambda_n} \prod_{l=1}^{M} (\lambda - \sigma_l^{(M)}) \left[\prod_{j=1}^{k} (\lambda - \tau_j^{(k)}) \right]^2 \mathrm{d}\omega(\lambda),$$

where $\lambda_1 < \eta < \lambda_n$. If $M = 0$, this leads to the Gauss rule with no prescribed nodes. If $M = 1$ and we fix the node at one of the end points, $\sigma_1^{(1)} = \lambda_1$ or $\sigma_1^{(1)} = \lambda_n$, we have the Gauss–Radau formula. If $M = 2$ and $\sigma_1^{(2)} = \lambda_1$, $\sigma_2^{(2)} = \lambda_n$, this is the Gauss–Lobatto formula.

As presented above, the nodes and weights in the Gauss rule are given by the eigenvalues of T_k (the Ritz values $\theta_j^{(k)}$) and the squared first elements of the normalized eigenvectors of T_k respectively (*cf.* Golub and Welsch (1969)), where T_k is the tridiagonal matrix of the recurrence coefficients generated by the Lanczos algorithm for A and the starting vector $v^1 = u$. For the Gauss quadrature rule, we have

$$\int_{\lambda_1}^{\lambda_n} f(\lambda)\,\mathrm{d}\omega(\lambda) \equiv L_G^{(k)}[f] + R_G^{(k)}[f],$$

with

$$L_G^{(k)}[f] = \sum_{j=1}^{k} \omega_j^{(k)} f(\theta_j^{(k)}) = (e^1)^T f(T_k)\, e^1,$$

$$R_G^{(k)}[f] = \frac{f^{(2k)}(\eta)}{(2k)!} \int_{\lambda_1}^{\lambda_n} \left[\prod_{j=1}^{k} (\lambda - \theta_j^{(k)}) \right]^2 \mathrm{d}\omega(\lambda).$$

Consequently, in order to compute the value of the quadrature, we do not need to determine its nodes and weights. Suppose f is such that $f^{(2k)}(\xi) > 0$, $\forall k$, $\forall \xi$, $\lambda_1 < \xi < \lambda_n$. Then

$$L_G[f] \le I[f], \quad k = 1, 2, \ldots$$

and the Gauss rule provides in this case a lower bound for the quadratic form. Note that this applies for $f(\lambda) = 1/\lambda$.

To summarize, for estimating the A-norm of the error in CG, we obtain

$$\|\epsilon^0\|_A^2 = (A^{-1}r^0, r^0) = \|r^0\|^2 (T_n^{-1} e^1, e^1),$$

$$L_G^{(k)}\left[\frac{1}{\lambda}\right] = (T_k^{-1} e^1, e^1),$$

$$\|r^0\|^2 \left[(T_n^{-1} e^1, e^1) - (T_k^{-1} e^1, e^1) \right] = \|r^0\|^2\, R_G^{(k)}\left[\frac{1}{\lambda}\right] \ge 0.$$

and we formulate the key point as a theorem.

Theorem 3.2. Using the previous notation, we get the following identities for the A-norm of the error in CG:

$$\|\epsilon^k\|_A^2 = \|r^0\|^2 \, R_G^{(k)}\left[\frac{1}{\lambda}\right] = \|r^0\, \|^2[(T_n^{-1}e^1, e^1) - (T_k^{-1}e^1, e^1)],$$

i.e.,

$$\|\epsilon^k\|_A^2 = \|r^0\|^2 \left[\sum_{j=1}^{n} \frac{(z^j_{(n)}, e^1)^2}{\lambda_j} - \sum_{j=1}^{k} \frac{(z^j_{(k)}, e^1)^2}{\theta_j^{(k)}}\right].$$

Proof. This result is known: see Dahlquist, Golub and Nash (1978). The proof given here is, however, different from the original one.

By using the definition of the A-norm and $A\epsilon^k = r^k = r^0 - AV_ky^k$, we have

$$\|\epsilon^k\|_A^2 = (A\epsilon^k, \epsilon^k) = (A^{-1}r^0, r^0) - 2(r^0, V_ky^k) + (AV_ky^k, V_ky^k).$$

Since $T_ky^k = \|r^0\|e^1$,

$$(r^0, V_ky^k) = \|r^0\|^2(T_k^{-1}e^1, e^1),$$

and

$$(AV_ky^k, V_ky^k) = (V_k^TAV_ky^k, y^k) = (T_ky^k, y^k) = \|r^0\|^2(T_k^{-1}e^1, e^1),$$

the first identity is proved. The rest follows from the spectral decomposition of T_n and T_k. □

We can conclude that the square of the A-norm of the CG error at the kth step divided by $\|r^0\|^2$ represents the remainder of the kth Gauss quadrature approximation of the corresponding Riemann–Stieltjes integral determined by A and $u = r^0/\|r^0\|$. Therefore the Gauss quadrature (here represented fully in the matrix form) gives lower bounds for the A-norm of the CG error. Upper bounds can be obtained with the Gauss–Radau rule if we have estimates of λ_1: see Golub and Meurant (1994). The second identity reflects the complicated relationship between the CG rate of convergence and the convergence of the Ritz values towards the eigenvalues of A. Another point on this is given by the following theorem.

Theorem 3.3. For all k, there exists ϑ_k, $\lambda_1 \leq \vartheta_k \leq \lambda_n$ such that the A-norm of the error is given by

$$\|\epsilon^k\|_A^2 = \frac{\|r^0\|^2}{\vartheta_k^{2k+1}} \sum_{i=1}^{n} \left[\prod_{j=1}^{k} (\lambda_i - \theta_j^{(k)})^2\right] \omega_i,$$

where $\omega_i = |(v^1, q^i)|^2$.

Proof. The remainder of approximation $\int_{\lambda_1}^{\lambda_n} f(\lambda)\,\mathrm{d}\omega(\lambda)$ with the Gauss quadrature is

$$\frac{f^{(2k)}(\vartheta)}{(2k!)}\int_{\lambda_1}^{\lambda_n}\prod_{j=1}^{k}(\lambda-\theta_j^{(k)})^2\,\mathrm{d}\omega(\lambda),$$

with $\lambda_1 \leq \vartheta \leq \lambda_n$. Using $f = 1/\lambda$, this gives the statement of the theorem. □

This shows that, in exact arithmetic, when a Ritz value has converged to an eigenvalue of A, we have eliminated the component of the initial residual in the direction of the corresponding eigenvector of A. For related results on this subject we refer in particular to Axelsson and Linskog (1986) and van der Sluis and van der Vorst (1986).

The statement from Theorem 3.2 can be written as

$$\|\epsilon^0\|_A^2 = \|r^0\|^2(T_k^{-1}e^1, e^1) + \|\epsilon^k\|_A^2.$$

This recalls (3.6); restated as a theorem it reads as follows.

Theorem 3.4.

$$\|\epsilon^0\|_A^2 = \sum_{l=1}^{k}\gamma_{l-1}\|r^{l-1}\|^2 + \|\epsilon^k\|_A^2.$$

This means that the Gauss quadrature approximation $(T_k^{-1}e^1, e^1)$ can easily be computed as

$$L_G^{(k)}\left[\frac{1}{\lambda}\right] = \sum_{l=1}^{k}\gamma_{l-1}\frac{\|r^{l-1}\|^2}{\|r^0\|^2}.$$

Theorem 3.4 is in fact proved in Theorem 6:1 of Hestenes and Stiefel (1952, p. 416). The result was later derived and used, independently of the original paper, by many authors: see, *e.g.*, Deuflhard (1994), Axelsson and Kaporin (2001), Greenbaum (1997*a*) and Arioli (2004). It was used without being explicitly stated in Golub and Meurant (1997). In some of the given references the motivation is estimation of the error in CG.

The importance of the formula in Theorem 3.4 was emphasized by Strakoš and Tichý (2002, 2005). The first paper points to the original reference Hestenes and Stiefel (1952), proves the equivalence with the Gauss quadrature and gives an elementary proof which does not use the global orthogonality of the residuals or the global A-orthogonality of the direction vectors

$$\begin{aligned}\|\epsilon^k\|_A^2 - \|\epsilon^{k+1}\|_A^2 &= \|x - x^{k+1} + x^{k+1} - x^k\|_A^2 - \|\epsilon^{k+1}\|_A^2\\ &= \|x^{k+1} - x^k\|_A^2 + 2(x - x^{k+1})^T A(x^{k+1} - x^k)\\ &= \gamma_k^2(p^k, Ap^k) + 2(r^{k+1}, x^{k+1} - x^k) = \gamma_k\|r^k\|^2.\end{aligned}$$

The independence of the result on the global orthogonality is fundamental: it allows one to perform a detailed rounding error analysis, and to build up a mathematically rigorous argument that justifies validity of the given identity in finite precision CG computations. Without such an analysis, results derived using assumptions violated because of rounding errors are numerically useless, since they can give misleading information. Estimating errors in CG will be reviewed in Section 5.

Regarding the Euclidean norm, Theorem 6:3 of Hestenes and Stiefel (1952, pp. 416–417) gives the following result.

Theorem 3.5.

$$\|\epsilon^k\|^2 - \|\epsilon^{k+1}\|^2 = \frac{\|\epsilon^k\|_A^2 + \|\epsilon^{k+1}\|_A^2}{\mu(p^k)},$$

with

$$\mu(p^k) = \frac{(p^k, Ap^k)}{\|p^k\|^2}.$$

Hence, the Euclidean norm of the error is monotonically decreasing.

There is another expression for the Euclidean norm of the error derived by Meurant (2005).

Theorem 3.6.

$$\|\epsilon^k\|^2 = \|r^0\|^2[(e^1, T_n^{-2}e^1) - (e^1, T_k^{-2}e^1)] - 2\frac{(e^k, T_k^{-2}e^1)}{(e^k, T_k^{-1}e^1)}\|\epsilon^k\|_A^2.$$

This result (which is, of course, equivalent to the expression obtained by Hestenes and Stiefel) allows us, by using the spectral decomposition of T_n and T_k, to relate the norm of the error to the eigenvalues of A and to the Ritz values.

4. The Lanczos algorithm in finite precision

As an example, we consider a matrix that was introduced by Strakoš (1991) and used by Strakoš and Greenbaum (1992). The matrix of dimension n is diagonal, with the eigenvalues

$$\lambda_i = \lambda_1 + \left(\frac{i-1}{n-1}\right)(\lambda_1 - \lambda_n)\,\rho^{n-i}, \quad i = 1, \dots, n.$$

The parameter ρ controls the distribution of the eigenvalues within the interval $[\lambda_1, \lambda_n]$. We shall use $n = 30$, $\lambda_1 = 0.1$, $\lambda_n = 100$ and $\rho = 0.8$, which gives well-separated large eigenvalues, and call this matrix D30. Figure 4.1 shows $\log_{10}$ of the elements of $|V_{30}^T V_{30}|$, each plotted against its index pair i, j, for the Lanczos algorithm applied to $A = \text{D30}$ with the initial vector v^1 having equal components. Ideally $V_{30}^T V_{30}$ should be the identity matrix.

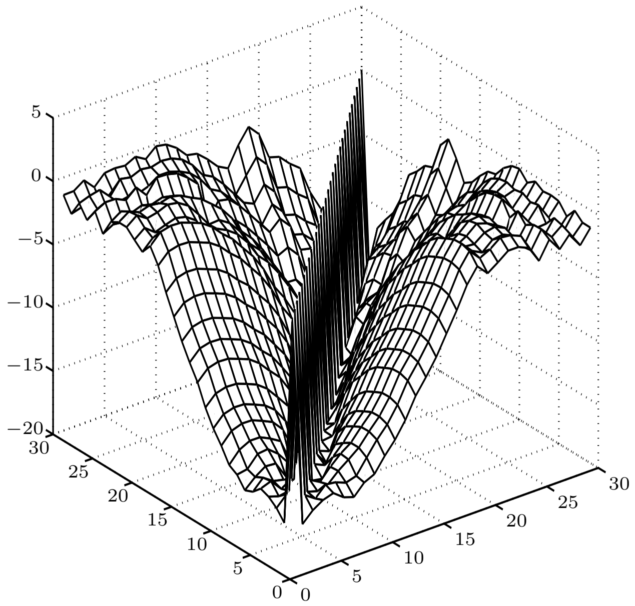

Figure 4.1. Matrix D30, $\log_{10}$ of $|V_{30}^T V_{30}|$.

But numerically this matrix is far from the identity. The magnitude of most nondiagonal entries is much larger than the level of elementary round-off. On the contrary, the magnitude of many of them is $\mathcal{O}(1)$.

It has been known since Lanczos (1950) that the behaviour of the algorithm in finite precision arithmetic is far from ideal. The Lanczos vectors v^k do not stay orthogonal as they ideally should. This also means that $V_k^T A V_k$ is no longer a tridiagonal matrix and the computed tridiagonal matrix T_k is not the projection of A on the computed Krylov subspace. Therefore the computed T_n is not similar to A, and the algorithm typically does not deliver sufficiently accurate numerical approximations to all eigenvalues of A in n iterations. Moreover, some eigenvalues of A can be numerically approximated by sets of very close Ritz values (called multiple copies) and it is difficult to decide whether such Ritz values are good approximations of the genuine close eigenvalues of A or just artifacts caused by rounding errors.

All these troubles are easily observable from numerical experiments; they are pointed out in practically all textbook expositions of the effect of rounding errors in the Lanczos algorithm. The same attention is, however, not paid to the analysis and resolution of the worst consequence of rounding errors, which every practical user of the Lanczos algorithm is inevitably faced with. Since numerically the Lanczos vectors are not orthogonal and they can very soon become linearly dependent, there is no guarantee whatsoever that the norm of the Ritz vector

$$x^j_{(k)} = V_k z^j_{(k)}$$

is close to one, where we assume for simplicity that the eigenvector $z^j_{(k)}$ of the computed T_k corresponding to the Ritz value $\theta_j^{(k)}$ is determined exactly. It can even numerically vanish, with norm close to the machine precision. In such a case it is absolutely unclear whether a small value of the convergence criterion $\eta_{k+1}|\zeta_{k,j}^{(k)}|$ described in Section 2.1 means convergence of $\theta_j^{(k)}$ to any eigenvalue λ_i of A using the finite precision Lanczos algorithm. Please notice that the trouble is not in computing $\eta_{k+1}|\zeta_{kj}^{(k)}|$ more or less accurately from T_k. We will assume (with negligible and easily quantifiable inaccuracy) that the exact spectral decomposition of T_k is known and that the quantity in question is computed exactly. The trouble consists in the fact that the derivation of the bound for $\min_i |\lambda_i - \theta_j^{(k)}|$ is based on the assumption that $V_k^T V_k = I$, which is usually drastically violated in finite precision computations.

It may seem that all the mathematical theory behind the Lanczos algorithm is lost because of rounding errors and the loss of orthogonality. Without a proper rounding error analysis we can not even interpret any Ritz value as a close approximation to an eigenvalue of A.

The first, and at the same time most original and most significant step in explaining the behaviour of the Lanczos algorithm in finite precision arithmetic was made by Chris Paige in his PhD thesis (Paige 1971). He proved the fundamental result that loss of orthogonality goes hand in hand with convergence of Ritz values, and developed a theory which formed a basis for practically all further progress in this area (except, perhaps, investigations of the maximal attainable accuracy of Krylov subspace linear algebraic solvers). His results were published in a series of papers, Paige (1972, 1976, 1980). Most of them are also included, together with some subsequent developments, in the classical monograph by Beresford Parlett (1980).

In this paper we would like to (partially) address the following questions.

- What theoretical properties of the Lanczos algorithm remain (with an insignificant inaccuracy) true in finite precision arithmetic?
- How can we describe the mechanics of the loss of orthogonality among the Lanczos vectors?
- What happens numerically to the equivalence of the Lanczos and CG algorithms as well as to the equivalence with orthogonal polynomials and Gauss quadrature?
- How do we evaluate convergence of CG in finite precision arithmetic?

The length of this paper is limited. Unless there is a good expository reason for presenting a part or the whole proof, in the following review we present theorems and statements without proofs. The reader interested in proofs or further details is referred to the original works.

4.1. Finite precision arithmetic

We use the standard model for floating point computations: see, *e.g.*, Higham (2002). Where needed we will denote by $fl(X)$ the result of the computation of X or denote the computed quantities by $\tilde{}$. For any of the four basic operations $(+, -, *, /)$ denoted by op, we have

$$fl(x \text{ op } y) = (x \text{ op } y)(1+\delta), \quad |\delta| \le u_M$$

u_M being the unit round-off which is $(1/2)\beta^{1-t}$, where β is the base and t is the number of digits in the mantissa. This bound is obtained using rounding to the nearest floating point number, but this is generally the case. In IEEE standard 754, double precision, $(\beta = 2, t = 53)$ and

$$u_M = 1.110223024625157 \times 10^{-16},$$

which is half of the machine precision unit (machine epsilon) $\varepsilon_M = \beta^{1-t}$ representing the distance from 1 to the next-larger floating point number.

4.2. Paige's theory, loss of orthogonality, stabilization and convergence

The fundamental work of Chris Paige started at the end of the 1960s with some technical reports and papers, Paige (1969*a*), (1969*b*), (1970*a*) and (1970*b*), whose results led to his PhD thesis, Paige (1971), which clearly stated and proved, contrary to the common wisdom of the time, that even though the Lanczos algorithm in finite precision arithmetic does not keep its theoretical properties, it nevertheless works well as a reliable and highly efficient numerical tool for computing highly accurate approximations of dominant, and often other, eigenvalues of large sparse matrices.

The main contributions of Chris Paige presented in his thesis, or further developed from it, can be described as follows. He derived bounds for the local rounding errors in the Lanczos algorithm. He showed that the last elements of the eigenvectors of the computed tridiagonal matrix T_k indeed reliably tell us how well the eigenvalues of A are approximated by Ritz values, and how we can always obtain useful intervals containing eigenvalues of A. The computed Ritz values always lie between the extreme eigenvalues of A to within a small multiple of the machine precision. Moreover, at least one small interval containing an eigenvalue of A is found by the nth iteration. The algorithm behaves numerically like the Lanczos algorithm with full reorthogonalization until a very close eigenvalue approximation is found. Of course, the most (and rightly) celebrated of the results from

Paige's thesis is his proof that loss of orthogonality follows a rigorous pattern and implies that some Ritz values have converged.

Paige used a handy notation to bound combinations of rounding errors. Given $\varepsilon_1, \dots, \varepsilon_p$ with each $|\varepsilon_i| \le u_M$, then there exists a value α such that

$$\prod_{i=1}^{p}(1+\varepsilon_i) = \alpha^p, \qquad |\alpha - 1| \le u_M.$$

Let $D(\alpha)$ represent a diagonal matrix with elements not necessarily equal but satisfying the above bounds. Rules for manipulating such quantities are

$$\alpha^p \alpha^q = \alpha^{p+q},$$
$$x = \alpha(y+z) \Rightarrow x = \alpha y + \alpha z,$$
$$x = \left(\frac{\alpha^p}{\alpha^q}\right) y \text{ or } x = \alpha^p \alpha^q y \Rightarrow x = [1 + (p+q)\varepsilon] y,$$

where $|\varepsilon| \le 1.01 u_M$. Using this notation, for the inner product we have, neglecting higher order terms in ε,

$$fl(x^T y) = x^T D(\alpha^n)\, y = x^T y + n\varepsilon\, |x^T|\, |y|,$$

and for the computation of the Euclidean norm,

$$fl(x^T x) = \alpha^n x^T x.$$

For the matrix vector product Paige used

$$fl(Ax) = (A + \delta A)\, x, \quad |\delta A| \le m_A \varepsilon\, |A|,$$

where m_A is the maximum number of nonzero elements per row. This leads to

$$\|\delta A\| \le m_A \varepsilon\, \|\, |A|\, \|.$$

Let β be such that $\|\, |A|\, \| = \beta \|A\|$. Then

$$\|\delta A\| \le m_A \varepsilon \beta\, \|A\|.$$

Then Paige's thesis analysed various implementations of the Lanczos algorithm. This part of his work was published and complemented in Paige (1972), which justifies (2.1) as the preferable variant of the Lanczos algorithm.

In the subsequent part of the thesis, published in Paige (1976), the implementation (2.1) was studied further. The results were gathered in a theorem; see also Paige (1980, (2.10)–(2.16)).

Theorem 4.1. *Let* $\varepsilon_0 = 2(n+4)\varepsilon_M < \frac{1}{12}$, $\varepsilon_1 = 2(7 + m_A\beta)\varepsilon_M$. *Then the computed results of the Lanczos algorithm in finite precision arithmetic satisfy the matrix identity*

$$AV_k = V_k T_k + \eta_{k+1} v^{k+1} (e^k)^T + \delta V_k,$$

where, for $j = 1, 2, \ldots, k$,

$$\begin{aligned} |(v^{j+1})^T v^{j+1} - 1| &\leq \varepsilon_0, \\ \|\delta v^j\| &\leq \varepsilon_1 \|A\|, \\ \eta_{j+1} |(v^j)^T v^{j+1}| &\leq 2\varepsilon_0 \|A\|, \\ |\eta_j^2 + \alpha_j^2 + \eta_{j+1}^2 - \|Av^j\|^2| &\leq 4j(3\varepsilon_0 + \varepsilon_1)\|A\|^2. \end{aligned}$$

Since the local errors collected in δV_k are minor, the computed quantities satisfy the identity which formally looks very close to its exact precision counterpart. The presence of the extra term δV_k has, however, significant consequences. As an immediate one we get the following theorem.

Theorem 4.2. (Paige 1971, 1976 (21)–(23)) If R_k is the strictly upper triangular part of $V_k^T V_k$ such that

$$V_k^T V_k = R_k^T + \operatorname{diag}((v^j)^T v^j) + R_k,$$

then

$$T_k R_k - R_k T_k = \eta_{k+1} V_k^T v^{k+1} (e^k)^T + \delta R_k,$$

where δR_k is upper triangular with elements such that $|(\delta R_k)_{1,1}| \leq 2\varepsilon_0 \|A\|$, and for $j = 2, 3, \ldots, k$

$$|(\delta R_k)_{j,j}| \leq 4\varepsilon_0 \|A\|, \qquad |(\delta R_k)_{j-1,j}| \leq 2(\varepsilon_0 + \varepsilon_1)\|A\|,$$
$$|(\delta R_k)_{i,j}| \leq 2\varepsilon_1 \|A\|, \quad i = 1, 2, \ldots, j-2.$$

This shows how the loss of orthogonality propagates through the algorithm.

A paper which finalizes publication of many of the results presented in Paige's thesis was published in 1980 in *Linear Algebra and its Applications* (Paige 1980). This paper is truly seminal; in this time of malign overemphasis on quantity of publications it should serve as a textbook example of a paper which could easily be split, although not for good reasons, into several publishable papers. The effect would have been similar to cutting a large diamond of superb quality into several pieces of more common size. The resulting pieces would still be easy to sell, but would be reduced to average quality. As a single brilliant piece, the paper Paige (1980) will continue to be read decades after its publication.

The paper starts by recalling the theorems of Paige (1976) quoted above (in Paige (1980) and here too the values of ε_0 and ε_1 are twice those of Paige (1976)). The matrix δR_k is bounded by

$$\|\delta R_k\|_F^2 \leq 2[2(5k-4)\varepsilon_0^2 + 4(k-1)\varepsilon_0\varepsilon_1 + k(k-1)\varepsilon_1^2]\,\|A\|^2,$$

where $\|\cdot\|_F$ denotes the Frobenius norm. If we denote $\varepsilon_2 = \sqrt{2}\max(6\varepsilon_0, \varepsilon_1)$ then

$$\|\delta R_k\|_F \leq k\varepsilon_2\|A\|.$$

The fundamental result relating the loss of orthogonality to eigenvalue convergence is given in the following theorem. We present the proof for its elegance and instructiveness.

Theorem 4.3. Let $z^j_{(k)} = (\zeta^{(k)}_{1,j}, \ldots, \zeta^{(k)}_{k,j})^T$ be the eigenvector of T_k corresponding to the Ritz value $\theta^{(k)}_j$ and $x^j_{(k)} = V_k z^j_{(k)}$ the corresponding Ritz vector, $j = 1, \ldots, k$. Let $\epsilon^{(k)}_{l,j} = (z^l_{(k)})^T \delta R_k z^j_{(k)}$. Then,

$$|\epsilon^{(k)}_{l,j}| \leq k\varepsilon_2\|A\|,$$

and

$$(x^j_{(k)})^T v^{k+1} = -\frac{\epsilon^{(k)}_{j,j}}{\eta_{k+1}|\zeta^{(k)}_{k,j}|}.$$

Proof. Multiplying the identity from Theorem 4.2 on both sides with a different eigenvector of T_k, we have

$$(z^l_{(k)})^T(T_k R_k - R_k T_k)z^j_{(k)} = \eta_{k+1}(x^l_{(k)})^T v^{k+1}\zeta^{(k)}_{k,j} + \epsilon^{(k)}_{l,j}.$$

Therefore,

$$(\theta^{(k)}_l - \theta^{(k)}_j)(z^l_{(k)})^T R_k z^j_{(k)} = \eta_{k+1}(x^l_{(k)})^T v^{k+1}\zeta^{(k)}_{k,j} + \epsilon^{(k)}_{l,j}.$$

If we take $l = j$, we obtain the result. The bound on $\epsilon^{(k)}_{l,j}$ is a consequence of the bound on the norm of δR_k. □

Hence, until $\eta_{k+1}|\zeta^{(k)}_{k,j}|$ is very small (at least proportional to $k\varepsilon_2\|A\|$), the scalar product of the Ritz vector $x^j_{(k)}$ and v^{k+1} is small.

We point out that here and elsewhere in this expository paper the actual values of the upper bounds for the quantities which are small are not at all tight for realistic problems. They do not represent indicators of the maximal attainable accuracy using the Lanczos algorithm. Most of the known worst case bound techniques inevitably produce values of the bounds which are largely oversized. But this has little effect, if any, to the value of the results obtained by the worst case rounding error analysis. Their importance and strength is in the *insight*, not in values of the bounds.

Now we come to the point. Ideally, small $\eta_{k+1}|\zeta^{(k)}_{k,j}|$ means convergence of $\theta^{(k)}_j$ to some eigenvalue λ_i of A. Numerically, however, we must take into account that $\|x^j_{(k)}\|$ can be significantly smaller than unity, and, as given in Paige (1980, relation (3.15)),

$$\min_i |\lambda_i - \theta^{(k)}_j| \leq \frac{\eta_{k+1}|\zeta^{(k)}_{k,j}|(1+\varepsilon_0) + \sqrt{k}\varepsilon_1\|A\|}{\|x^j_{(k)}\|}.$$

A bound for the accuracy of the Ritz vector is then (see Paige (1971) and also Strakoš and Greenbaum (1992, Lemma 3.4))

$$\|x^j_{(k)} - (x^j_{(k)}, q^i)q^i\| \leq \frac{\eta_{k+1}|\zeta^{(k)}_{k,j}| + \sqrt{k}\varepsilon_1\|A\|}{\min_{l\neq i}|\lambda_l - \theta^{(k)}_j|}.$$

Up to now, the analysis has been relatively simple and straightforward. This is no longer true for the remainder. In order to prove convergence of $\theta^{(k)}_j$ for $\|x^j_{(k)}\|$ significantly different from unity, Paige has ingeniously exploited properties of unreduced symmetric tridiagonal matrices. In particular, his concept of stabilized eigenvalues of T_k (see Section 2) plays the main role here. Paige has proved that if $\|x^j_{(k)}\|$ is significantly different from unity, then for some step $t < k$ there must be an eigenvalue of the left principal submatrix T_t of T_k which has stabilized to within a small δ and is close to $\theta^{(k)}_j$. This has further been used to prove that if $\eta_{k+1}|\zeta^{(k)}_{k,j}|$ is small, *i.e.*, if $\theta^{(k)}_j$ has stabilized to within a small δ, then it is always close to some eigenvalue λ_i of A, regardless the size of $\|x^j_{(k)}\|$. Consequently, although the Lanczos algorithm can produce multiple Ritz approximations of single original eigenvalues, it can never produce any 'spurious' eigenvalues, *i.e.*, Ritz values for which the convergence test $\eta_{k+1}|\zeta^{(k)}_{k,j}|$ is small and $\theta^{(k)}_j$ does not correspond to any eigenvalue λ_i of A. We summarize the result of Paige (1980, pp. 241–249) in the following theorem.

Theorem 4.4. Using the previous notation, for an eigenvalue $\theta^{(k)}_j$ of the matrix T_k computed via the Lanczos algorithm in finite precision arithmetic, we have

$$\min_i |\lambda_i - \theta^{(k)}_j| \leq \max\{2.5(\eta_{k+1}|\zeta^{(k)}_{k,j}| + \sqrt{k}\|A\|\varepsilon_1), [(k+1)^3 + \sqrt{3}n^2]\,\|A\|\varepsilon_2\}.$$

In the particular case when $\eta_{k+1}|\zeta^{(k)}_{k,j}|$ is small, the statement can be strengthened.

Theorem 4.5. If

$$\eta_{k+1}|\zeta^{(k)}_{k,j}| \leq \sqrt{3}k^2\|A\|\varepsilon_2,$$

then there exists a step $1 \leq t \leq k$ and an index $1 \leq s \leq t$ such that

$$\eta_{t+1}|\zeta_{s,t}^{(t)}| \leq \sqrt{3}t^2\|A\|\varepsilon_2 \quad \text{and} \quad \|x_{(t)}^s\| \geq \frac{1}{2},$$
$$\min_i |\lambda_i - \theta_s^{(t)}| \leq 5t^2\|A\|\varepsilon_2,$$

and $\theta_s^{(t)}, x_{(t)}^s$ is an exact eigenpair for a matrix within $5t^2\|A\|\varepsilon_2$ of A.

Please note that we are unable to prove a similar result for the given $\theta_j^{(k)}$. The difficulty is related to the possible existence of other Ritz values $\theta_l^{(k)}$ close to $\theta_j^{(k)}$. Using (2.8), Theorems 2.5, 2.7 and 4.3, Paige has proved that if $\theta_j^{(k)}$ is well separated from the other Ritz values at the same step, then $\|x_{(k)}^j\|$ cannot be significantly different from unity; see Paige (1980, (3.21), p. 243). In particular, if

$$\min_{l \neq j} |\theta_j^{(k)} - \theta_l^{(k)}| \geq k^{5/2}\|A\|\varepsilon_2,$$

we have

$$0.42 < \|x_{(k)}^j\| < 1.4.$$

Then, the result proved for $\theta_s^{(t)}$ will also hold for $\theta_j^{(k)}$. The strength of Theorem 4.4 is in the fact that the statement holds for $\theta_j^{(k)}$ no matter how many other eigenvalues of T_k are close to it.

The following theorem (see Paige (1980, Theorem 4.1)) shows that at least one eigenvalue of T_n (please recall that in exact arithmetic n represents the maximal number of steps of the Lanczos algorithm applied to A with any initial vector) must approximate some eigenvalue of A.

Theorem 4.6. If $n(3\varepsilon_0 + \varepsilon_1) \leq 1$, then at least one eigenvalue $\theta_j^{(n)}$ of T_n must be within $(n+1)^3\|A\|\varepsilon_2$ of an eigenvalue λ_i of the (n by n) matrix A. Moreover, there exist $1 \leq s \leq t \leq n$ such that

$$\eta_{t+1}|\zeta_{t,s}^{(t)}| \leq 5t^2\|A\|\varepsilon_2,$$

i.e., $\theta_s^{(t)}$ is within $5t^2\|A\|\varepsilon_2$, of λ_i.

One may question whether an analogous result can be proved for some Lanczos step $k < n$. The answer is negative, as follows from the beautiful result published in Scott (1979), which we now recall. Please note that the previous theory must hold for any initial vector v^1, $\|v^1\| = 1$. Scott's suggestion is to find, using the idea of reconstructing the unreduced symmetric tridiagonal matrix from the spectral data (for the history of this classical problem see Strakoš and Greenbaum (1992, p. 8)) a particular initial vector constructed in the following way.

Consider the diagonal matrix $A = \mathrm{diag}(\lambda_i)$. Then (remember the assumption that the eigenvalues of A are distinct) the weights in the corresponding Riemann–Stieltjes integral (2.4) are determined by $\omega_l = |(v^1, e^l)|^2$. Using (2.6) for the last step of the ideal Lanczos algorithm, the same weights are given by

$$\omega_l = |(v^1, e^l)|^2 = -\frac{\hat{\eta}}{\chi_{n-1}(\lambda_l)\chi_n'(\lambda_l)}, \quad l = 1, \ldots, n, \tag{4.1}$$

where $\hat{\eta}$ is a proper normalization constant chosen such that the constructed vector v^1 will have $\|v^1\| = 1$. Clearly, prescribing the eigenvalues of T_{n-1} (polynomial χ_{n-1}), (4.1) allows us to construct

$$v^1 \equiv (\sqrt{\omega_1}, \ldots, \sqrt{\omega_n})^T$$

such that the ideal Lanczos algorithm applied to $A = \mathrm{diag}(\lambda_i)$ with this v^1 gives T_n in the last step (and T_{n-1} in the step $n-1$). The point is that the eigenvalues T_{n-1} (Ritz values $\theta_l^{(n-1)}$) can be chosen, *e.g.*, as the midpoints of the intervals determined by the (distinct) eigenvalues of T_n (and A). Then no Ritz value $\theta_l^{(n-1)}$ at the step $n-1$ of the ideal Lanczos algorithm applied to A, with v^1 constructed as above, approximates an eigenvalue of A and no $\eta_n|\zeta_{n-1,l}^{(n-1)}|$ is small, $l = 1, \ldots, n-1$. But this means, by Theorem 2.5 (the Persistence Theorem), that no $\eta_{t+1}|\zeta_{t,s}^{(t)}|$ can be small for any choice $1 \le s \le t \le n-1$. Consequently, for this A and v^1 no Ritz value converges until step n.

A variant of the above construction of v^1 works for any given symmetric matrix A. Moreover, using a clever argument, Scott quantified the result in the following theorem; see Scott (1979, Section 4, Theorem 4.3).

Theorem 4.7. Let A be a symmetric n by n matrix with eigenvalues $\lambda_1 < \lambda_2 < \cdots < \lambda_n$, $\delta_A \equiv \min_{l \neq i} |\lambda_i - \lambda_l|$. Then there exists a starting vector v^1 such that, for the exact Lanczos algorithm applied to A with v^1, at any step $j < n$ the residual norm

$$\|Ax_{(j)} - \theta^{(j)}x_{(j)}\|$$

of any Ritz pair $\theta^{(j)}, x_{(j)}$ will be larger than $\delta_A/4$.

It should be emphasized that this result does not imply that no Ritz value can be close to an eigenvalue of A before step n. Under some lucky circumstances this can happen. Theorem 4.7 proves that such a situation cannot be revealed by the residual norm $\|Ax_{(j)} - \theta^{(j)}x_{(j)}\|$ or by the value $\eta_{j+1}|\zeta_{j,l}^{(j)}|$.

The previous results have remarkable consequences.

- First, since small $\eta_{k+1}|\zeta_{k,j}^{(k)}|$ means convergence of $\theta_j^{(k)}$ to some eigenvalue λ_i of A, Theorem 4.3 may be restated:

  Orthogonality can be lost only in the directions of converged Ritz vectors.

  In contrast to this, we do not have a *proof* that convergence of a Ritz value is necessarily accompanied by the loss of orthogonality of v^{k+1} in the direction of the corresponding Ritz vector, since $\epsilon_{j,j}^{(k)}$ in the numerator in the statement of Theorem 4.3 can vanish. We have, however, not seen an example of such behaviour.

- Second, in the example constructed by Scott there is ideally no convergence of Ritz values until the final step. If this also remains true numerically, then for the particular A, v^1 given by Scott there is no significant loss of orthogonality among the computed Lanczos vectors $v^1, \dots, v^n$! This means that loss of orthogonality in the finite precision Lanczos algorithm significantly depends for a given A on the choice of v^1. It should be admitted, though, that the particular initial vectors for which the loss of orthogonality is suppressed typically have rather weird components varying by many orders of magnitude. Interested readers can check the validity of the above statements and the illustrative properties of the initial vectors suggested by Scott by numerical experiments.

An argument derived from the investigation of the accuracy of $\theta_j^{(k)}$ as the Rayleigh quotient shows (Paige 1980, (3.48))

$$\lambda_{\min}(A) - k^{\frac{5}{2}}\|A\|\varepsilon_2 \leq \theta_j^{(k)} \leq \lambda_{\max}(A) + k^{\frac{5}{2}}\|A\|\varepsilon_2,$$

which is true whether or not $\theta_j^{(k)}$ has stabilized to within a small δ.

We have seen that until some Ritz value stabilizes to within small δ, the orthogonality of numerically computed Lanczos vectors cannot be lost. This poses the question as to how closely the Lanczos algorithm in finite precision arithmetic can resemble the ideal one. Paige gives an elegant answer in terms of the backward error. In fact, if at step k

$$\eta_{l+1}|\zeta_{l,j}^{(l)}| \geq \sqrt{3}k^2\|A\|\varepsilon_2, \quad 1 \leq j \leq l \leq k, \tag{4.2}$$

then $\|R_k\|_F^2 < 1/12$ and all singular values of V_k lie in the open interval $(0.41, 1.6)$: see Paige (1980, p. 250). If the Lanczos algorithm is applied with full reorthogonalization at every step, implemented via the modified Gram–Schmidt algorithm, then under a mild restriction the computed columns V_k span the exact Krylov subspace of $A + \delta A$ (starting with the same v^1), where $\|\delta A\|$ is a multiple of $\|A\|\varepsilon_M$ (Paige 1970*a*). The following theorem (see Paige (1980, Theorem 4.2)) completes the argument.

Theorem 4.8. Using the previous notation, let (4.2) hold at step k of the finite precision Lanczos algorithm applied to A with v^1. Then there exists a matrix

$$A'(k) \quad \text{within} \quad (3k)^{1/2}\|A\|\varepsilon_2 \quad \text{of} \quad A$$

such that, for all $l = 1, \ldots, k+1$, the Lanczos vectors $v^1, \ldots, v^l$ span the Krylov subspaces of $A'(k)$ with the initial vector v^1.

Consequently, until the computed Krylov subspace contains an exact eigenvector of a matrix to within $5k^2\|A\|\varepsilon_2$ of the original matrix A (see Theorem 4.5), this subspace is the same as the Krylov subspace generated by a slightly perturbed matrix, *i.e.*, it is numerically stable in the backward error sense. It should, however, be noted that generally Krylov subspaces can be very sensitive to small changes in the matrix A.

We conclude the journey through the fascinating paper of Paige (1980) with the following comment. Until a Ritz value in steps 1 to k has stabilized to within $\sqrt{3}k^2\|A\|\varepsilon_2$, the Lanczos algorithm behaves numerically like the algorithm with full modified Gram–Schmidt reorthogonalization.

4.3. Backward-like analysis of Greenbaum and subsequent results

Consider a fixed step k of the finite precision Lanczos algorithm applied to A and v^1. We ask whether the results computed in steps 1 to k can be interpreted in some sense as results of the ideal Lanczos algorithm applied to some matrix B with an initial vector v_B^1. Indeed, as we have seen in Section 2, the numerically computed matrix T_k storing the Lanczos recurrence coefficients can be obtained in k steps of the k-dimensional ideal Lanczos algorithm applied to T_k with the initial vector e^1. Components of e^1 in the basis of the (orthonormal) eigenvectors $Z_k = (z_{(k)}^1, \ldots, z_{(k)}^k)$ of T_k are equal to the elements of the first row of the matrix Z_k; their squares representing weights in the corresponding k-dimensional Riemann–Stieltjes integral: see (2.6). Consequently, the matrix T_k can be obtained as a result of the exact Lanczos algorithm applied to any k by k matrix B having the same eigenvalues as T_k with the initial vector v_B^1 having the components in the corresponding eigenspaces of B equal to the elements of $Z_k^T e^1$.

This relationship, although interesting, does not tell us much, since T_k (or B) can have some eigenvalues close to the eigenvalues of A, but others can be very different from the eigenvalues of A. As we have seen, the finite precision Lanczos algorithm may form multiple copies of several eigenvalues of A, with the multiplicities growing as the number of iteration steps increases. But the algorithm will never give a Ritz value stabilized to within a small δ that does not approximate any eigenvalue of A. It therefore seems necessary to impose additional conditions on B and v_B^1.

Given T_k computed in k steps of the finite precision Lanczos algorithm applied to A with v^1, we look for B and v_B^1 such that *all* eigenvalues of B lie close to the eigenvalues of A. In addition to that, the sum of squares of the components of v_B^1 in the invariant subspaces corresponding to close approximations of some eigenvalue λ_i of A is required to be equal to the squared component of v^1 in the direction of the original eigenvector q^i. Finally, the point is that we require T_k to be determined in the first k steps of the *exact* Lanczos algorithm applied to B with v_B^1.

When Anne Greenbaum developed her highly original and deeply thought theory on the foundations laid by Paige, she supported the previous intuitive argument with a rigorous mathematical derivation. She showed that the exact Lanczos recurrence for a matrix whose eigenvalues are clustered in small intervals can be thought of as a slightly perturbed recurrence, analogous to that of Theorem 4.1, for a *new problem.* This new problem has, for each original cluster interval, one eigenvalue from this interval representing the whole cluster. The sum of the weights of the original eigenvalues in each cluster is equal to the weight of its chosen representing eigenvalue: see Greenbaum (1989). From that she set the goal of proving that *every* slightly perturbed Lanczos recurrence, including the finite precision Lanczos algorithm described in Theorem 4.1, is in the sense described above equivalent to an exact Lanczos recurrence for a matrix whose eigenvalues lie in small intervals about the eigenvalues of the given original matrix.

While the details of the theorems and proofs of Greenbaum (1989) are quite involved and have probably not been read carefully by many people, the basic ideas are ingenious, and the paper should be considered obligatory classical reading together with Paige (1980). We will try to recall the main points in order to reveal, within our abilities, the beauty of the construction given by Greenbaum.

To show that the matrix T_k generated at step k of the finite precision Lanczos recurrence applied to A with v^1 is the same as that given by the exact Lanczos algorithm applied to some B with v_B^1, where all eigenvalues of B are close to those of A, it is sufficient and also necessary to show that T_k can be extended to a larger unreduced symmetric tridiagonal matrix (having positive subdiagonal entries)

$$T_{k+K} = \begin{pmatrix} T_k & \eta_{k+1} & & & & & \\ \eta_{k+1} & \alpha_{k+1} & \eta_{k+2} & & & & \\ & \eta_{k+2} & \alpha_{k+2} & \eta_{k+3} & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & \eta_{k+K-1} & \alpha_{k+K-1} & \eta_{k+K} \\ & & & & \eta_{k+K} & \alpha_{k+K} \end{pmatrix}$$

whose eigenvalues are all close to those of A. Then we can simply take $B \equiv T_{k+K}$, $v_B^1 \equiv e^1$.

Greenbaum has constructed T_{k+K} by a hypothetical continuation of the first k steps of the finite precision Lanczos algorithm applied to A with v^1. The needed situation $\eta_{k+K+1} = 0$ for some K can be reached in the following way. Based on the theory of Paige describing the loss of orthogonality among the Lanczos and Ritz vectors, Greenbaum has identified a set of $k - m_k$ vectors in the subspace generated by the computed $\{v^1, \dots, v^k\}$ such that the chosen vectors are mutually (exactly) orthogonal and normalized, and the newly computed v^{k+1} is also approximately orthogonal to all of them. Let these vectors be stored as the columns of the n by $(k - m_k)$ matrix Y_{k-m_k}. Exact orthogonalization of v^{k+1} against them adds a small additional contribution into the error term. Then the Lanczos recurrence can hypothetically be continued with the exact orthogonalization of the newly generated Lanczos vectors against each other and with exact orthogonalization of them against Y_{k-m_k}. From the exact orthogonalization we must get $\eta_{k+K+1} = 0$ since $(Y_{k-m_k}, v^{k+1}, \dots, v^{k+K})$, where $K = n + m_k - k$, represents a set of n orthogonal vectors in the n-dimensional space. Summarizing, we get

$$AV_{k+K} = V_{k+K}T_{k+K} + F_{k+K},$$

where in $F_{k+K} = \big(f^1, \dots, f^{k-1}, f^k, \dots, f^{k+K}\big)$ the first $k-1$ columns are the perturbations in the steps of the original finite precision Lanczos algorithm and the other columns $f^k, \dots, f^{k+K}$ are perturbations arising from reorthogonalizations in Greenbaum's construction. The way this is done cannot be described here since it involves many details which cannot be included in this expository paper. The key point is in the choice of the orthonormal vectors Y_{k-m_k}; they cannot contain, *e.g.*, any vector in the subspace of the converged Ritz vectors corresponding to well-separated Ritz values, since these represent well-defined directions in which the orthogonality is definitely lost. More substantively, the clever choice of Y_{k-m_k} described in Greenbaum (1989) allows her to prove that the perturbation vectors $f^k, \dots, f^{k+K}$, introduced in the hypothetical continuation of the finite precision Lanczos algorithm, are *small*. Paige's results summarized in Theorem 4.4 can then be applied to the $k + K = n + m_k$ steps of the hypothetically extended finite precision Lanczos recurrence described above with $\eta_{n+m_k+1} = 0$, where from the proofs in Paige (1980) it follows that the size of the errors corresponding to the perturbations $f^k, \dots, f^{k+K}$ can be expressed in term of their norms.

Theorem 4.9. The matrix T_k generated at step k of the finite precision Lanczos algorithm applied to A with v^1 is equal to that generated by an exact Lanczos recurrence applied to an $(n + m_k)$ by $(n + m_k)$ matrix B whose eigenvalues lie within

$$\mathcal{O}(n + m_k)^3 \max\{\varepsilon_M \|A\|, \|f^k\|, \dots, \|f^{n+m_k}\|\}$$

of some of the eigenvalues of A, where $f^k, \dots, f^{n+m_k}$ are the smallest perturbations that will cause a coefficient η_{j+1} to be zero at or before step $n + m_k$.

The particular $f^k, \dots, f^{n+m_k}$ given via the construction of Greenbaum (1989) are perhaps not the optimal ones, but they are small enough to justify this approach. Finally, Theorem 4.2 of Strakoš (1991) proves the intuitively expected fact that any matrix B with the property of Theorem 4.9 must have at least one eigenvalue close to each eigenvalue of the original matrix A for which the component of the initial vector v^1 in the corresponding invariant subspace is nonzero.

We remark that the value m_k and the matrix B depend on k. The matrix B with the required property described above is not unique; there might be other constructions giving similar results with matrices of different sizes. If we limit the number of steps in the application of the Lanczos algorithm to some reasonable number N, say, much smaller than $(n\varepsilon_M \|A\|)^{-1}$, then it is legitimate to ask whether one can take a matrix B with v_B^1 such that the exact Lanczos algorithm applied to B with v_B^1 will give in steps 1 to N not necessarily identical, but *very close* Ritz values, to those provided by the finite precision Lanczos algorithm applied to A, v^1. Here we do not mean determining B (and v_B^1) *a posteriori* for the step N, but *a priori* using the spectral decomposition of A and the components of v^1 in the individual invariant subspaces. This idea was thoroughly illustrated in Greenbaum and Strakoš (1992), where B was constructed by spreading sufficiently many eigenvalues in tiny intervals around each eigenvalue of A. Numerical experiments suggest that the size of such intervals is much smaller than the technically complicated bounds from Greenbaum (1989) would suggest. A rigorous mathematical quantification of this approach is still incomplete. When completed, it would also lead to a possibly very elegant matrix-free description of the Lanczos algorithm behaviour in finite precision arithmetic in terms of the Gauss quadratures of a Riemann–Stieltjes integral with a slightly blurred distribution function (see Section 5 of Golub and Strakoš (1994), Section 4.5 of Greenbaum (1997*a*), and Section 5 of Strakoš and Tichý (2002)). This must, however, include a sensitivity analysis of Gauss quadrature to small perturbations of the Riemann–Stieltjes integral, which appears to be a rather difficult problem (O'Leary and Strakoš 2004). A different but somewhat related problem concerning sensitivity of the Lanczos coefficients to perturbations of the distribution function in the Riemann–Stieltjes integral is investigated in Kautsky and Golub (1983); see also Gragg and Harrod (1984), Laurie (1999, 2001) and Druskin, Borcea and Knizhnerman (2005).

A frequently asked question is whether the finite precision Lanczos algorithm can simply miss an eigenvalue because it is constantly forming copies of others. This is known as the Lanczos phenomenon (see Cullum and Willoughby (1985)) and it can be considered resolved by the series of works by Druskin and Knizhnerman (Druskin and Knizhnerman 1991, Knizhnerman 1995*a*, 1995*b*, 1996); see also Druskin, Greenbaum and Knizhnerman (1998) and Greenbaum (1994). Using some technical assumptions, it is proved that each eigenvalue of A will indeed eventually be approximated by a Ritz value. The proven statement is, however, more of theoretical than practical interest. A considerable part of these papers is also devoted to approximation of matrix functions.

Existence of tight clusters of Ritz values is linked to most of the technical difficulties that complicate the bounds and proofs of Paige (1980) and Greenbaum (1989). We know that a Ritz value can stabilize to within a small δ only close to an eigenvalue of A. If the stabilized Ritz value is well separated, then the norm of the Ritz vector cannot significantly differ from unity, and the Ritz vector closely approximates the corresponding eigenvector of A. When a Ritz value is a part of a tight cluster, then some or *all* Ritz pairs corresponding to the cluster can have weird properties.

In Strakoš and Greenbaum (1992) several conjectures have been formulated, but not proved (except for some simple cases). In particular, it is important to ask the following questions.

C1 (Stabilization of clusters.) Does any tight well-separated cluster consisting of at least two Ritz values approximate an eigenvalue of A?

C2 (Stabilization of Ritz values in a cluster.) Is any Ritz value in a tight well-separated cluster stabilized to within a small δ? In particular, Strakoš and Greenbaum (1992) conjectured that the answer is positive, and that δ is proportional to the square root of the size of the cluster interval divided by the square root of the separation of the cluster from the other Ritz values.

C3 (Stabilization of weights.) Let Ritz values in a tight well-separated cluster, which may consist of one or more Ritz values, closely approximate some eigenvalue λ_i of A. Does the sum of weights of these Ritz values in the corresponding Riemann–Stieltjes integral closely approximate the weight of the original eigenvalue λ_i?

Similar questions can be formulated solely in terms of unreduced symmetric tridiagonal matrices, and they are therefore not specific to the finite precision Lanczos algorithm. In the latter case they are, however, of particular importance. We will not specify the intuitive meaning of the terms ‘tight cluster’, ‘size of the cluster interval’ and ‘separation of the cluster’ since that would need detailed notation which we cannot afford, because of lack of space. The intuitive meaning is clear; a technical quantification can be found in the papers by Wülling, which we are now going to recall.

The conjectures were investigated in Wülling (2005) and (2006) with the following outcome.

- Every tight well-separated cluster of at least two Ritz values must stabilize, *i.e.*, the answer to **C1** is positive.
- There are tight well-separated clusters of Ritz values (which, according to the previous point, must approximate an eigenvalue of A) in which none of the Ritz values is stabilized to within a small δ, *i.e.*, the answer to **C2** is negative.
- The weights in the Riemann–Stieltjes integral corresponding to the kth Gauss quadrature of the original Riemann–Stieltjes integral determined by A and v^1 must stabilize, *i.e.*, the answer to **C3** is positive. This is not proved directly in Wülling (2005), but it can be obtained by a combination of Wülling (2005) with the inequality (8.21) in Greenbaum (1989): see Wülling (2005, Section 5).

In contrast with Strakoš and Greenbaum (1992), where the results are based on relatively simple algebraic manipulations of the known formulas for eigenvalues and eigenvector elements of unreduced symmetric tridiagonal matrices, Wülling (2005) and (2006) are based on the following very clever observation. The bottom and top elements of the eigenvectors of T_k, which determine the stabilization criterion and the weights respectively, are expressed in terms of the values of polynomials $\chi_{k-1}(\theta)$ and $\chi_k'(\theta)$: see (2.6)–(2.8). Moreover, $\chi_{k-1}(\theta)$ and $\chi_k(\theta)$ have simple roots in the corresponding Ritz values. Therefore, using the residue theorem from complex analysis, the sum of squares of the bottom elements of the (normalized) eigenvectors of T_k, which correspond to the Ritz values in a cluster C, can be viewed as the result of the line integral

$$\sum_C (\zeta_{k,l}^{(k)})^2 = -\sum_C \frac{\chi_{k-1}(\theta_l^{(k)})}{\chi_k'(\theta_l^{(k)})} = \frac{1}{2\pi}\left|\int_{\partial D_C} \frac{\chi_{k-1}(z)}{\chi_k(z)}\,\mathrm{d}z\right|, \tag{4.3}$$

where ∂D_C is the circle which contains all Ritz values belonging to C in its interior and all other eigenvalues of T_k in its exterior: see Wülling (2006). Similarly, omitting technicalities, the changes in the weights can be investigated using the line integral

$$\frac{1}{2\pi}\left|\int_{\partial D_C} \frac{\eta_2^2\eta_3^2\ldots\eta_k^2}{\chi_{k-1}(z)\chi_k(z)}\,\mathrm{d}z\right|; \tag{4.4}$$

see Wülling (2005, (4.5)). The results are then obtained by bounding the line integrals (4.3) and (4.4), which represent an example of nontrivial technical work. We also point out that, concerning **C1** and **C2**, the results of Wülling (2006) are stronger than the formulations of the conjectures in Strakoš and Greenbaum (1992) have assumed.

The analysis of Wülling gives another example of interplay between analysis (here complex analysis, which is used to obtain bounds for algebraic expressions formulated in terms of values of orthogonal polynomials) and algebra, often observable while dealing with the Lanczos algorithm.

4.4. Intermediate quantities and the accuracy of Ritz approximations

As we have already seen, the finite precision Lanczos algorithm serves as an instructive example illustrating several fundamental principles. Its rounding error analysis is perhaps complicated, lengthy and full of unpleasant technical details, bounds and formulas. However, it reveals the pattern rigorously, and the conclusions can be formulated clearly, simply and in an elegant way.

In addition, the whole rounding error analysis reveals the following principal fact of 'philosophical' importance. The Ritz values as approximations to eigenvalues of the original matrix A can be computed to high accuracy despite the fact that the intermediate quantities, *i.e.*, the computed Lanczos coefficients stored in the matrix T_k, $k = 1, 2, \ldots$, can have from some (typically rather modest) value of k not a single digit of accuracy. In other words, the number of correct digits in the computed entries of T_k (in comparison with their ideal counterparts) is absolutely irrelevant for the obtainable accuracy of the approximations to the eigenvalues of A determined from T_k. Here we see the power of the backward-like analysis (*cf.* Parlett (1990, pp. 22 and 24)), and the limitations of the mechanically applied forward error analysis, when it considers comparison of *all* computed and ideal quantities.

4.5. Reorthogonalization strategies and rewards for maintaining semiorthogonality

Although the inaccuracy of T_k does not prevent accurate approximation of eigenvalues of A by Ritz values, it has rather unpleasant effects: multiple approximations of some eigenvalues of A, and delays in the approximation of another ones. The way to suppress these side effects, which is sometimes desirable, is to apply a correction procedure which preserves maximally, or to some suitable level, the mutual orthogonality of the computed Lanczos vectors. Reorthogonalization strategies and the rewards for maintaining a proper level of mutual orthogonality are thoroughly described in Scott (1978), Parlett and Scott (1979), Scott (1981), Simon (1982, 1984*a*) and Parlett (1994), and excellently summarized in Simon (1984*b*), Parlett (1992). Here we will briefly recall some main ideas. An extended exposition can be found in the last two papers.

We start with the PhD thesis of Grcar (1981), which, to our knowledge, was not published. In contrast to other researchers, his considerations are based on the forward error of the computed Lanczos vectors. Grcar's results suggest, though the formal proofs have not been completed, that until the above-mentioned forward error exceeds the level proportional to $\sqrt{\varepsilon_M}$, the computed Krylov subspace is correct to the level proportional to ε_M (the error stays largely within the ideal Krylov subspace). In order to maintain this so-called projection property, Grcar suggested periodic reorthogonalization. The forward approach of Grcar (1981) has to deal with some theoretical and practical difficulties. The way Grcar uses nonhomogeneous three-term recurrences inspired later solutions of other problems: see Gutknecht and Strakoš (2000) and Meurant (2006).

Beresford Parlett and his PhD students played the instrumental role in the other reorthogonalization strategies, which have been conveniently based on Paige's results and backward error analysis. It was discovered that, in order to largely suppress the unpleasant effects of round-off on the approximation of the eigenvalues of A, full reorthogonalization of the Lanczos vectors (in order to maintain their mutual orthogonality close to ε_M) is not necessary. It suffices to maintain some 'strong linear independence' of the computed Lanczos vectors. Scott has shown (see Parlett and Scott (1979), Scott (1978, 1981)) that it is beneficial to maintain *semi-orthogonality* of the numerically computed Lanczos vectors, *i.e.*, to satisfy

$$\|V_k^T v^{k+1}\| \leq \sqrt{\varepsilon_M}, \quad k = 1, 2, \dots. \tag{4.5}$$

Since Theorem 4.3 proved by Paige shows that orthogonality can be lost only in the direction of converged Ritz vectors, one suggestion is to maintain semi-orthogonality by reorthogonalizing at each step k the newly computed Lanczos vector against all Ritz vectors for which

$$\eta_{k+1}|\zeta_{k,l}^{(k)}| < k\sqrt{\varepsilon_M}\|A\|;$$

cf. Simon (1984*b*, Theorem 6, p. 126). This strategy, called *selective reorthogonalization* (SO) requires computing Ritz vectors. That is avoided in the *partial reorthogonalization strategy* (PRO) of Simon. Based on the underlying rigorous analysis of Paige, Simon has suggested and justified a simplified model of finite precision behaviour of the Lanczos algorithm. His strategy is based on monitoring the loss of orthogonality among the Lanczos vectors via a three-term recurrence: see Simon (1984*b*, Theorem 1, p. 107). It reorthogonalizes the newly computed Lanczos vector at step k against those previously computed Lanczos vectors related through some heuristic to the threshold criteria for the loss of orthogonality proportional to $\sqrt{\varepsilon_M}/k$. Simon then proved the following theorem.

Theorem 4.10. Let T_k be the unreduced symmetric tridiagonal matrix computed by the Lanczos algorithm applied to A with v^1 that uses some reorthogonalization in order to maintain semi-orthogonality among the computed Lanczos vectors. Then, up to a (full) perturbation matrix having norm proportional to $\varepsilon\|A\|$, T_k is the orthogonal projection of A onto the subspace spanned by the computed Lanczos vectors.

This means (see also Simon (1984*b*, pp. 119–122), Parlett (1992, pp. 255–257)) that in the above sense semi-orthogonality is as good as orthogonality maintained proportional to full machine precision. Finally, Theorem 4.4 of Parlett (1992) proves that an additional full reorthonalization at a step k guarantees an improvement of the mutual orthogonality only if semi-orthogonality is maintained in steps 1 to k.

As mentioned above, in our exposition we assume that the exact spectral decomposition of the unreduced symmetric tridiagonal matrix T_k is known. Here such an assumption is reasonable, since an investigation of further issues related to computing this spectral decomposition is out of the scope of this review. Nevertheless, since T_k can have tight clusters of eigenvalues, we wish at least to point out several publications devoted to interesting issues arising from this problem; see Ye (1995), Parlett (1996), Parlett and Dhillon (2000) and Dhillon and Parlett (2003, 2004). .

4.6. Recent results on the loss of orthogonality and multiple approximation of eigenvalues

As we have said before, an attempt at forward error analysis of the Lanczos algorithm was given by Grcar (1981). Grcar obtained expressions for the computed Lanczos vectors in terms of the exact Lanczos vectors. In this section, we will summarize works (Zemke 2003, Meurant 2006) interested in the components of the Lanczos vectors in the directions of the eigenvectors of A. The goal of these works is to understand the behaviour of the projections of the Lanczos vectors, their relation to the loss of orthogonality, and the appearance of multiple copies of the eigenvalues. This problem leads to investigation of perturbed three-term scalar recurrences. There are different ways to write the solution of these recurrences (for instance, using polynomials or using the Lanczos matrix T_k). They show what equation (2.2), giving Lanczos vectors as polynomials in A applied to the initial vector, becomes in finite precision arithmetic.

Let us start by considering the D30 example. We look at components of the Lanczos vectors in the directions of the eigenvectors of A. Since the matrix D30 is diagonal, we simply consider the components of the Lanczos vectors. The initial vector has all its components equal. The eigenvalue which is first approximated by a Ritz value is the largest one, $\lambda_{30} = 100$. In Figure 4.2 the solid line is $\log_{10}(|v_{30}^k|)$ as a function of k, computed by the Lanczos algorithm using full reorthogonalization of the newly computed

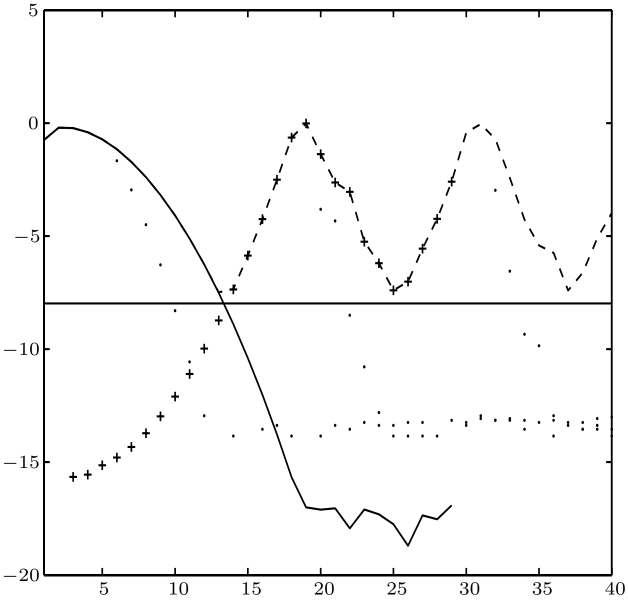

Figure 4.2. D30, $\log_{10}$ of the absolute value of the last component of the Lanczos vectors.

Lanczos vector against the previously computed Lanczos vectors, with the reorthogonalization done twice (which we call double reorthogonalization). Before the component of interest v^k_{30} reaches the square root of machine precision, the computed results of this example can be considered close approximations to the exact precision ones. As predicted by theory, the last component of the Lanczos vector (with double reorthogonalization) converges to machine precision. The dashed line (which is hidden behind the solid line until it is nearly at the horizontal line) is $\log_{10}(|\tilde{v}^k_{30}|)$ computed by the standard finite precision Lanczos algorithm. The + signs represent $\log_{10}$ of the absolute value of the differences $v^k_{30} - \tilde{v}^k_{30}$. The horizontal line is $\log_{10}(\sqrt{\varepsilon_M})$. The dots give the distances of the Ritz values to λ_{30} after they become smaller than a threshold of 0.1.

The computed component is almost equal to the ideal result down to $\sqrt{\varepsilon_M}$ but then, instead of continuing to go down, it starts going back up to $\mathcal{O}(1)$. The difference is increasing almost from the beginning of the iterations up to iteration 18. After that there is an almost periodic behaviour. Each time the last component reaches $\mathcal{O}(1)$, a new copy of the largest eigenvalue appears. This simple example shows there is an interesting structure in the components of the Lanczos vectors in the directions of the eigenvectors of A. Similar pictures with different examples are given and analysed in Zemke (2003, pp. 210–217).

In exact arithmetic we have the relation

$$\eta_{k+1}v^{k+1} = Av^k - \alpha_k v^k - \eta_k v^{k-1}.$$

In finite precision computations, this relation becomes

$$\tilde{\eta}_{k+1}\tilde{v}^{k+1} = A\tilde{v}^k - \tilde{\alpha}_k\tilde{v}^k - \tilde{\eta}_k\tilde{v}^{k-1} + f^k, \tag{4.6}$$

where f^k represents the rounding errors that occurred while computing step $k+1$. Of course, the coefficients $\tilde{\alpha}_k$ and $\tilde{\eta}_k$ are different from those in exact arithmetic since they are determined (numerically) using the computed Lanczos vectors. This is what makes a forward analysis of the finite precision Lanczos algorithm difficult. Let $\bar{v}^k = Q^T\tilde{v}^k$ be the vector of the projections of the computed Lanczos vector on the eigenvectors of A. We have

$$\tilde{\eta}_{k+1}\bar{v}_i^{k+1} = \lambda_i\bar{v}_i^k - \tilde{\alpha}_k\bar{v}_i^k - \tilde{\eta}_k\bar{v}_i^{k-1} + \bar{f}_i^k, \tag{4.7}$$

where $\bar{f}^k = Q^T f^k$. Solutions of such three-term recurrences are studied in Meurant (2006) where the following result is proved.

Theorem 4.11. Let j be given and let $p_{j,k}$ be the polynomials determined by

$$p_{j,j-1}(\lambda) \equiv 0, \qquad p_{j,j}(\lambda) \equiv 1,$$
$$\zeta_{k+1}p_{j,k+1}(\lambda) = (\lambda - \tau_k)p_{j,k}(\lambda) - \zeta_k p_{j,k-1}(\lambda), \; k = j, j+1, \ldots .$$

The solution of the perturbed scalar recurrence

$$\zeta_{k+1}s_{k+1} = (\lambda - \tau_k)s_k - \zeta_k s_{k-1} + f_k, \tag{4.8}$$

starting from $s_0 = 0$ and s_1 is given by

$$s_{k+1} = p_{1,k+1}(\lambda)s_1 + \sum_{l=1}^{k} p_{l+1,k+1}(\lambda)\frac{f_l}{\zeta_{l+1}}.$$

The polynomials $p_{j,k}$, $j > 1$ are usually called the associated polynomials. They are orthogonal with respect to a Riemann–Stieltjes integral with a distribution function that depends on j. When applying this to the Lanczos algorithm, we use the following result.

Lemma 4.12. The polynomial $p_{j,k}$, $k \geq j$ is given by

$$p_{j,k}(\lambda) = (-1)^{k-j}\frac{\chi_{j,k-1}(\lambda)}{\tilde{\eta}_{j+1}\cdots\tilde{\eta}_k},$$

where $\chi_{j,k}(\lambda)$ is the determinant of $\tilde{T}_{j,k} - \lambda I$, where $\tilde{T}_{j,k}$ is the tridiagonal matrix obtained from the computed Lanczos matrix $\tilde{T}_k$ by deleting the first $j-1$ rows and columns.

The possible growth of the local round-off perturbations is therefore linked to the eigenvalues of the matrices $\tilde{T}_{j,k}$ for all $j \leq k$. A similar technique has also been used by Gutknecht and Strakoš (2000) in the investigation of the maximal attainable accuracy.

Applying these results to the finite precision Lanczos algorithm, that is, to (4.7), we obtain the following result.

Theorem 4.13. Let j be given and $\tilde{p}_{j,k}$ be the polynomials given by

$$\tilde{p}_{j,j-1}(\lambda) \equiv 0, \qquad \tilde{p}_{j,j}(\lambda) \equiv 1,$$
$$\tilde{\eta}_{k+1}\tilde{p}_{j,k+1}(\lambda) = (\lambda - \tilde{\alpha}_k)\tilde{p}_{j,k}(\lambda) - \tilde{\eta}_k\tilde{p}_{j,k-1}(\lambda), \; k = j, j+1, \dots .$$

Then, the computed Lanczos vector at iteration $k+1$ is

$$\tilde{v}^{k+1} = \tilde{p}_{1,k+1}(A)v^1 + \sum_{l=1}^{k} \tilde{p}_{l+1,k+1}(A)\frac{f^l}{\tilde{\eta}_{l+1}}. \tag{4.9}$$

This is to be compared with (2.2) which gives the result in exact arithmetic. We note that the first term $\tilde{p}_{1,k+1}(A)v^1$ is different from what we have in exact arithmetic since the coefficients of the recurrence are different. If we want to pursue the forward analysis and consider the difference between ideal and computed Lanczos vectors, we have to link $\tilde{v}^{k+1}$ to v^{k+1}. Looking at the three-term recurrences for the ideal and computed polynomials we have

$$\eta_{k+1}p_{1,k+1}(\lambda) = (\lambda - \alpha_k)p_{1,k}(\lambda) - \eta_k p_{1,k-1}(\lambda),$$

and

$$\tilde{\eta}_{k+1}\tilde{p}_{1,k+1}(\lambda) = (\lambda - \tilde{\alpha}_k)\tilde{p}_{1,k}(\lambda) - \tilde{\eta}_k\tilde{p}_{1,k-1}(\lambda).$$

Setting $\Delta p_k(\lambda) = p_{1,k}(\lambda) - \tilde{p}_{1,k}(\lambda)$, this difference satisfies a three-term recurrence relation,

$$\tilde{\eta}_{k+1}\Delta p_{k+1}(\lambda) = (\lambda - \tilde{\alpha}_k)\Delta p_k(\lambda) - \tilde{\eta}_k\Delta p_{k-1}(\lambda) + g_k(\lambda), \tag{4.10}$$

with

$$g_k(\lambda) = (\tilde{\eta}_{k+1} - \eta_{k+1})p_{1,k+1}(\lambda) + (\tilde{\alpha}_k - \alpha_k)p_{1,k}(\lambda) + (\tilde{\eta}_k - \eta_k)p_{1,k-1}(\lambda).$$

From Theorem 4.11 we can obtain the solution of (4.10) and then derive an expression for the difference between the ideal and computed Lanczos vectors: see Meurant (2006).

Theorem 4.14. As long as $k < n$,

$$\tilde{v}^{k+1} = v^{k+1} + \sum_{l=1}^{k} \tilde{p}_{l+1,k+1}(A)g_l(A)\frac{v^1}{\tilde{\eta}_{l+1}} + \sum_{l=1}^{k} \tilde{p}_{l+1,k+1}(A)\frac{f^l}{\tilde{\eta}_{l+1}}. \tag{4.11}$$

Theorem 4.14 shows that the difference between the ideal and the computed Lanczos vectors arises from two sources: the local rounding errors f^l and the differences of the coefficients (which, of course, come from the differences of the previous Lanczos vectors). From the D30 example, we have seen that it is interesting to consider the behaviour of $(\bar{v}^{k+1})_i = (Q^T\tilde{v}^{k+1})_i$. This is given by

$$(Q^T\tilde{v}^{k+1})_i = \tilde{p}_{1,k+1}(\lambda_i)(Q^Tv^1)_i + \sum_{l=1}^{k} \tilde{p}_{l+1,k+1}(\lambda_i)\frac{(Q^Tf^l)_i}{\tilde{\eta}_{l+1}}. \tag{4.12}$$

It is difficult to study the behaviour of the sum in (4.12). This shows again the limitations of a forward analysis. However, in order to get some insight, one can look at each term individually.

What can be shown is the fact that, for a given λ_i towards which a Ritz value is converging, the absolute value of the polynomials $|\tilde{p}_{1,k}(\lambda_i)|$, as a function of k, first decreases to the level $\sqrt{\varepsilon_M}$, and then increases back to $\mathcal{O}(1)$. The values $|\tilde{p}_{j,k}(\lambda_i)|$ for $j > 1$ increase as a function of k up to a maximum of $\mathcal{O}(1)$, and then decrease down to $\sqrt{\varepsilon_M}$. This can be proved rigorously for the beginning of the process until the first Ritz value has converged and $|\tilde{p}_{1,k}(\lambda_i)|$ is back to $\mathcal{O}(1)$. This is done by investigating the product $|\tilde{p}_{1,k}(\lambda)\tilde{p}_{j,k}(\lambda)|$ for $k > j > 1$: see Meurant (2006).

The approach using polynomials offers some insight into the numerical behaviour of the Lanczos algorithm. In the beginning, the growth of the individual terms in the sum representing the influence of the round-off on the components of the Lanczos vectors in the directions of the eigenvectors of A goes hand in hand with the decrease of the original component. But, the argument is incomplete since we cannot analyse the whole sums defining a component of $\bar{v}^k$.

One can also consider other ways to write the solution of a three-term nonhomogeneous recurrence: see Meurant (2006). We consider once again the recurrence (4.8) with s_1 given and $\zeta_2 s_2 = (\lambda - \tau_1)s_1 + f_1$. For simplicity we take $\lambda = 0$ and let

$$L_{k+1} = \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 & 0 \\ \tau_1 & \zeta_2 & 0 & \cdots & 0 & 0 \\ \zeta_2 & \tau_2 & \zeta_3 & & \vdots & \vdots \\ & \ddots & \ddots & \ddots & 0 & \vdots \\ & & \zeta_{k-1} & \tau_{k-1} & \zeta_k & 0 \\ & & & \zeta_k & \tau_k & \zeta_{k+1} \end{pmatrix}.$$

This matrix is written as

$$L_{k+1} = \begin{pmatrix} (e^1)^T & 0 \\ T_k & \zeta_{k+1}e^k \end{pmatrix},$$

where T_k is the tridiagonal matrix of the recurrence coefficients. Let $s^{k+1} = \big(s_1, \dots, s_{k+1}\big)^T$ and $g = s_1$, $h = \big(f_1, \dots, f_k\big)^T$ then the non homogeneous recurrence (4.8) can be written as

$$L_{k+1} s^{k+1} = \begin{pmatrix} g \\ h \end{pmatrix}.$$

In the following we shall use this for the Lanczos algorithm with $\tilde{T}_k - \lambda_i I$ instead of T_k. To obtain the solution of the recurrence, the first step is to find an expression for the inverse of L_{k+1} involving T_k. This is given in the next theorem in which we only give the entries we are interested in, and with the proof left to Meurant (2006).

Theorem 4.15.

$$(L_{k+1}^{-1})_{(1:k,1)} = \frac{1}{(T_k^{-1})_{1,k}} T_k^{-1} e^k,$$
$$(L_{k+1}^{-1})_{(1:k,2:k+1)} = T_k^{-1} - \frac{1}{(T_k^{-1})_{1,k}} T_k^{-1} e^k (e^1)^T T_k^{-1}.$$

From Theorem 4.15 we have a characterization of the solution of the three-term recurrence (4.8) involving the inverse of T_k.

Theorem 4.16. The k first elements of the solution of the three-term recurrence (4.8) are given by

$$s^k = \left(L_{k+1}^{-1} \begin{pmatrix} s_1 \\ h \end{pmatrix} \right)_{1:k} = \frac{s_1}{(T_k^{-1})_{1,k}} T_k^{-1} e^k + \left[I - \frac{1}{(T_k^{-1})_{1,k}} T_k^{-1} e^k (e^1)^T \right] T_k^{-1} h. \tag{4.13}$$

Moreover, the last element is

$$s_k = (T_k^{-1} h)_k - \frac{(T_k^{-1})_{k,k}}{(T_k^{-1})_{1,k}} (T_k^{-1} h)_1 + \frac{(T_k^{-1})_{k,k}}{(T_k^{-1})_{1,k}} s_1.$$

The solution can also be written as

$$s_k = \frac{(T_k^{-1})_{k,k}}{(T_k^{-1})_{1,k}} s_1 + \frac{1}{\zeta_k (T_{k-1}^{-1})_{1,k-1}} \sum_{j=1}^{k-1} (T_{k-1}^{-1})_{j,1} f_j.$$

For the components of the Lanczos vectors in the directions of the eigenvectors of A we apply Theorem 4.16 with $\breve{T}_k = \tilde{T}_k - \lambda_i I$ (which is nonsingular) instead of T_k, where $\tilde{T}_k$ is the computed Lanczos matrix. This gives

$$\bar{v}_i^{k+1} = \frac{(\breve{T}_{k+1}^{-1})_{k+1,k+1}}{(\breve{T}_{k+1}^{-1})_{1,k+1}} \bar{v}_i^1 + \frac{1}{\eta_{k+1} (\breve{T}_k^{-1})_{1,k}} \sum_{j=1}^{k} (\breve{T}_k^{-1})_{j,1} \bar{f}_i^j.$$

It can be shown that the first term on the right-hand side of the last identity is

$$\frac{(\breve{T}_{k+1}^{-1})_{k+1,k+1}}{(\breve{T}_{k+1}^{-1})_{1,k+1}}\bar{v}_i^1 = \tilde{p}_{1,k+1}(\lambda_i)\bar{v}_i^1,$$

where $\tilde{p}_{1,k}$ is the polynomial defined in Theorem 4.13.

We will finish this section by showing that the previous results are useful when bounding perturbation terms. Going back to (4.13) and denoting

$$U_k = I - \frac{1}{(\breve{T}_k{}^{-1})_{1,k}}\breve{T}_k{}^{-1}e^k(e^1)^T$$

and $h^{(i)} = \begin{pmatrix}\bar{f}_i^1 & \cdots & \bar{f}_i^k\end{pmatrix}^T$, we can bound the perturbation term $U_k\breve{T}_k{}^{-1}h^{(i)}$ by

$$\|U_k\breve{T}_k{}^{-1}h^{(i)}\| \leq \|U_k\|\,\|\breve{T}_k{}^{-1}\|\,\|h^{(i)}\|.$$

It can be shown (see Meurant (2006)) that $\|U_k\|$ is bounded by $C\sqrt{k}/|\bar{v}_i^1|$, where C is a constant independent of k, when the component of the initial vector in the direction of the ith eigenvector $|\bar{v}_i^1| = (q^i, v^1)$ is different from zero. This result seems not to be optimal since, when $|\bar{v}_i^1|$ is small, the bound can be large. This can possibly reflect the fact that, in this case, $\tilde{T}_k - \lambda_i I$ can be close to singular. Using this bound, we have the following result.

Theorem 4.17. Using the previous notation and supposing $|\bar{v}_i^1| \neq 0$, the perturbation term in (4.13) is bounded by

$$\|U_k\breve{T}_k{}^{-1}h^{(i)}\| \leq \frac{C\sqrt{k}}{|\bar{v}_i^1|}\,\frac{\|h^{(i)}\|}{\min_j(\theta_j^{(k)} - \lambda_i)}.$$

We note that

$$\|h^{(i)}\|^2 = \sum_{j=1}^{k}(q^i, f^j)^2 \leq \sum_{j=1}^{k}\|f^j\|^2.$$

Theorem 4.17 shows that if $\min_j(\theta_j^{(k)} - \lambda_i)$ is large (no Ritz value is close to λ_i), the perturbation term for the ith component $(Q^T\tilde{v}^{k+1})_i$ of the projection of the finite precision Lanczos vector stays bounded and small, as long as $|\bar{v}_i^1|$ is not too small.

This represents a different point of view to the behaviour of the finite precision Lanczos algorithm, which also helps in understanding some properties of CG convergence in presence of round-off errors. However, the approach here does not allow us to study how $|\bar{v}_i^k|$ varies, since $(\breve{T}_k^{-1})_{j,1}$ seems to be difficult to analyse.

5. The conjugate gradient algorithm in finite precision

Let us start with an example. Figure 5.1 depicts the Euclidean norm of the residual when the conjugate gradient algorithm is applied to a linear system with the matrix D30, a right-hand side of all ones and starting vector equal to zero. The solid line corresponds to the finite precision CG computation and the dashed line to CG with full reorthogonalization of the iteratively computed residual vectors at each step. As expected, in the latter case the residual vanishes at iteration 30. However, in finite precision arithmetic it takes many more iterations to get a small residual. Notice that even to reach a modest decrease, the number of iterations is considerably larger than the order of the matrix.

In finite precision arithmetic CG exhibits similar problems to the Lanczos algorithm: the residual vectors lose their orthogonality. Moreover in comparison to what happens in exact arithmetic or with reorthogonalization, convergence of the CG approximate solution is delayed. Intuitively, this observed fact is closely related to convergence of Ritz values. In CG the tridiagonal matrix T_k and the Ritz values do not appear explicitly, therefore the appearance of multiple Ritz approximations to single original eigenvalues is hard to notice for a practical user of the algorithm. Since we know that ideally CG behaviour depends on convergence of the Ritz values to eigenvalues (see Section 3), we may also expect the same numerically. An appearance

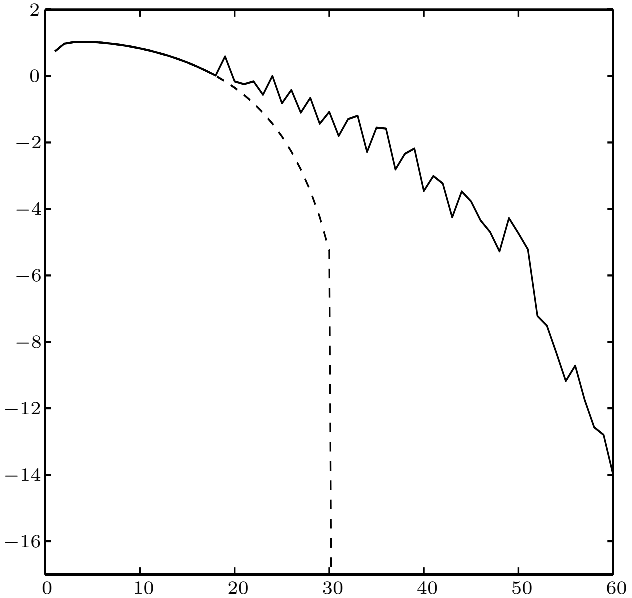

Figure 5.1. D30, $\log_{10}$ of the norms of residuals.

of multiple Ritz approximations to some eigenvalues delays convergence of Ritz values to other eigenvalues. Consequently it also delays convergence of the approximate solutions in the finite precision CG algorithm.

We first recall the relationships between the Lanczos and CG algorithms in finite precision arithmetic. For the finite precision CG algorithm we present, based on the existing literature, the corresponding CG–Lanczos recurrence which resembles, apart from the different perturbation error terms, the finite precision Lanczos algorithm from Section 4. Using the established correspondence, we then use the knowledge about the finite precision Lanczos algorithm in order to understand the finite precision CG behaviour.

In practical applications it is important to estimate the errors of computed approximate solutions. We recall the state-of-the-art error estimates and explain the instrumental role of rounding error analysis in convergence evaluation and in formulation of a meaningful stopping criteria. Finally, in addition to delaying convergence, rounding errors also limit the maximal attainable accuracy of the computed approximate solutions. We address this issue and end the section by pointing out some recent developments.

5.1. Local rounding errors and the CG–Lanczos recurrence

In analogy to the finite precision Lanczos algorithm, the recurrences for the CG quantities (*cf.* (3.2)) computed in finite precision arithmetic can be written in the form

$$\begin{aligned}
\overline{\gamma}_{k-1} &\equiv \gamma_{k-1} + \delta_\gamma^{k-1} \equiv \frac{\|r^{k-1}\|^2}{(p^{k-1}, Ap^{k-1})} + \delta_\gamma^{k-1},\\
x^k &= x^{k-1} + \gamma_{k-1} p^{k-1} + \delta_x^k,\\
r^k &= r^{k-1} - \gamma_{k-1} A p^{k-1} + \delta_r^k,\\
\overline{\beta}_k &\equiv \beta_k + \delta_\beta^k \equiv \frac{\|r^k\|^2}{\|r^{k-1}\|^2} + \delta_\beta^k,\\
p^k &= r^k + \beta_k p^{k-1} + \delta_p^k,
\end{aligned} \tag{5.1}$$

where the perturbation terms also depend, in addition to ε_M, n and $\|A\|$, on the norms and absolute values of the computed vector and scalar quantities respectively. The detailed bounds for the perturbation terms can be found in relations (7.9)–(7.14) of Strakoš and Tichý (2002, p. 71); see also Meurant (2006). The local orthogonality between the vectors r^{k+1} and r^k, r^{k+1} and p^k, p^{k+1} and Ap^k can also be bounded analogously to the local orthogonality among the computed subsequent Lanczos vectors in Theorem 4.1, but the bounds (and the proofs) are considerably more complicated. They depend, in addition to ε_M, n and $\|A\|$ also on $\kappa(A)$ and $\|r^k\|^2$: see Strakoš and Tichý (2002, Section 9).

As in the ideal CG algorithm in Section 3.1 we can write a three-term recurrence for the computed residuals:

$$r^k = -\gamma_{k-1} A r^{k-1} + \left(1 + \frac{\gamma_{k-1}\beta_{k-1}}{\gamma_{k-2}}\right) r^{k-1} - \frac{\gamma_{k-1}\beta_{k-1}}{\gamma_{k-2}} r^{k-2} + \Delta_r^k, \quad k \geq 2,$$
$$r^1 = r^0 - \gamma_0 A r^0 + \Delta_r^0.$$

Introducing the CG–Lanczos vectors w^k determined from the iteratively computed CG residuals (in finite precision arithmetic w^k is not generally identical to the vector v^k computed via the finite precision Lanczos algorithm)

$$w^{k+1} = (-1)^k \frac{r^k}{\|r^k\|}, \quad k = 1, 2, \ldots,$$

we get the following theorem.

Theorem 5.1. The three-term recurrence for the CG–Lanczos vectors determined from the finite precision CG algorithm is

$$\eta_{k+1} w^{k+1} = A w^k - \alpha_k w^k - \eta_k w^{k-1} + \Delta_w^k, \quad k = 2, 3, \ldots, \tag{5.2}$$

where

$$\eta_{k+1} = \frac{\sqrt{\beta_k}}{\gamma_{k-1}}, \qquad \alpha_k = \frac{1}{\gamma_{k-1}} + \frac{\beta_{k-1}}{\gamma_{k-2}}, \qquad \alpha_1 = \frac{1}{\gamma_0}$$

with

$$\gamma_k = \frac{\|r^k\|^2}{(Ap^k, p^k)}, \qquad \beta_k = \frac{\|r^k\|^2}{\|r^{k-1}\|^2}$$

and the initial vectors given by

$$w^1 = r^0/\|r^0\| + \Delta_w^0, \qquad \eta_2 w^2 = A w^1 - \alpha_1 w^1 - \Delta_w^1.$$

Here the recurrence is based on the coefficients γ_k and β_k determined from the computed r^k, r^{k-1} and p^k *exactly.* If we want to refer to the computed coefficients $\overline{\gamma}_k$ and $\overline{\beta}_k$, we still have the same kind of relationship but with slightly different perturbation terms. We leave the bound on the perturbation terms Δ_w^k to Meurant (2006).

5.2. Results of the backward-like analysis of CG

Based on Theorem 5.1, the analysis of Section 4.3 will also apply to the finite precision CG algorithm: see Greenbaum (1989, p. 24). The tridiagonal matrix T_k, with the entries defined by the finite precision CG algorithm as described in Theorem 5.1, is equal to that generated by the *exact* CG algorithm for a matrix whose eigenvalues lie within small intervals of the original eigenvalues. This relationship implies that the Euclidean norms of the residuals in the finite precision CG algorithm are the same as those in the correspondingly constructed exact CG recurrence. With the A-norm of the error, which is minimized at each step of the ideal CG algorithm, the situation is technically more complicated, as the reader can find in Greenbaum (1989, Theorem 3, pp. 26–29), since the definition of the norm depends on the matrix. It can still be concluded, however, that the A-norm of the error in the finite precision CG algorithm is reduced at approximately the same rate as the corresponding energy-norm of the error in the constructed exact CG recurrence. This has been further discussed and illustrated numerically in Greenbaum and Strakoš (1992).

Here it is assumed that the maximal attainable accuracy, which is limited because of rounding errors, is far away. We are solely interested in the delay of convergence. In principle, the delay at step k is given by the rank-deficiency of a basis of the computed Krylov subspace. This is in fact determined from the numerical rank (for some appropriate threshold criterion) of the computed matrix $W_k = (w^1, \dots, w^k)$, where the w^j are the CG–Lanczos vectors: *cf.* Paige and Strakoš (1999).

The results can be quantified in various ways using the polynomial formulation of the CG algorithm. Instead of working with orthogonal polynomials corresponding to the distribution function with n points of increase $\lambda_1 < \lambda_2 < \cdots < \lambda_n$ (we again assume, for simplicity of notation, that the eigenvalues of A are distinct), one must, however, consider orthogonal polynomials with respect to distribution functions having possibly many points of increase close to some or each λ_j.

In constructing the bounds one must consider the minimax polynomials on the union of tiny intervals containing the eigenvalues λ_j: see Greenbaum (1989), Greenbaum and Strakoš (1992), Greenbaum (1994). This seemingly small difference generally has a dramatic impact. We notice this from the fact that rounding errors can make a dramatic difference to the behaviour of the CG errors and residuals: see, *e.g.*, the example presented in Figure 5.1 above. The last fact is obvious, but in terms of polynomials it is not always correctly understood. This sometimes leads to misleading statements relating convergence behaviour of finite precision CG to incorrectly interpreted and simplified approximations to the minimal polynomial of A.

An example of a rigorous and instructive extension of the results from Greenbaum (1989) and Greenbaum and Strakoš (1992) can be found in Notay (1993), where the author presents bounds for the delay of convergence of the finite precision CG algorithm in the presence of isolated outlying eigenvalues.

5.3. Estimates of the error norms

As we have seen in Section 3.3, the initial error $\epsilon^0 = x - x^0$ and the kth error $\epsilon^k = x - x^k$, measured in the A-norm, are in exact precision CG ideally related by the identity

$$\frac{\|\epsilon^0\|_A^2}{\|r^0\|^2} = k\text{th Gauss quadrature } + \frac{\|\epsilon^k\|_A^2}{\|r^0\|^2},$$

where ϵ^0 and ϵ^k are unknowns and the kth Gauss quadrature can be determined by

$$(e^1)^T T_k^{-1} e^1 = \sum_{l=1}^{k} \gamma_{l-1} \|r^{l-1}\|^2.$$

In order to get an estimate for $\|\epsilon^k\|_A^2$, we have to eliminate $\|\epsilon^0\|_A^2$: see Golub and Strakoš (1994, pp. 262–263). Subtracting the identities for k and $k+d$,

$$\frac{\|\epsilon^k\|_A^2}{\|r^0\|^2} = (k+d)\text{th Gauss quadrature} - k\text{th Gauss quadrature } + \frac{\|\epsilon^{k+d}\|_A^2}{\|r^0\|^2}.$$

Since the last term on the right-hand side is always nonnegative (and strictly smaller than the term on the left-hand side), the difference between the Gauss quadratures determines in exact arithmetic the square of the lower bound for $\|\epsilon^k\|_A/\|r^0\|$.

Based on the analysis of the Gauss quadrature, Golub and Strakoš (1994) proved that this bound also works in finite precision CG computations until $\|\epsilon^k\|_A/\|r^0\|$ drops below the level $\sqrt{\varepsilon_M}$. An appropriate numerically stable implementation of this estimate was proposed by Golub and Meurant (1997). Experimental evidence shows that estimates obtained with this implementation are not significantly affected by rounding errors until the finite precision CG algorithm reaches its maximal attainable accuracy level. The proof from Golub and Strakoš (1994) cannot, however, be extended in order to justify that.

As mentioned in Section 3.3, using some simple algebraic manipulations and a lengthy rounding error analysis Strakoš and Tichý (2002) proved that in the finite precision CG algorithm the A-norm of the error satisfies

$$\|\epsilon^k\|_A^2 - \|\epsilon^{k+1}\|_A^2 = \gamma_k \|r^k\|^2 + \delta_\epsilon^k, \tag{5.3}$$

where δ_ϵ^k depends on the loss of orthogonality between r^{k+1} and p^k. Based on (5.3),

$$\nu_{k,k+d} = \sum_{l=k}^{k+d-1} \gamma_l \|r^l\|^2 \tag{5.4}$$

can be used as a lower bound for $\|\epsilon^k\|_A^2$, and this lower bound is not significantly affected by rounding errors until $\|\epsilon^k\|_A/\|\epsilon^0\|_A$ reaches a level proportional to the machine precision: see Strakoš and Tichý (2002, Section 10).

We wish to emphasize an important point. The numerical justification for (5.4) as the squared lower bound for $\|\epsilon^k\|_A$ is in no way based on the fact that in finite precision arithmetic this term is evaluated with negligible additional errors (here we do not even consider them). It is based on the nontrivial fact that (5.3) holds for the finite precision CG approximate solutions, and that δ_ϵ^k is small. We see an analogy with the rounding error analysis of the accuracy of Ritz values in the finite precision Lanczos algorithm given by Paige: see Section 4. Here again, the error estimate is also valid in finite precision computations, but we know this *only because of rigorous and nontrivial mathematical proofs.* It can be easily shown that ideally equivalent but numerically different formulas can lead to highly misleading results: see Strakoš and Tichý (2002, Figure 6.1, p. 69) and Strakoš and Liesen (2005, Figure 8, p. 319). Error estimates without appropriate rounding error analyses represent a highly hazardous pursuit.

In order to get a lower bound for the A-norm of the error at step k, we need to perform d extra steps. If the A-norm of the error reasonably drops at around step k, then d can be small. If on the other hand the A-norm of the error almost stagnates, then a small d will not ensure a close lower bound. Of course, the actual convergence behaviour is not known: it is to be estimated. Therefore the choice of d represents a difficult open problem. In any case, the proposed lower bound offers extra information which is computable at negligible additional cost, and which can with great benefit complement the commonly used measures of convergence: see Arioli (2004), Arioli, Noulard and Russo (2001), Strakoš and Liesen (2005), Strakoš and Tichý (2005) and Meurant (1999*a*). Moreover, if we agree to store one additional real number per iteration, we can easily update the previous estimates at each step. Together with the estimate for $\|\epsilon^k\|_A$ based on d, we can get (at step $k+d$) an estimate for $\|\epsilon^{k-1}\|_A$ based on $d+1$, an estimate for $\|\epsilon^{k-2}\|_A$ based on $d+2$, *etc.* In this way, the convergence of CG measured by the A-norm of the error can be 'reconstructed' using lower bounds: see Strakoš and Tichý (2005, Figure 5.4).

In linear systems arising from finite element discretizations of self-adjoint elliptic partial differential equations, it is natural to evaluate CG convergence via the relative A-norm of the error

$$\frac{\|\epsilon^k\|_A}{\|x\|_A} = \frac{\|x - x^k\|_A}{\|x\|_A}$$

(see Arioli (2004)). Subtracting the ideal identities

$$\|\epsilon^0\|_A^2 = \nu_{0,k+d} + \|\epsilon^{k+d}\|_A^2,$$
$$\|\epsilon^0\|_A^2 = \|x - x^0\|_A^2 = \|x\|_A^2 - b^T x^0 - (r^0)^T x^0$$

gives

$$\|x\|_A^2 = \nu_{0,k+d} + b^T x^0 + (r^0)^T x^0 + \|\epsilon^{k+d}\|_A^2,$$
$$\|x\|_A^2 \ge \mu_{k+d} \equiv \nu_{0,k+d} + b^T x^0 + (r^0)^T x^0.$$

We will assume that $\|x - x^0\|_A \leq \|x\|_A$. This represents a very natural assumption which should never be violated in practical computations. Indeed, it is meaningless to use a nonzero x^0 without justification that guarantees that a nonzero initial approximation is better than taking $x^0 = 0$. For CG, the A-norm of the error represents the proper measure of 'goodness'. If in doubt, it is always possible to scale an initial approximation such that $\|x - \alpha x^0\|_A$ is minimal, which gives

$$\alpha = \frac{b^T x^0}{(x^0)^T A x^0}$$

(see Strakoš and Tichý (2005)). If $\|x - x^0\|_A \leq \|x\|_A$, then it is easy to show that $\mu_{k+d} > 0$, and an algebraic manipulation ideally gives

$$\frac{\|\epsilon^k\|_A^2}{\|x\|_A^2} \geq \frac{\nu_{k,d}}{\mu_{k+d}} > 0,$$

i.e., in exact precision CG, $\nu_{k,d}/\mu_{k+d}$ is a lower bound for the squared relative A-norm of the error. Since numerically all considerations leading to this bound are based on *local orthogonality only*, this estimate is also well established (though not always a lower bound) for the finite precision CG algorithm. For further details we refer to Strakoš and Tichý (2005) and also to Strakoš and Tichý (2002), who also describe estimation of the Euclidean norm of the error and presents open problems. For the estimation of the Euclidean norm see also Meurant (2005), and for that norm in finite precision see Meurant (2006).

Various other options for computing the error bounds in the CG algorithm are summarized by Calvetti, Morigi, Reichel and Sgallari (2000). Based on quadrature considerations, the bounds are more complicated. They cannot be easily justified for finite precision CG computations. Still, they can prove useful in some particular applications. Interesting ideas concerning the upper bounds for the A-norm of the error can also be found in Greenbaum (1997*a*, p. 108) and in Golub and Meurant (1997).

5.4. Maximal attainable accuracy

Rounding errors generally do not allow the finite precision CG algorithm to produce approximate solutions to arbitrary accuracy. It is therefore important to find out the maximal attainable accuracy which can be reached for a given A and b. The importance of this question is, however, more in the impact which the corresponding analysis has on understanding the CG algorithm and its implementations, than in practical applications of the results. In most applications, perhaps with the exception of some inner CG iterations used in nonlinear optimization, or difficult problems with $\|A\|$ large, the computation is stopped much before the maximal attainable accuracy is reached.

Here we will assume, as above, that A is symmetric positive definite and not close to singular, and we will concentrate on limitations on the maximal attainable accuracy caused by the possible amplification of elementary round-off throughout the recurrences. We leave other effects, which can be observed in indefinite systems or near-singular systems, to the literature: see, *e.g.*, Sleijpen, van der Vorst and Modersitzki (2001). Work on maximal attainable accuracy has focused on the residual as the easiest and most common measure of convergence. Based on the residual, bounds for the maximal attainable accuracy measured by the Euclidean or the A-norm of the error can easily be obtained using the obvious relationships, together with the characterization of conditioning of the matrix A.

In the CG algorithm, the residual vector is recursively computed at each step as a part of the recurrence. In finite precision arithmetic, this recursively computed residual r^k (see (5.1)) can differ from the directly computed quantity $b - Ax^k$, which is generally called the true residual. Convergence of the recursive residuals was analysed by Wozniakowski (1978, 1980) and Bollen (1984), for example. Although some assumptions used there cannot in general be satisfied by the CG recurrence (5.1), the results proved useful in a further analysis: see Greenbaum (1994, 1997*b*). For a survey of the early developments see Higham (2002).

In Theorem 2 of Greenbaum (1989), the question of the difference between the true residual and the recursively computed residual was analysed for the first time, to our knowledge. It was shown that this difference at step k can be bounded by a simple sum of the elementary perturbation terms at steps 0 (which means computation of the initial residual) to k, *i.e.*,

$$\|r^k - (b - Ax^k)\| \leq \|\delta_r^0\| + \sum_{l=1}^{k} \big(\|\delta_r^l\| + \|A\delta_x^l\|\big).$$

Sleijpen, van der Vorst and Fokkema (1994), Greenbaum (1994), and slightly later Greenbaum (1997*b*) studied this problem further, resulting in the bound

$$\frac{\|r^k - (b - Ax^k)\|}{\|A\|\,\|x\|} \leq \mathcal{O}(k)\varepsilon_M \left(1 + \max_{l \leq k} \frac{\|x^l\|}{\|x\|}\right). \tag{5.5}$$

If $\|r^k\|$ becomes of the order of the machine precision, which is often observed numerically for large k but which has not yet been completely proved in the given literature, then (5.5) gives a bound for the maximal attainable accuracy measured by the true residual norm divided by $\|A\|\ \|x\|$.

This result offers the following insight into the behaviour of the finite precision CG algorithm. One can expect a high maximal attainable accuracy with the finite precision CG algorithm if the norms of the iterates do not significantly exceed the norm of the true solution. Since ideally the Euclidean norm of the error is strictly decreasing, $\|x - x^k\| < \|x - x^0\|$ implies

$$\|x^k\| \leq 2\|x\| + \|x^0\|.$$

Using the backward-like error analysis of Greenbaum described above, this upper bound holds true, to within a small error, in the finite precision CG algorithm. With a reasonable choice of $\|x^0\|$, the finite precision CG algorithm can therefore be expected to achieve a high maximal attainable accuracy if $\|A\|$ is not too large.

The situation is dramatically different in CG-like algorithms applied to nonsymmetric systems, to many of which the above analysis can also be applied. For detailed discussions see Greenbaum (1997*b*) and Greenbaum (1997*a*, Section 7.3).

When the CG algorithm is implemented via the mathematically equivalent three-term recurrence (for examples see Rutishauser (1959) and Hageman and Young (1981)), the maximal attainable accuracy is much more vulnerable to local errors. As shown by Gutknecht and Strakoš (2000), the difference $r^k - (b - Ax^k)$ is then equal to a sum of local error terms (different from those in the analysis of the two term recurrences above) plus multiples of the same terms by factors which can become large if the norm of the iteratively computed residual oscillates, *i.e.*, if

$$\max_{0 \leq l < j \leq k} \frac{\|r^j\|^2}{\|r^l\|^2} \quad \text{is large.}$$

Consequently a large increase in the norm of the computed iterative residuals can damage the maximal attainable accuracy. Moreover, damage caused at an early stage of the computation cannot in general be compensated for in the subsequent iterations. The technique used in Gutknecht and Strakoš (2000) is based on writing k steps of the second-order nonhomogeneous difference equation for the gap $r^k - (b - Ax^k)$ as a superposition of the $k+1$ homogeneous difference equations, which resembles the technique used in a different context by Grcar (1981). For further details we refer to Gutknecht and Strakoš (2000). As pointed out in the concluding part of the last paper, the same result can also be attained by using matrix approach analogous to that of Paige (1980). The matrix approach allows easier further generalizations. In some applications the matrix A is not explicitly available, and the matrix–vector multiplication is performed by solving an auxiliary problem. It might therefore be convenient to relax the accuracy of this operation. That can, however, affect convergence behaviour and the maximal attainable accuracy. Analysis of this problem goes far beyond the investigation of numerical stability. Several authors have recently presented interesting results focused mostly on maximal attainable accuracy: see, *e.g.*, Bouras and Frayssé (2005), and the surveys in Simoncini and Szyld (2005, Section 11) and van den Eshof (2003, Chapter 5).

5.5. Recent developments

In this section we summarize some recent results about CG convergence in finite precision arithmetic: see Meurant (2006).

For the recurrence of w^k given in (5.2) we can directly apply the results we have reviewed for general three-term recurrences: see Theorem 4.16. Let us denote by $\bar{w}_i^k$ the component of w^k in the direction of the ith eigenvector of A, $i = 1, 2, \ldots, n$.

Theorem 5.2. Let

$$\bar{\delta}^k \equiv (\bar{\delta}_1^k, \ldots, \bar{\delta}_n^k)^T = Q^T \Delta_w^k,$$

let j be given and let $p_{j,k}$ be the polynomial determined by

$$p_{j,j-1}(\lambda) = 0, \qquad p_{j,j}(\lambda) = 1,$$
$$\eta_{k+1} p_{j,k+1}(\lambda) = (\lambda - \alpha_k) p_{j,k}(\lambda) - \eta_k p_{j,k-1}(\lambda), \quad k = j, j+1, \ldots.$$

The solution of the perturbed recurrence

$$\eta_{k+1} \bar{w}_i^{k+1} = (\lambda_i - \alpha_k) \bar{w}_i^k - \eta_k \bar{w}_i^{k-1} + \bar{\delta}_i^k$$

starting from $w_i^0 = 0$ and w_i^1 is given by

$$\bar{w}_i^{k+1} = p_{1,k+1}(\lambda_i) \bar{w}_i^1 + \sum_{l=1}^{k} p_{l+1,k+1}(\lambda_i) \frac{\bar{\delta}_i^l}{\eta_{l+1}}, \quad i = 1, \ldots, n.$$

This immediately leads to an expression for w^{k+1}:

$$w^{k+1} = p_{1,k+1}(A) w^1 + \sum_{l=1}^{k} p_{l+1,k+1}(A) \frac{\Delta_w^l}{\eta_{l+1}}.$$

Then, using the correspondence between w^{k+1} and r^k, we can express the recursively determined CG residual vector computed in finite precision arithmetic in the following form.

Theorem 5.3. Using the notation of Theorem 5.2,

$$r^k = (-1)^k \frac{\|r^k\|}{\|r^0\|} p_{1,k+1}(A) r^0 + (-1)^k \|r^k\| \sum_{l=1}^{k} p_{l+1,k+1}(A) \frac{\Delta_w^l}{\eta_{l+1}}.$$

In exact arithmetic, after a Ritz value has converged, the corresponding projections of the residual and of the error on the corresponding eigenvector vanish. This is not the case in finite precision arithmetic. After decreasing for a while, the projection of the residual on the subspace generated by the corresponding eigenvector of A rises back to contribute to the norm of the residual, because of the amplification of the local round-off. Once a new Ritz copy is formed, the component again decreases, *etc.* This can delay convergence and lead to oscillations of the residual components in the directions of the individual eigenvectors of A, and, as a consequence, to oscillations of the residual norm. In comparison with the expression for the finite precision Lanczos–CG vector w^{k+1}, the perturbation term in Theorem 5.3 is multiplied by $\|r^k\|$. Therefore, for small $\|r^k\|$ the possible oscillations caused by possible amplification of the error terms are typically much less pronounced in r^k than in w^{k+1}.

When considering CG convergence, we have to be careful on how to link the error to the computed quantities. Ideally, the error is $\epsilon^k = x - x^k$ where x is the exact solution and it is related to the residual by $A\epsilon^k = r^k$. But this is only true if the residual is $b - Ax^k$. We have seen in Section 5.4 that the computed iterative residual can be different from $b - Ax^k$. Hence there are more alternatives. Considering the ultimate stagnation of $\|b - Ax^k\|$, it seems reasonable to work (besides the true error linked to the true residual) with $(A^{-1}r^k, r^k) = \|A^{-\frac{1}{2}}r^k\|^2$, where r^k is the (recursively) computed iterative residual, as another useful measure. We denote it $\varepsilon^k \equiv A^{-1}r^k$. Then, we have the following result whose proof is based on a lengthy analysis of local orthogonality: see Meurant (2006) and Strakoš and Tichý (2002, (10.1)).

Proposition 5.4.

$$\|\varepsilon^{k+1}\|_A^2 = \|\varepsilon^k\|_A^2 - \gamma_k\|r^k\|^2 + \varepsilon_M C_1^k\|r^k\|^2 + \varepsilon_M^2 C_2^k\|r^k\|^2,$$

where $|C_1^k|$ and $|C_2^k|$ are bounded by quantities involving $\|r^k\|$ and $\|p^k\|$.

This proposition leads to a result about strict decrease of the error norm under a restriction on the condition number of A.

Theorem 5.5. If

$$\kappa(A) < \frac{1}{\varepsilon_M \lambda_1 |C_1^k|} + \mathcal{O}(\varepsilon_M), \quad \text{for all } k,$$

then

$$\|\varepsilon^{k+1}\|_A < \|\varepsilon^k\|_A.$$

Hence, if the condition number is not too large, $\|\varepsilon^k\|_A$ is, as in exact arithmetic, strictly decreasing. However, having a limitation on $\kappa(A)$ is not satisfactory, since in numerical computations we hardly observe an increase or oscillation of $\|\varepsilon^k\|_A$.

This result complements those of Anne Greenbaum (1989) who obtained a decrease of the A-norm of the error without an explicit restriction on the condition number but with additional small terms. A proof of the strict decrease of $\|\varepsilon^k\|_A$ and a proof that the computed iterative residual must ultimately vanish, *i.e.*, $\|r^k\| \to 0$, without any restriction on the condition number, remains open.

6. Conclusions

The Lanczos and conjugate gradient algorithms are considered effective numerical tools for computing eigenvalues, approximating matrix functions and quadratic forms, and for solving (linear) algebraic equations. As we have seen, they also represent interesting mathematical objects with very deep links reaching far beyond the borders of numerical linear algebra, numerical mathematics or algebraic structures. This is perhaps why the investigation into their behaviour in exact and in finite precision arithmetic is leading to results which, piece by piece, are being assembled into a rigorous, consistent, rich and beautiful mathematical theory. In this way the Lanczos and conjugate gradient algorithms represent another example along the lines drawn by Baxter and Iserles (2003). The rigour and beauty of their mathematical structure, including the effects of rounding errors, reveals once again the presence of such attributes in the field called computational mathematics.

Acknowledgements

We sincerely thank Iveta Hnětynková, Jörg Liesen, Jurjen Duintjer Tebbens and Petr Tichý for comments and corrections, Chris Paige for a careful reading of earlier versions of the manuscript, suggesting valuable improvements and correcting our wording in many places, and Gene Golub for suggesting the writing of this paper.

REFERENCES

M. Arioli (2004), A stopping criterion for the conjugate gradient algorithms in a finite element method framework, *Numer. Math.* **97**, 1–24.

M. Arioli, E. Noulard and A. Russo (2001), Stopping criteria for iterative methods: applications to PDE's, *Calcolo* **38**, 97–112.

W. E. Arnoldi (1951), The principle of minimized iterations in the solution of the matrix eigenvalue problem, *Quart. Appl. Math.* **9**, 17–29.

O. Axelsson and I. Kaporin (2001), Error norm estimations and stopping criteria in preconditioned conjugate gradient iterations, *Numer. Linear Algebra Appl.* **8**, 265–286.

O. Axelsson and G. Linskog (1986), On the rate of convergence of the preconditioned conjugate gradient methods, *Numer. Math.* **51**, 209–227.

B. J. C. Baxter and A. Iserles (2003), On the foundations of computational mathematics, in Vol. XI of *Handbook of Numerical Analysis*, North-Holland, Amsterdam, pp. 3–34.

B. Beckermann and A. Kuijlaars (2002), Superlinear CG convergence for special right-hand sides, *Electron. Trans. Numer. Anal.* **14**, 1–19.

Å. Björck, T. Elfving and Z. Strakoš (1998), Stability of conjugate gradients and Lanczos methods for linear least squares problems, *SIAM J. Matrix Anal. Appl.* **19**, 720–736.

J. A. M. Bollen (1980), Round-off error analysis of descent methods for solving linear equations, PhD thesis, Technische Hogeschool Eindhoven, the Netherlands.

J. A. M. Bollen (1984), Numerical stability of descent methods for solving linear equations, *Numer. Math.* **43**, 361–377.

A. Bouras and V. Frayssé (2005), Inexact matrix–vector products in Krylov methods for solving linear systems: A relaxation strategy, *SIAM J. Matrix Anal. Appl.* **26**, 660–678.

D. Calvetti, S. Morigi, L. Reichel and F. Sgallari (2000), Computable error bounds and estimates for the conjugate gradient, *Numer. Algorithms* **25**, 79–88.

E. B. Christoffel (1877), Sur une classe particulière de fonctions entières et de fractions continues, *Ann. Mat. Pura Appl.* **8**, 1–10.

J. K. Cullum and R. A. Willoughby (1985), *Lanczos Algorithms for Large Symmetric Eigenvalue Computations*, Vol. I, Theory, Vol. II, Programs, Birkhäuser. Reprinted by SIAM in the series *Classics in Applied Mathematics*.

G. Dahlquist, S. C. Eisenstat and G. H. Golub (1972), Bounds for the error of linear systems of equations using the theory of moments, *J. Math. Anal. Appl.* **37**, 151–166.

G. Dahlquist, G. H. Golub and S. G. Nash (1978), Bounds for the error in linear systems, in *Proc. Workshop on Semi-Infinite Programming* (R. Hettich, ed.), Springer, pp. 154–172.

J. W. Daniel (1967), The conjugate gradient method for linear and nonlinear operator equations, *SIAM J. Numer. Anal.* **4**, 10–26.

G. Darboux (1878), Mémoire sur l'approximation des fonctions de très grand nombres et sur une classe étendue de développements en série, *J. Mat. Pures Appl.* **4**, 5–56, 377–416.

P. Davis and P. Rabinowitz (1984), *Methods of Numerical Integration*, second edition, Academic Press.

P. Deuflhard (1994), Cascadic conjugate gradient methods for elliptic partial differential equations: Algorithm and numerical results, in *Domain Decomposition Methods in Scientific and Engineering Computing* (University Park PA, 1993), Vol. 180 of *Contemporary Mathematics*, AMS, Providence, RI, pp. 29–42.

I. Dhillon and B. N. Parlett (2003), Orthogonal eigenvectors and relative gaps, *SIAM J. Matrix Anal. Appl.* **25**, 858–899.

I. Dhillon and B. N. Parlett (2004), Multiple representations to compute orthogonal eigenvectors of symmetric tridiagonal matrices, *Linear Algebra Appl.* **387**, 1–28.

V. Druskin, A. Greenbaum and L. Knizhnerman (1998), Using nonorthogonal Lanczos vectors in the computation of matrix functions, *SIAM J. Sci. Comput.* **19**, 38–54.

V. Druskin and L. Knizhnerman (1991), Error bounds in the simple Lanczos procedure for computing functions of symmetric matrices and eigenvalues, *Comput. Math. Math. Phys.* **31**, 20–30.

V. Druskin, L. Borcea and L. Knizhnerman (2005), On the sensitivity of Lanczos recursions to the spectrum, *Linear Algebra Appl.* **396**, 103–125.

S. Elhay, G. M. L. Gladwell, G. H. Golub and Y. M. Ram (1999), On some eigenvector-eigenvalue relations, *SIAM J. Matrix Anal. Appl.* **20**, 563–574.

J. van den Eshof (2003), Nested iteration methods for nonlinear matrix problems. PhD thesis, University of Utrecht.

B. Fischer (1996), *Polynomial Based Iteration Methods for Symmetric Linear Systems*, Wiley, Chichester.

R. Fletcher (1976), Conjugate gradient methods for indefinite systems, in *Numerical Analysis: Proc. 6th Biennial Dundee Conf., Univ. Dundee* (Dundee, 1975), Vol. 506 of *Lecture Notes in Mathematics*, Springer, Berlin, pp. 73–89.

V. M. Fridman (1963), The method of minimum iterations with minimum errors for a system of linear algebraic equations with a symmetrical matrix, *USSR Comput. Math. Math. Phys.* **2**, 362–363.

F. R. Gantmacher (1959), *The Theory of Matrices*, Vol. 1 and 2, Chelsea Publishing Co., New York.

W. Gautschi (1968), Construction of Gauss–Christoffel quadrature formulas, *Math. Comp.* **22**, 251–270.

W. Gautschi (1981), A survey of Gauss–Christoffel quadrature formulae, in *E. B. Christoffel: The Influence of His Work on Mathematics and the Physical Sciences* (P. L. Bultzer and F. Fehér, eds), Birkhauser, Boston, pp. 73–157.

W. Gautschi (1982), On generating orthogonal polynomials, *SIAM J. Sci. Statist. Comput.* **3**, 289–317.

W. Gautschi (1985), Orthogonal polynomials: Constructive theory and applications, *J. Comput. Appl. Math.* **12 & 13**, 61–76.

W. Gautschi (2002), The interplay between classical analysis and (numerical) linear algebra: A tribute to Gene H. Golub, *Electron. Trans. Numer. Anal.* **13**, 119–147.

W. Gautschi (2004), *Orthogonal Polynomials, Computation and Approximation*, Oxford University Press, Oxford.

G. H. Golub (1973), Some uses of the Lanczos algorithm in numerical linear algebra, in *Topics in Numerical Analysis* (J. H. H. Miller, ed.), Springer, Heidelberg/New York, pp. 23–31.

G. H. Golub and U. von Matt (1991), Quadratically constrained least squares and quadratic problem, *Numer. Math.* **59**, 561–580.

G. H. Golub and G. Meurant (1994), Matrices, moments and quadrature, in *Numerical Analysis 1993* (D. F. Griffiths and G. A. Watson, eds), *Pitman Research Notes in Mathematics*, pp. 105–156.

G. H. Golub and G. Meurant (1997), Matrices, moments and quadrature II: How to compute the norm of the error in iterative methods, *BIT* **37**, 687–705.

G. H. Golub and D. P. O'Leary (1989), Some history of the conjugate gradient and Lanczos algorithms: 1948–1976, *SIAM Rev.* **31**, 50–102.

G. H. Golub and Z. Strakoš (1994), Estimates in quadratic formulas, *Numer. Algorithms* **8**, 241–268.

G. H. Golub and J. H. Welsch (1969), Calculation of Gauss quadrature rules, *Math. Comp.* **23**, 221–230.

W. B. Gragg and W. J. Harrod (1984), The numerically stable reconstruction of Jacobi matrices from spectral data, *Numer. Math.* **44**, 317–335.

J. Grcar (1981), Analysis of the Lanczos algorithm and of the approximation problem in Richardson's method, PhD thesis, University of Illinois at Urbana–Champaign.

A. Greenbaum (1979), Comparison of splittings used with the conjugate gradient algorithm, *Numer. Math.* **33**, 181–194.

A. Greenbaum (1981), Convergence properties of the conjugate gradient algorithm in exact and finite precision arithmetic, PhD thesis, University of California, Berkeley.

A. Greenbaum (1989), Behavior of slightly perturbed Lanczos and conjugate gradient recurrences, *Linear Algebra Appl.* **113**, 7–63.

A. Greenbaum (1994), The Lanczos and conjugate gradient algorithms in finite precision arithmetic, in *Proc. Cornelius Lanczos International Centenary Conference, 1993* (J. D. Brown, M. T. Chu, D. C. Ellison and R. J. Plemmons, eds), SIAM, pp. 49–60.

A. Greenbaum (1997*a*), *Iterative Methods for Solving Linear Systems*, SIAM.

A. Greenbaum (1997*b*), Estimating the attainable accuracy of recursively computed residual methods, *SIAM J. Matrix Anal. Appl.* **18**, 535–551.

A. Greenbaum and Z. Strakoš (1992), Predicting the behavior of finite precision Lanczos and conjugate gradient computations, *SIAM J. Matrix Anal. Appl.* **13**, 121–137.

M. Gutknecht and Z. Strakoš (2000), Accuracy of two three-term and three two-term recurrences for Krylov space solvers, *SIAM J. Matrix Anal. Appl.* **22**, 213–229.

W. Hackbusch (1994), *Iterative Solution of Large Sparse Systems of Equations*, Vol. 95 of *Applied Mathematical Sciences*, Springer, New York. Translated and revised from the 1991 German original.

L. Hageman and D. Young (1981), *Applied Iterative Methods*, Academic Press, Orlando.

M. R. Hestenes and E. Stiefel (1952), Methods of conjugate gradients for solving linear systems, *J. Nat. Bur. Standards* **49**, 409–436.

M. R. Hestenes and J. Todd (1991), *Mathematicians Learning to Use Computers, National Institute of Standards and Technology Special Publication* **730**, US department of Commerce, National Institute of Standards and Technology, Washington, DC.

N. J. Higham (2002), *Accuracy and Stability of Numerical Algorithms*, second edition, SIAM.

A. S. Householder (1975), *The Theory of Matrices in Numerical Analysis*, Dover, New York. Reprint of 1964 edition.

S. Kaniel (1966), Estimates of some computational techniques in linear algebra, *Math. Comp.* **20**, 369–378.

J. Kautsky and G. H. Golub, (1983), On the calculation of Jacobi matrices, *Linear Algebra Appl.* **53/53**, 439–455.

L. Knizhnerman (1995*a*), The quality of approximations to an isolated eigenvalue and the distribution of 'Ritz numbers' in the simple Lanczos procedure, *Comput. Math. Math. Phys.* **35**, 1175–1187.

L. Knizhnerman (1995*b*), On adaptation of the Lanczos method to the spectrum. Report EMG-001-95-12, Schlumberger–Doll–Research.

L. Knizhnerman (1996), The simple Lanczos procedure: estimates of the error of the Gauss quadrature formula and their applications, *Comput. Math. Math. Phys.* **36**, 1481–1492.

A. N. Krylov (1931), O Čislemnon rešenii uravnenija, kotorym v techničeskih voprasah opredeljajutsja častoy malyh kolebaniĭ material'nyh., *Izv. Adad. Nauk SSSR old. Mat. Estet.*, pp. 491–539.

A. B. J. Kuijlaars (2006), Convergence analysis of Krylov subspace iterations with methods from potential theory, *SIAM Review* **48**, 3–40.

C. Lanczos (1950), An iteration method for the solution of the eigenvalue problem of linear differential and integral operators, *J. Res. Nat. Bur. Standards* **45**, 255–282.

C. Lanczos (1952), Solution of systems of linear equations by minimized iterations, *J. Res. Nat. Bur. Standards* **49**, 33–53.

D. P. Laurie (1999), Accurate recovery of recursion coefficients from Gaussian quadrature formulas, *J. Comput. Appl. Math.* **112**, 165–180.

D. P. Laurie (2001), Computation of Gauss-type quadrature formulas, *J. Comput. Appl. Math.* **127**, 201–217.

J. G. Lewis (1977), Algorithms for sparse matrix eigenvalue problems. PhD thesis, Report STAN-CS-77-595, Computer Science Department, Stanford University, Stanford, CA.

Ren-Cang Li (2005), On Meinardus' examples for the conjugate gradient method. Technical Report 2005-06, Department of Mathematics, University of Kentucky.

J. Liesen and P. Tichý (2005), On the worst case convergence of MR and CG for symmetric positive definite tridiagonal Toeplitz matrices, *Electron. Trans. Numer. Anal.* **20**, 180–197.

D. G. Luenberger (1969), Hyperbolic pairs in the method of conjugate gradients, *SIAM J. Appl. Math.* **17**, 1263–1267.

D. G. Luenberger (1970), The conjugate residual method for constrained minimization problems, *SIAM J. Numer. Anal.* **7**, 390–398.

G. Meinardus (1963), Über eine Verallgemeinerung einer Ungleichung von L. V. Kantorowitsch, *Numer. Math.* **5**, 14–23.

G. Meurant (1997), The computation of bounds for the norm of the error in the conjugate gradient algorithm, *Numer. Algorithms* **16**, 77–87.

G. Meurant (1999*a*), Numerical experiments in computing bounds for the norm of the error in the preconditioned conjugate gradient algorithm, *Numer. Algorithms* **22**, 353–365.

G. Meurant (1999*b*), *Computer Solution of Large Linear Systems*, North-Holland.

G. Meurant (2005), Estimates of the l_2 norm of the error in the conjugate gradient algorithm, *Numer. Algorithms* **40**, 157–169.

G. Meurant (2006), *The Lanczos and Conjugate Gradient Algorithms: From Theory to Finite Precision Computations*, SIAM, book to appear.

Y. Notay (1993), On the convergence rate of the conjugate gradients in the presence of rounding errors, *Numer. Math.* **65**, 301–317.

D. P. O'Leary and Z. Strakoš (2004), On sensitivity of Gauss–Christoffel quadrature estimates. Computer Science Department Report CS-TR-4622, Institute for Advanced Computer Studies Report UMIACS-2004-64, University of Maryland.

M. M. Overton (2001), *Numerical Computing with IEEE Floating Point Arithmetic*, SIAM.

C. C. Paige (1969*a*), Error analysis of the generalized Hessenberg processes. Technical Note ICSI 179, London University Institute of Computer Science.

C. C. Paige (1969*b*), Eigenvalues of perturbed Hermitian matrices. Technical Note ICSI 179, London University Institute of Computer Science.

C. C. Paige (1970*a*), Error analysis of the symmetric Lanczos process for the eigenproblem. Technical Note ICSI 248, London University Institute of Computer Science.

C. C. Paige (1970*b*), Practical use of the symmetric Lanczos process with reorthogonalization, *BIT* **10**, 183–195.

C. C. Paige (1971), The computation of eigenvalues and eigenvectors of very large sparse matrices, PhD thesis, University of London.

C. C. Paige (1972), Computational variants of the Lanczos method for the eigenproblem, *J. Inst. Math. Appl.* **10**, 373–381.

C. C. Paige (1976), Error analysis of the Lanczos algorithm for tridiagonalizing a symmetric matrix, *J. Inst. Math. Appl.* **18**, 341–349.

C. C. Paige (1980), Accuracy and effectiveness of the Lanczos algorithm for the symmetric eigenproblem, *Linear Algebra Appl.* **34**, 235–258.

C. C. Paige and M. Saunders (1975), Solution of sparse indefinite systems of linear equations, *SIAM J. Numer. Anal.* **12**, 617–629.

C. C. Paige and Z. Strakoš (1999), Correspondence between exact arithmetic and finite precision behavior of Krylov space methods, in *XIV Householder Symposium* (J. Varah, ed.), University of British Columbia, pp. 250–253.

C. C. Paige, B. N. Parlett and H. van der Vorst (1995), Approximate solutions and eigenvalue bounds from Krylov subspaces, *Numer. Linear Algebra Appl.* **2**, 115–133.

B. N. Parlett (1980), *The Symmetric Eigenvalue Problem*, Prentice-Hall.

B. N. Parlett (1990), The contribution of J. H. Wilkinson to numerical analysis, in *A History of Scientific Computing* (Princeton, NJ, 1987), ACM Press Hist. Ser., ACM, New York, pp. 17–30.

B. N. Parlett (1992), The rewards for maintaining semi-orthogonality among Lanczos vectors, *Numer. Linear Algebra Appl.* **1**, 234–267.

B. N. Parlett (1994), Do we fully understand the symmetric Lanczos algorithms yet?, in *Proc. Cornelius Lanczos International Centenary Conference, 1993* (J. D. Brown, M. T. Chu, D. C. Ellison and R. J. Plemmons, eds), SIAM, pp. 93–108.

B. N. Parlett (1996), Invariant subspaces for tightly clustered eigenvalues of tridiagonals, *BIT* **36**, 542–562.

B. N. Parlett and I. Dhillon (2000), Relatively robust representations of symmetric tridiagonals, *Linear Algebra Appl.* **309**, 121–151.

B. N. Parlett and D. S. Scott (1979), The Lanczos algorithm with selective orthogonalization, *Math. Comp.* **33**, 217–238.

H. Rutishauser (1959), Theory of gradient methods, in *Refined Iterative Mehods for Computation of the Solution and the Eigenvalues of Self-Adjoint Boundary Value Problems*, Mitt. Inst. Angew. Math. ETH Zürich, Birkhäuser, Basel, pp. 24–49.

Y. Saad (1980), On the rates of convergence of the Lanczos and the block Lanczos methods, *SIAM J. Numer. Anal.* **17**, 687–706.

Y. Saad (1992), *Numerical Methods for Large Eigenvalue Problems*, Wiley.

D. S. Scott (1978), Analysis of the symmetric Lanczos algorithm, PhD thesis, University of California, Berkeley.

D. S. Scott (1979), How to make the Lanczos algorithm converge slowly, *Math. Comp.* **33**, 239–247.

D. S. Scott (1981), The Lanczos algorithm, in *Sparse Matrices and Their Use* (I. S. Duff, ed.), Academic Press, pp. 139–159.

H. D. Simon (1982), The Lanczos algorithm for solving symmetric linear systems, PhD thesis, University of California, Berkeley.

H. D. Simon (1984*a*), The Lanczos algorithm with partial reorthogonalization, *Math. Comp.* **42**, 115–142.

H. D. Simon (1984*b*), Analysis of the symmetric Lanczos algorithm with reorthogonalization methods, *Linear Algebra Appl.* **61**, 101–131.

V. Simoncini and D. Szyld (2005), Recent developments in Krylov subspace methods for linear systems. Research Report 05-9-25. Department of Mathematics, Temple University.

G. L. G. Sleijpen, H. A. van der Vorst and D. R. Fokkema (1994), BiCGstab(l) and other hybrid Bi-CG methods, *Numer. Algorithms* **7**, 75–109.

G. L. G. Sleijpen, H. A. van der Vorst and J. Modersitzki (2001), Difference in the effects of rounding errors in Krylov solvers for symmetric indefinite systems, *SIAM J. Matrix Anal. Appl.* **22**, 726–751.

A. van der Sluis and H. A. van der Vorst (1986), The rate of convergence of conjugate gradients, *Numer. Math.* **48**, 543–560.

A. van der Sluis and H. A. van der Vorst (1987), The convergence behavior of Ritz values in the presence of close eigenvalues, *Linear Algebra Appl.* **88**, 651–694.

G. W. Stewart (2001), *Matrix Algorithms*, Vol. II: *Eigensystems*, SIAM.

T. J. Stieltjes (1884), Quelques recherches sur la théorie des quadratures mécaniques, *Ann. Sci. Ecole Norm. Paris* **1**, 409–426. [Oeuvres I, 377–396.]

J. Stoer (1983), Solution of large linear systems of equations by conjugate gradient type methods, in *Mathematical Programming: The State of the Art* (A. Bachem, M. Grötschel and B. Korte, eds), Springer, pp. 540–565.

J. Stoer and R. Bulirsch (1983), *Introduction to Numerical Analysis*, second edition, Springer.

J. Stoer and R. W. Freund (1982), On the solution of large linear systems of equations by conjugate gradient type methods, in *Computer Methods in Applied Science and Engineering V* (R. Glowinski and J. L. Lions, eds), North-Holland, pp. 35–53.

Z. Strakoš (1991), On the real convergence rate of the conjugate gradient method, *Linear Algebra Appl.* **154–156**, 535–549.

Z. Strakoš (1998), Convergence and numerical behavior of the Krylov space methods, *NATO ASI Institute Algorithms for Large Sparse Linear Algebraic Systems: The State of the Art and Applications in Science and Engineering* (G. Winter Althaus and E. Spedicato, eds), Kluwer Academic, pp. 175–197.

Z. Strakoš and A. Greenbaum (1992), Open questions in the convergence analysis of the Lanczos process for the real symmetric eigenvalue problem, *IMA Preprint Series* **934**, University of Minnesota.

Z. Strakoš and J. Liesen (2005), On numerical stability in large scale linear algebraic computations, *ZAMM Z. Angew. Math. Mech.* **85**, No. 5, 307–325.

Z. Strakoš and P. Tichý (2002), On error estimation in the conjugate gradient method and why it works in finite precision computation, *Electron. Trans. Numer. Anal.* **13**, 56–80.

Z. Strakoš and P. Tichý (2005), Error estimation in preconditioned conjugate gradients, *BIT Numerical Mathematics* **45**, 789–817.

R. C. Thompson and P. McEnteggert (1968), Principal submatrices II: The upper and lower quadratic inequalities, *Linear Algebra Appl.* **1**, 211–243.

H. A. van der Vorst (1982), Preconditioning by incomplete decompositions, PhD thesis, Academic Computer Centrum Utrecht.

H. A. van der Vorst (2003), *Iterative Krylov Methods for Large Linear Systems*, Cambridge University Press.

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

H. Wozniakowski (1978), Round-off error analysis of iterations for large linear systems, *Numer. Math.* **30**, 301–314.

H. Wozniakowski (1980), Round-off error analysis of a new class of conjugate gradient algorithms, *Linear Algebra Appl.* **29**, 507–529.

W. Wülling (2005), The stabilization of weights in the Lanczos and conjugate gradient methods, *BIT Numerical Mathematics* **45**, 395–414.

W. Wülling (2006), On stabilization and convergence of clustered Ritz values in the Lanczos method, *SIAM J. Matrix Anal. Appl.* **27**, 891–908.

Q. Ye (1995), On close eigenvalues of tridiagonal matrices, *Numer. Math.* **70**, 507–514.

J.-P. M. Zemke (2003), Krylov subspace methods in finite precision: a unified approach, PhD thesis, Technical University of Hamburg–Harburg.