Electronic Transactions on Numerical Analysis.
Volume 13, pp. 56-80, 2002.

ISSN 1068-9613.

ETNA
Kent State University
etna@mcs.kent.edu

ON ERROR ESTIMATION IN THE CONJUGATE GRADIENT METHOD AND WHY IT WORKS IN FINITE PRECISION COMPUTATIONS *

ZDENĚK STRAKOŠ† AND PETR TICHÝ*

Abstract. In their paper published in 1952, Hestenes and Stiefel considered the conjugate gradient (CG) method an iterative method which terminates in at most n steps if no rounding errors are encountered [24, p. 410]. They also proved identities for the A-norm and the Euclidean norm of the error which could justify the stopping criteria [24, Theorems 6:1 and 6:3, p. 416]. The idea of estimating errors in iterative methods, and in the CG method in particular, was independently (of these results) promoted by Golub; the problem was linked to Gauss quadrature and to its modifications [7], [8]. A comprehensive summary of this approach was given in [15], [16]. During the last decade several papers developed error bounds algebraically without using Gauss quadrature. However, we have not found any reference to the corresponding results in [24]. All the existing bounds assume exact arithmetic. Still they seem to be in a striking agreement with finite precision numerical experiments, though in finite precision computations they estimate quantities which can be orders of magnitude different from their exact precision counterparts! For the lower bounds obtained from Gauss quadrature formulas this nontrivial phenomenon was explained, with some limitations, in [17].

In our paper we show that the lower bound for the A-norm of the error based on Gauss quadrature ([15], [17], [16]) is mathematically equivalent to the original formula of Hestenes and Stiefel [24]. We will compare existing bounds and we will demonstrate necessity of a proper rounding error analysis: we present an example of the well-known bound which can fail in finite precision arithmetic. We will analyse the simplest bound based on [24, Theorem 6:1], and prove that it is numerically stable. Though we concentrate mostly on the lower bound for the A-norm of the error, we describe also an estimate for the Euclidean norm of the error based on [24, Theorem 6:3]. Our results are illustrated by numerical experiments.

Key words. conjugate gradient method, Gauss quadrature, evaluation of convergence, error bounds, finite precision arithmetic, rounding errors, loss of orthogonality.

AMS subject classifications. 15A06, 65F10, 65F25, 65G50.

1. Introduction.

Consider a symmetric positive definite matrix $A \in \mathbb{R}^{n\times n}$ and a right-hand side vector $b \in \mathbb{R}^n$ (for simplicity of notation we will assume A, b real; generalization to complex data will be obvious). This paper investigates numerical estimation of errors in iterative methods for solving linear systems

$$Ax = b. \tag{1.1}$$

In particular, we focus on the conjugate gradient method (CG) of Hestenes and Stiefel [24] and on the lower estimates of the A-norm (also called the energy norm) of the error, which has important meaning in physics and quantum chemistry, and plays a fundamental role in evaluating convergence [1], [2].

Starting with the initial approximation x_0, the conjugate gradient approximations are determined by the condition

$$\begin{aligned} x_j &\in x_0 + \mathcal{K}_j(A, r_0) \\ \|x - x_j\|_A &= \min_{u \in x_0 + \mathcal{K}_j(A,r_0)} \|x - u\|_A, \end{aligned} \tag{1.2}$$

i.e. they minimize the A-norm of the error

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

*Received June 7, 2001. Accepted for publication August 18, 2002. Recommended by L. Reichel.

†Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 2, 182 07 Praha 8, Czech Republic. E-mail: strakos@cs.cas.cz and petr.tichy@centrum.cz. This research was supported by the Grant Agency of the Czech Republic under grant No. 201/02/0595.

over all methods generating approximations in the manifold $x_0 + \mathcal{K}_j(A, r_0)$. Here

$$\mathcal{K}_j(A, r_0) = \text{span}\{r_0, Ar_0, \dots A^{j-1}r_0\}$$

is the j-th Krylov subspace generated by A with the initial residual r_0, $r_0 = b - Ax_0$, and x is the solution of (1.1). The standard implementation of the CG method was given in [24, (3:1a)-(3:1f)]:

Given x_0, $r_0 = b - Ax_0$, $p_0 = r_0$, and for $j = 1, 2, \dots$, let

$$\begin{aligned} \gamma_{j-1} &= (r_{j-1}, r_{j-1})/(p_{j-1}, Ap_{j-1}), \\ x_j &= x_{j-1} + \gamma_{j-1}\, p_{j-1}, \\ r_j &= r_{j-1} - \gamma_{j-1}\, Ap_{j-1}, \\ \delta_j &= (r_j, r_j)/(r_{j-1}, r_{j-1}), \\ p_j &= r_j + \delta_j\, p_{j-1}. \end{aligned} \tag{1.3}$$

The residual vectors $\{r_0, r_1, \dots, r_{j-1}\}$ form an orthogonal basis and the direction vectors $\{p_0, p_1, \dots, p_{j-1}\}$ an A-orthogonal basis of the j-th Krylov subspace $\mathcal{K}_j(A, r_0)$.

In [24] Hestenes and Stiefel considered CG as an iterative procedure. They presented relations [24, (6:1)-(6:3) and (6:5), Theorems 6:1 and 6:3] as justifications of a possible stopping criterion for the algorithm. In our notation these relations become

$$\|x - x_{j-1}\|_A^2 - \|x - x_j\|_A^2 \;=\; \gamma_{j-1}\|r_{j-1}\|^2, \tag{1.4}$$

$$\|x - x_j\|_A^2 - \|x - x_k\|_A^2 \;=\; \sum_{i=j}^{k-1} \gamma_i \|r_i\|^2, \quad 0 \le j < k \le n, \tag{1.5}$$

$$x_j = x_0 + \sum_{l=0}^{j-1} \gamma_l p_l \;=\; x_0 + \sum_{l=0}^{j-1} \frac{\|x - x_l\|_A^2 - \|x - x_j\|_A^2}{\|r_l\|^2}\, r_l, \tag{1.6}$$

$$\|x - x_{j-1}\|^2 - \|x - x_j\|^2 \;=\; \frac{\|x - x_{j-1}\|_A^2 + \|x - x_j\|_A^2}{\mu(p_{j-1})}, \qquad \mu(p_{j-1}) = \frac{(p_{j-1}, Ap_{j-1})}{\|p_{j-1}\|^2}. \tag{1.7}$$

Please note that (1.5) represents an identity describing the decrease of the A-norm of the error in terms of quantities available in the algorithm, while (1.7) describes decrease of the Euclidean norm of the error in terms of the A-norm of the error in the given steps.

Hestenes and Stiefel did not give any particular stopping criterion. They emphasized, however, that while the A-norm of the error and the Euclidean norm of the error had to decrease monotonically at each step, the residual norm oscillated and might even increase in each but the last step. An example of this behaviour was used in [23].

The paper [24] is frequently referenced, but some of its results has not been paid much attention. Residual norms have been (and still are) commonly used for evaluating convergence of CG. The possibility of using (1.4)–(1.7) for constructing a stopping criterion has not been, to our knowledge, considered.

An interest in estimating error norms in the CG method reappeared with works of Golub and his collaborators. Using some older results [7], Dahlquist, Golub and Nash [8] related error bounds to Gauss quadrature (and to its modifications). The approach presented in that paper became a basis for later developments. It is interesting to note that the relationship of the CG method to the Riemann-Stieltjes integral and Gauss quadrature was described in detail in [24, Section 14], but without any link to error estimation. The work of Golub and his collaborators was independent of [24].

The paper [8] brought also into attention an important issue of rounding errors. The authors noted that in order to guarantee the numerical stability of the computed Gauss quadrature nodes and weights, the computed basis vectors had to be reorthogonalized. That means that the authors of that paper were from the very beginning aware of the fact that rounding errors might play a significant role in the application of their bounds to practical computations. In the numerical experiments used in [8] the effect of rounding errors were, however, not noticeable. This can be explained using the results by Paige ([33], [34], [35] and [36]). Due to the distribution of eigenvalues of the matrix used in [8] Ritz values do not converge to the eigenvalues until the last few steps. Before this convergence takes place there is no significant loss of orthogonality and the effects of rounding errors are not visible.

Error bounds in iterative methods were intensively studied or used in many later papers and in several books, see, e.g. [9], [10], [12], [15], [17], [16], [11], [21], [28], [29], [30], [4], [6]. Except for [17], effects of rounding errors were not analysed in these publications.

Frommer and Weinberg [13] pointed out the problem of applying exact precision formulas to finite precision computations, and proposed to use interval arithmetic for computing verified error bounds. As stated in [13, p. 201], this approach had serious practical limitations. Axelsson and Kaporin [3] considered preconditioned conjugate gradients and presented (1.5) independently of [24]. Their derivation used (global) mutual A-orthogonality among the direction vectors p_j, $j = 0, 1, \ldots, n-1$. They noticed that the numerical values found from the resulting estimate were identical to those obtained from Gauss quadrature, but did not prove this coincidence. They also noticed the potential difficulty due to rounding errors. They presented an observation that loss of orthogonality did not destroy applicability of their estimate. Calvetti et al. [5] presented several bounds and estimates for the A-norm of the error, and addressed a problem of cancellation in their computations [5, relation (46)].

In our paper we briefly recall some error estimates published after (and independently of) (1.4)-(1.7) in [24]. For simplicity of our exposition we will concentrate mostly on the A-norm of the error $\|x - x_j\|_A$. We will show that the simplest possible estimate for $\|x-x_j\|_A$, which follows from the relation (1.4) published in the original paper [24], is mathematically (in exact arithmetic) equivalent to the corresponding bounds developed later. In finite precision arithmetic, rounding errors in *the whole computation, not only in the computation of the convergence bounds*, must be taken into account. We emphasize that rounding error analysis of formulas for computation of the convergence bounds represents in almost all cases a simple and unimportant part of the problem. Almost all published convergence bounds (including those given in [5]) can be computed accurately (i.e. computation of the bounds using given formulas is not significantly affected by rounding errors). But this does not prove that these bounds give anything reasonable when they are applied to finite precision CG computations. We will see an example of the accurately computed bound which gives no useful information about the convergence of CG in finite precision arithmetic in Section 6.

An example of rounding error analysis for the bounds based on Gauss quadrature was presented in [17]. The results from [17] rely on the work by Paige and Greenbaum ([36], [19] and [22]). Though [17] gives a strong qualitative justification of the bounds in finite precision arithmetic, this justification is applicable only until $\|x - x_j\|_A$ reaches the square root of the machine precision. Moreover, quantitative expressions for the rounding error terms are very complicated. They contain factors which are not tightly estimated (see [19], [22]). Here we complement the analysis from [17] by substantially stronger results. We prove that the simplest possible lower bound for $\|x - x_j\|_A$ based on (1.4) works also for numerically

ETNA
Kent State University
etna@mcs.kent.edu

computed quantities till $\|x - x_j\|_A$ reaches its ultimate attainable accuracy.

The paper is organized as follows. In Section 2 we briefly describe relations between the CG and Lanczos methods. Using the orthogonality of the residuals, these algorithms are related to sequences of orthogonal polynomials, where the inner product is defined by a Riemann-Stieltjes integral with some particular distribution function $\omega(\lambda)$. The value of the j-th Gauss quadrature approximation to this Riemann-Stieltjes integral for the function $1/\lambda$ is the complement to the error in the j-th iteration of the CG method measured by $\|x - x_j\|_A^2/\|r_0\|^2$. In Section 3 we reformulate the result of the Gauss quadrature using quantities that are at our disposal during the CG iterations. In Section 4 we use the identities from Section 3 for estimation of the A-norm of the error in the CG method, and we compare the main existing bounds. Section 5 describes delay of convergence due to rounding errors. Section 6 explains why applying exact precision convergence estimates to finite precision CG computations represents a serious problem which must be properly addressed. Though exact precision CG and finite precision CG can dramatically differ, some exact precision bounds seem to be in good agreement with the finite precision computations. Sections 7–10 explain this paradox. The individual terms in the identities which the convergence estimates are based on can be strongly affected by rounding errors. *The identities as a whole, however, hold true (with small perturbations) also in finite precision arithmetic*. Numerical experiments are presented in Section 11.

When it will be helpful we will use the word “ideally” (or “mathematically”) to refer to a result that would hold using exact arithmetic, and “computationally” or “numerically” to a result of a finite precision computation.

2. Method of conjugate gradients and Gauss quadrature. For A and r_0 the Lanczos method [27] generates ideally a sequence of orthonormal vectors $v_1, v_2, \dots$ via the recurrence

Given $v_1 = r_0/\|r_0\|$, $\beta_1 \equiv 0$, and for $j = 1, 2, \dots$, let

$$\begin{aligned} \alpha_j &= (Av_j - \beta_j v_{j-1}, v_j), \\ w_j &= Av_j - \alpha_j v_j - \beta_j v_{j-1}, \\ \beta_{j+1} &= \|w_j\|, \\ v_{j+1} &= w_j/\beta_{j+1}. \end{aligned} \tag{2.1}$$

Denoting by $V_j = [v_1, \dots, v_j]$ the n by j matrix having the Lanczos vectors $\{v_1, \dots, v_j\}$ as its columns, and by T_j the symmetric tridiagonal matrix with positive subdiagonal

$$T_j = \begin{pmatrix} \alpha_1 & \beta_2 & & \\ \beta_2 & \alpha_2 & \ddots & \\ & \ddots & \ddots & \beta_j \\ & & \beta_j & \alpha_j \end{pmatrix} \tag{2.2}$$

the formulas (2.1) are written in the matrix form

$$AV_j = V_j T_j + \beta_{j+1} v_{j+1} e_j^T, \tag{2.3}$$

where e_j is the j-th column of the n by n identity matrix. Comparing (1.3) with (2.1) gives

$$v_{j+1} = (-1)^j \frac{r_j}{\|r_j\|}, \tag{2.4}$$

and also relations between the recurrence coefficients:

$$\alpha_j = \frac{1}{\gamma_{j-1}} + \frac{\delta_{j-1}}{\gamma_{j-2}}, \qquad \delta_0 \equiv 0,\ \gamma_{-1} \equiv 1,$$
$$\beta_{j+1} = \frac{\sqrt{\delta_j}}{\gamma_{j-1}}\,. \tag{2.5}$$

Finally, using the change of variables

$$x_j = x_0 + V_j\, y_j, \tag{2.6}$$

and the orthogonality relation between r_j and the basis $\{v_1, v_2, \ldots, v_j\}$ of $\mathcal{K}_j(A, r_0)$, we see that

$$\begin{aligned} 0 = V_j^T r_j &= V_j^T(b - Ax_j) = V_j^T(r_0 - AV_j\, y_j) \\ &= e_1\|r_0\| - V_j^T A V_j\, y_j = e_1\|r_0\| - T_j\, y_j\,. \end{aligned}$$

Ideally, the CG approximate solution x_j can therefore be determined by solving

$$T_j\, y_j = e_1\|r_0\|\,, \tag{2.7}$$

with subsequent using of (2.6).

Orthogonality of the CG residuals creates the elegance of the CG method which is represented by its link to the world of classical orthogonal polynomials. Using (1.3), the j-th error resp. residual can be written as a polynomial in the matrix A applied to the initial error resp. residual,

$$x - x_j = \varphi_j(A)\ (x - x_0), \qquad r_j = \varphi_j(A)\ r_0, \qquad \varphi_j \in \Pi_j, \tag{2.8}$$

where Π_j denotes the class of polynomials of degree at most j having the property $\varphi(0) = 1$ (that is, the constant term equal to one). Consider the eigendecomposition of the symmetric matrix A in the form

$$A = U\Lambda U^T, \qquad UU^T = U^T U = I, \tag{2.9}$$

where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ and $U = [u_1, \ldots, u_n]$ is the matrix having the normalized eigenvectors of A as its columns. Substituting (2.9) and (2.8) into (1.2) gives

$$\begin{aligned} \|x - x_j\|_A = \|\varphi_j(A)(x - x_0)\|_A &= \min_{\varphi\in\Pi_j} \|\varphi(A)(x - x_0)\|_A = \min_{\varphi\in\Pi_j} \|\varphi(A) r_0\|_{A^{-1}} \\ &= \min_{\varphi\in\Pi_j} \left\{ \sum_{i=1}^{n} \frac{(r_0, u_i)^2}{\lambda_i}\, \varphi^2(\lambda_i) \right\}^{1/2}. \end{aligned} \tag{2.10}$$

Consequently, for A symmetric positive definite the rate of convergence of CG is determined by the distribution of eigenvalues of A and by the size of the components of r_0 in the direction of the individual eigenvectors.

Similarly to (2.8), v_{j+1} is linked with some monic polynomial ψ_j,

$$v_{j+1} = \psi_j(A)\, v_1 \cdot \frac{1}{\beta_2\beta_3 \ldots \beta_{j+1}}\,. \tag{2.11}$$

Using the orthogonality of v_{j+1} to $v_1, \ldots, v_j$, the polynomial ψ_j is determined by the minimizing condition

$$\|\psi_j(A)v_1\| = \min_{\psi \in \mathcal{M}_j} \|\psi(A)v_1\| = \min_{\psi \in \mathcal{M}_j} \left\{ \sum_{i=1}^{n} (v_1, u_i)^2 \, \psi^2(\lambda_i) \right\}^{1/2}, \tag{2.12}$$

where $\mathcal{M}_j$ denotes the class of monic polynomials of degree j.

We will explain what we consider the essence of the CG and Lanczos methods.

Whenever the CG or the Lanczos method (defined by (1.3) resp. by (2.1)) is considered, there is a sequence $1, \psi_1, \psi_2, \ldots$ of the monic orthogonal polynomials determined by (2.12). These polynomials are orthogonal with respect to the discrete inner product

$$(f, g) = \sum_{i=1}^{n} \omega_i f(\lambda_i) g(\lambda_i)\,, \tag{2.13}$$

where the weights ω_i are determined as

$$\omega_i = (v_1, u_i)^2, \qquad \sum_{i=1}^{n} \omega_i = 1\,, \tag{2.14}$$

($v_1 = r_0/\|r_0\|$). For simplicity of notation we assume that all the eigenvalues of A are distinct and increasingly ordered (an extension to the case of multiple eigenvalues will be obvious). Let ζ, ξ be such that $\zeta \le \lambda_1 < \lambda_2 < \ldots < \lambda_n \le \xi$. Consider the distribution function $\omega(\lambda)$ with the finite points of increase $\lambda_1, \lambda_2, \ldots, \lambda_n$,

$$\begin{aligned} \omega(\lambda) &= 0 && \text{for} && \lambda < \lambda_1\,, \\ \omega(\lambda) &= \sum_{l=1}^{i} \omega_l && \text{for} && \lambda_i \le \lambda < \lambda_{i+1}\,, \\ \omega(\lambda) &= 1 && \text{for} && \lambda_n \le \lambda\,, \end{aligned} \tag{2.15}$$

see Fig. 2.1, and the corresponding Riemann-Stieltjes integral

$$\int_{\zeta}^{\xi} f(\lambda)\, d\omega(\lambda) = \sum_{i=1}^{n} \omega_i f(\lambda_i)\,. \tag{2.16}$$

Then (2.12) can be rewritten as

$$\psi_j = \arg\min_{\psi \in \mathcal{M}_j} \left\{ \int_{\zeta}^{\xi} \psi^2(\lambda)\, d\omega(\lambda) \right\}, \qquad j = 0, 1, 2, \ldots, n\,. \tag{2.17}$$

The j steps of the CG resp. the Lanczos method starting with $\|r_0\|v_1$ resp. v_1 determine a symmetric tridiagonal matrix (with a positive subdiagonal) T_j (2.2). Consider, analogously to (2.9), the eigendecomposition of T_j in the form

$$T_j = S_j \Theta_j S_j^T\,, \qquad S_j^T S_j = S_j S_j^T = I, \tag{2.18}$$

$\Theta_j = \mathrm{diag}(\theta_1^{(j)}, \ldots, \theta_j^{(j)})$, $S_j = [s_1^{(j)}, \ldots, s_j^{(j)}]$. Please note that we can look at T_j also as determined by the CG or the Lanczos method applied to the j-dimensional problem $T_j y_j = e_1\|r_0\|$ resp. T_j with initial residual $e_1\|r_0\|$ resp. starting vector e_1. Clearly, we can construct

ETNA
Kent State University
etna@mcs.kent.edu

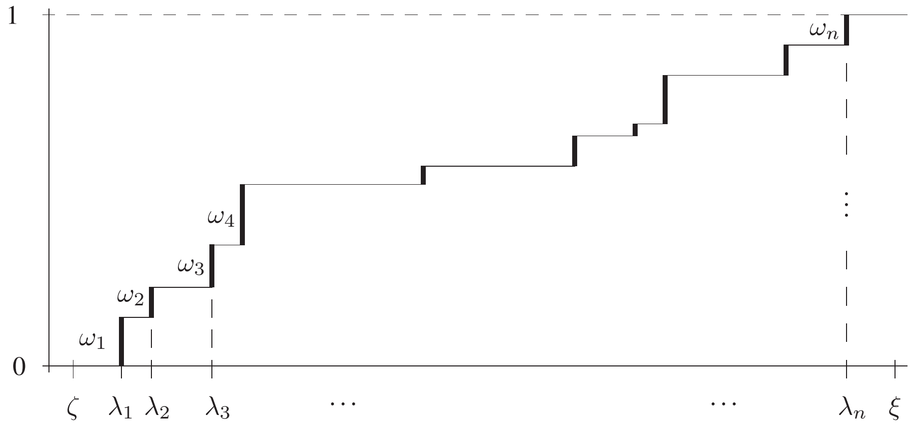

FIG. 2.1. *Distribution function* $\omega(\lambda)$

Riemann-Stieltjes integral for this j-dimensional problem similarly as above. Let $\zeta \leq \theta_1^{(j)} < \theta_2^{(j)} < \ldots < \theta_j^{(j)} \leq \xi$ be the eigenvalues of T_j (Ritz values, they must be distinct, see, e.g. [38, Chapter 7]). Let

$$\omega_i^{(j)} = (e_1, s_i^{(j)})^2, \qquad \sum_{i=1}^{j} \omega_i^{(j)} = 1 \tag{2.19}$$

be the weights determined by the squared size of the components of e_1 in the direction of T_j's eigenvectors, and

$$\begin{aligned} \omega^{(j)}(\lambda) &= 0 && \text{for} && \lambda < \theta_1^{(j)}\,, \\ \omega^{(j)}(\lambda) &= \sum_{l=1}^{i} \omega_l^{(j)} && \text{for} && \theta_i^{(j)} \leq \lambda < \theta_{i+1}^{(j)}, \\ \omega^{(j)}(\lambda) &= 1 && \text{for} && \theta_j^{(j)} \leq \lambda\,. \end{aligned}$$

Then the first j polynomials from the set $\{1, \psi_1, \ldots, \psi_n\}$ determined by (2.17) are also determined by the condition based on the Riemann-Stieltjes integral with the distribution function $\omega^{(j)}(\lambda)$

$$\psi_l = \arg \min_{\psi \in \mathcal{M}_l} \left\{ \int_\zeta^\xi \psi^2(\lambda)\, d\omega^{(j)}(\lambda) \right\}, \qquad l = 0, 1, \ldots, j\,, \tag{2.20}$$

(we can look at the subsequence $\{1, \psi_1, \ldots, \psi_j\}$ as determined by the CG or the Lanczos method applied to the j-dimensional problem described above). The integral

$$\int_\zeta^\xi f(\lambda)\, d\omega^{(j)}(\lambda) = \sum_{i=1}^{j} \omega_i^{(j)} f(\theta_i^{(j)}) \tag{2.21}$$

is the well-known j-th Gauss quadrature approximation of the integral (2.16), see, e.g., [14]. Thus, the CG and Lanczos methods determine the sequence of distribution functions $\omega^{(1)}(\lambda), \omega^{(2)}(\lambda), \ldots, \omega^{(j)}(\lambda), \ldots$ approximating in an optimal way (in the sense of Gauss quadrature, i.e. $\omega^{(l)}(\lambda)$ ensures that for any polynomial of degree less than of equal to $2l - 1$ the value of the original integral (2.16) is approximated by (2.21) exactly) the original distribution function $\omega(\lambda)$, cf. [26], [46, Chapter XV], [45].

ETNA
Kent State University
etna@mcs.kent.edu

All this is well-known. Gauss quadrature represents a classical textbook material and the connection of CG to Gauss quadrature was pointed out in the original paper [24]. This connection is, however, a key to understanding both mathematical properties and finite precision behaviour of the CG method.

Given A and r_0, (2.16) and its Gauss quadrature approximations (2.21) are for $j = 1, 2, \ldots, n$ uniquely determined (remember we assumed that the eigenvalues of A are positive and distinct). Conversely, the distribution function $\omega^{(j)}(\lambda)$ uniquely determines the symmetric tridiagonal matrix T_j, and, through (2.7) and (2.6), the CG approximation x_j. With $f(\lambda) = \lambda^{-1}$ we have from (2.10)

$$\|x - x_0\|_A^2 = \|r_0\|^2 \sum_{i=1}^{n} \frac{\omega_i}{\lambda_i} = \|r_0\|^2 \int_\zeta^\xi \lambda^{-1}\, d\omega(\lambda)\,, \tag{2.22}$$

and, using (2.3) with $j = n$,

$$\|x - x_0\|_A^2 = (r_0, A^{-1} r_0) = \|r_0\|^2 (e_1, T_n^{-1} e_1) \equiv \|r_0\|^2\, (T_n^{-1})_{11}\,.$$

Consequently,

$$\int_\zeta^\xi \lambda^{-1}\, d\omega(\lambda) = (T_n^{-1})_{11}\,. \tag{2.23}$$

Repeating the same considerations using the CG method for T_j with the initial residual $\|r_0\| e_1$, or the Lanczos method for T_j with e_1

$$\int_\zeta^\xi \lambda^{-1}\, d\omega^{(j)}(\lambda) = (T_j^{-1})_{11}\,. \tag{2.24}$$

Finally, applying the j-point Gauss quadrature to (2.16) gives

$$\int_\zeta^\xi f(\lambda)\, d\omega(\lambda) = \int_\zeta^\xi f(\lambda)\, d\omega^{(j)}(\lambda) + R_j(f), \tag{2.25}$$

where $R_j(f)$ stands for the (truncation) error in the Gauss quadrature. In the next section we present several different ways of expressing (2.25) with $f(\lambda) = \lambda^{-1}$.

3. Basic Identities.

Multiplying the identity (2.25) by $\|r_0\|^2$ gives

$$\|r_0\|^2 \int_\zeta^\xi f(\lambda)\, d\omega(\lambda) \;=\; \|r_0\|^2 \int_\zeta^\xi f(\lambda)\, d\omega^{(j)}(\lambda) + \|r_0\|^2 R_j(f). \tag{3.1}$$

Using (2.22), (2.23) and (2.24), (3.1) can for $f(\lambda) = \lambda^{-1}$ be written as

$$\|x - x_0\|_A^2 \;=\; \|r_0\|^2 (T_n^{-1})_{11} \;=\; \|r_0\|^2 (T_j^{-1})_{11} + \|r_0\|^2 R_j(\lambda^{-1})\,.$$

In [17, pp. 253-254] it was proved that for $f(\lambda) = \lambda^{-1}$ the truncation error in the Gauss quadrature is equal to

$$R_j(\lambda^{-1}) = \frac{\|x - x_j\|_A^2}{\|r_0\|^2}\,,$$

which gives

$$\|x - x_0\|_A^2 \;=\; \|r_0\|^2 (T_j^{-1})_{11} + \|x - x_j\|_A^2 \,. \tag{3.2}$$

Summarizing, the value of the j-th Gauss quadrature approximation to the integral (2.23) is the complement of the error in the j-th CG iteration measured by $\|x - x_j\|_A^2 / \|r_0\|^2$,

$$\frac{\|x - x_0\|_A^2}{\|r_0\|^2} \;=\; j\text{-point Gauss quadrature} \;+\; \frac{\|x - x_j\|_A^2}{\|r_0\|^2}. \tag{3.3}$$

This relation was developed in [8] in the context of moments; it was a subject of extensive work motivated by estimation of the error norms in CG in the papers [12], [15] and [17]. Work in this direction continued and led to the papers [16], [28], [30], [5].

An interesting form of (3.2) was noticed by Warnick in [47]. In the papers mentioned above the values of $\|x - x_0\|_A^2 / \|r_0\|^2 = (T_n^{-1})_{11}$ and $(T_j^{-1})_{11}$ were approximated from the actual Gauss quadrature calculations (or from the related recurrence relations). Using (2.7) and (2.6), the identities

$$\begin{aligned} \|r_0\|^2 (T_j^{-1})_{11} &= \|r_0\| \, e_1^T \, T_j^{-1} \, e_1 \|r_0\| \\ &= \|r_0\| \, v_1^T V_j \, T_j^{-1} e_1 \|r_0\| = (\|r_0\| v_1)^T \left(V_j T_j^{-1} e_1 \|r_0\| \right) \\ &= r_0^T (x_j - x_0) \end{aligned}$$

show that $(T_j^{-1})_{11}$ is given by a simple inner product. Indeed,

$$\|x - x_0\|_A^2 \;=\; r_0^T (x_j - x_0) + \|x - x_j\|_A^2 \,. \tag{3.4}$$

This remarkable identity was pointed out to us by Saylor [41], [40]. Please note that derivation of the identity (3.4) from the Gauss quadrature-based (3.2) uses the orthogonality relation $v_1^T V_j = e_1$. In finite precision computations this orthogonality relation does not hold. Consequently, (3.4) does not hold in finite precision arithmetic. We will return to this point in Section 6.

A mathematically equivalent identity can be derived by simple algebraic manipulations without using Gauss quadrature,

$$\begin{aligned} (x - x_0)^T A (x - x_0) &= (x - x_j + x_j - x_0)^T A (x - x_0) \\ &= (x - x_j)^T A (x - x_0) + (x_j - x_0)^T A (x - x_0) \\ &= (x - x_j)^T A (x - x_j + x_j - x_0) + (x_j - x_0)^T r_0 \\ &= \|x - x_j\|_A^2 + (x - x_j)^T A (x_j - x_0) + r_0^T (x_j - x_0) \\ &= \|x - x_j\|_A^2 + r_j^T (x_j - x_0) + r_0^T (x_j - x_0), \end{aligned}$$

hence

$$\|x - x_0\|_A^2 \;=\; r_j^T (x_j - x_0) + r_0^T (x_j - x_0) + \|x - x_j\|_A^2. \tag{3.5}$$

The right-hand side of (3.5) contains, in comparison with (3.4), the additional term $r_j^T (x_j - x_0)$. This term is in exact arithmetic equal to zero, but it has an important correction effect in finite precision computations (see Section 6).

Relations (3.2), (3.4) and (3.5) represent various mathematically equivalent forms of (3.1). While in (3.2) the j-point Gauss quadrature is evaluated as $(T_j^{-1})_{11}$, in (3.4) and (3.5) this quantity is computed using inner products of the vectors that are at our disposal during the iteration process. But, as mentioned in Introduction, there is much simpler identity (1.5) mathematically equivalent to (3.1). It is very surprising that, though (1.5) is present in the Hestenes and Stiefel paper [24, Theorem 6.1, relation (6:2), p. 416], this identity has (at least to our knowledge) never been related to Gauss quadrature. Its derivation is very simple. Using (1.3)

$$\begin{aligned}\|x-x_i\|_A^2-\|x-x_{i+1}\|_A^2 &= \|x-x_{i+1}+x_{i+1}-x_i\|_A^2-\|x-x_{i+1}\|_A^2\\ &= \|x_{i+1}-x_i\|_A^2+2(x-x_{i+1})^TA(x_{i+1}-x_i)\\ &= \gamma_i^2\,p_i^TAp_i+2r_{i+1}^T(x_{i+1}-x_i)\\ &= \gamma_i\|r_i\|^2\,. \end{aligned}\tag{3.6}$$

Consequently, for $0\le l<j\le n$,

$$\|x-x_l\|_A^2-\|x-x_j\|_A^2=\sum_{i=l}^{j-1}\left(\|x-x_i\|_A^2-\|x-x_{i+1}\|_A^2\right)=\sum_{i=l}^{j-1}\gamma_i\|r_i\|^2,\tag{3.7}$$

and (3.1) can be written in the form

$$\|x-x_0\|_A^2 \;=\; \sum_{i=0}^{j-1}\gamma_i\|r_i\|^2+\|x-x_j\|_A^2.\tag{3.8}$$

The numbers $\gamma_i\|r_i\|^2$ are trivially computable; both γ_i and $\|r_i\|^2$ are available at every iteration step. Please note that in the derivation of (3.7) we used the local orthogonality among the consecutive residuals and direction vectors only. We avoided using mutual orthogonality among the vectors with generally different indices. This fact will be very important in the rounding error analysis of the finite precision counterparts of (3.7) in Sections 7–10.

4. Estimating the A-norm of the error. Using $\|x-x_0\|_A^2=\|r_0\|^2(T_n^{-1})_{11}$, (3.2) is written in the form

$$\|x-x_j\|_A^2 \;=\; \|r_0\|^2\left[(T_n^{-1})_{11}-(T_j^{-1})_{11}\right].$$

As suggested in [17, pp. 28–29], the unknown value $(T_n^{-1})_{11}$ can be replaced, at a price of $m-j$ extra steps, by a computable value $(T_m^{-1})_{11}$ for some $m>j$. The paper [17], however, did not properly use this idea and did not give a proper formula for computing the difference $(T_m^{-1})_{11}-(T_j^{-1})_{11}$ without cancellation, which limited the applicability of the proposed result. Golub and Meurant cleverly resolved this trouble in [16] and proposed an algorithm for estimating the A-norm of the error in the CG method called CGQL. This section will briefly summarize several important estimates.

Consider, in general, (3.1) for j and $j+d$, where d is some positive integer. The idea is simply to eliminate the unknown term $\int_\zeta^\xi f(\lambda)\,d\omega(\lambda)$ by subtracting the identities for j and $j+d$ which results in

$$\|r_0\|^2R_j(f)=\|r_0\|^2\left(\int_\zeta^\xi f(\lambda)\,d\omega^{(j+d)}(\lambda)-\int_\zeta^\xi f(\lambda)\,d\omega^{(j)}(\lambda)\right)+\|r_0\|^2R_{j+d}(f).$$

In particular, using (3.2), (3.4), (3.5), and (3.8) we obtain the mathematically equivalent identities

$$\|x-x_j\|_A^2 \;=\; \|r_0\|^2\left[(T_{j+d}^{-1})_{11}-(T_j^{-1})_{11}\right]+\|x-x_{j+d}\|_A^2\,,\tag{4.1}$$

$$\|x-x_j\|_A^2 \;=\; r_0^T(x_{j+d}-x_j)+\|x-x_{j+d}\|_A^2\,,\tag{4.2}$$

$$\begin{aligned}\|x-x_j\|_A^2 \;=\; & r_0^T(x_{j+d}-x_j)-r_j^T(x_j-x_0)+r_{j+d}^T(x_{j+d}-x_0)\\ & +\|x-x_{j+d}\|_A^2\,,\end{aligned}\tag{4.3}$$

and

$$\|x - x_j\|_A^2 \;=\; \sum_{i=j}^{j+d-1} \gamma_i \|r_i\|^2 + \|x - x_{j+d}\|_A^2 \,. \tag{4.4}$$

Now recall that the A-norm of the error is in the CG method strictly decreasing. If d is chosen such that

$$\|x - x_j\|_A^2 \gg \|x - x_{j+d}\|_A^2 \,, \tag{4.5}$$

then neglecting $\|x - x_{j+d}\|_A^2$ on the right-hand sides of (4.1), (4.2), (4.3) and (4.4) gives lower bounds (all mathematically equal) for the squared A-norm of the error in the j-th step. Under the assumption (4.5) these bounds are reasonably tight (their inaccuracy is given by $\|x - x_{j+d}\|_A^2$). We denote them

$$\eta_{j,d} \;=\; \|r_0\|^2 \; [(T_{j+d}^{-1})_{11} - (T_j^{-1})_{11}], \tag{4.6}$$

where the difference $(T_{j+d}^{-1})_{11} - (T_j^{-1})_{11}$ is computed by the algorithm CGQL from [16],

$$\mu_{j,d} \;=\; r_0^T (x_{j+d} - x_j), \tag{4.7}$$

which refers to the original bound due to Warnick,

$$\vartheta_{j,d} \;=\; r_0^T (x_{j+d} - x_j) - r_j^T (x_j - x_0) + r_{j+d}^T (x_{j+d} - x_0), \tag{4.8}$$

which is the previous bound modified by the correction terms and

$$\nu_{j,d} \;=\; \sum_{i=j}^{j+d-1} \gamma_i \|r_i\|^2. \tag{4.9}$$

Clearly, the last bound, which is a direct consequence of [24, Theorem 6:1], see (1.5), is much simpler than the others.

Mathematically (in exact arithmetic)

$$\eta_{j,d} = \mu_{j,d} = \vartheta_{j,d} = \nu_{j,d} \,. \tag{4.10}$$

In finite precision computations (4.10) does not hold in general, and the different bounds may give substantially different results. *Does any of the identities (4.1)–(4.4) have any relevance for the quantities computed in finite precision arithmetic?* The work described in this subsection and the papers published on this subject would be of little practical use without answering this question.

5. Delay of convergence. For more than 20 years the effects of rounding errors to the Lanczos and CG methods seemed devastating. Orthogonality among the computed vectors $v_1, v_2, \ldots$ was usually lost very quickly, with a subsequent loss of linear independence. Consequently, the finite termination property was lost. Still, despite a total loss of orthogonality among the vectors in the Lanczos sequence $v_1, v_2, \ldots$, and despite a possible regular appearance of Lanczos vectors which were linearly dependent on the vectors computed in preceding iterations, the Lanczos and the CG methods produced reasonable results.

A fundamental work which brought light into this darkness was done by Paige. He proved that loss of orthogonality among the computed Lanczos vectors $v_1, v_2, \ldots$ was possible only in the directions of the converged Ritz vectors $z_l^{(j)} \equiv V_j s_l^{(j)}$. For more details

ETNA
Kent State University
etna@mcs.kent.edu

see [33], [34], [35], [36], the review paper [44, Section 3.1] and the works quoted there (in particular [38], [39], [32] and [45]). Little was known about rounding errors in the Krylov subspace methods before the Ph.D. thesis of Paige [33], and almost all results (with the exception of works on ultimate attainable accuracy) published on the subject after this thesis and the papers [34], [35], [36] were based on them.

Another step, which can compete in originality with that of Paige, was made by Greenbaum in [19]. If CG is used to solve a linear symmetric positive definite system $Ax = b$ on a computer with machine precision ε, then [19] shows that the A-norms of the errors $\|x-x_l\|_A$, $l = 1, 2, \ldots, j$ are very close to the $\overline{A}$-norms of the errors $\|\overline{x} - \overline{x}_l\|_{\overline{A}}$, $l = 1, 2, \ldots, j$ determined by the *exact* CG applied to some particular symmetric positive definite system $\overline{A}(j)\overline{x}(j) = \overline{b}(j)$ (see [19, Theorem 3, pp. 26-27]). This system and the initial approximation $\overline{x}_0(j)$ depend on the iteration step j. The matrix $\overline{A}(j)$ is larger than the matrix A. Its eigenvalues must lie in tiny intervals about the eigenvalues of A, and there must be at least one eigenvalue of $\overline{A}(j)$ close to each eigenvalue of A (the last result was proved in [43]). Moreover, for each eigenvalue λ_i of A, $i = 1, \ldots, n$ (similarly to Section 2 we assume, with no loss of generality, that the eigenvalues of A are distinct), the weight $\omega_i = (v_1, u_i)^2$ closely approximates the sum of weights corresponding to the eigenvalues of $\overline{A}(j)$ clustered around λ_i (see [19, relation (8.21) on p. 60]).

The quantitative formulations of the relationships between A, b, x_0 and $\overline{A}(j)$, $\overline{b}(j)$, $\overline{x}_0(j)$ contains some terms related in various complicated ways to machine precision ε (see [19], [43] and [17, Theorems 5.1–5.3 and the related discussion on pp. 257–260]). The actual size of the terms given in the quoted papers documents much more difficulties of handling accurately peculiar technical problems of rounding error analysis than it says about the accuracy of the described relationships. The fundamental concept to which the (very often weak) rounding error bounds lead should be read: the first j steps of a *finite precision* CG computation for $Ax = b$ can be viewed as the first j steps of the *exact* CG computation for some particular $\overline{A}(j)\overline{x}(j) = \overline{b}(j)$. This relationship was developed and proved theoretically. Numerical experiments show that its tightness is much better than the technically complicated theoretical calculations in [19] would suggest. We will not continue with describing the results of the subsequent work [22]. We do not need it here. Moreover, a rigorous theoretical description of the model from [22] in the language of Riemann-Stieltjes integral and Gauss quadrature still needs some clarification. We hope to return to that subject elsewhere.

As a consequence of the loss of orthogonality caused by rounding errors, convergence of the CG method is delayed. In order to illustrate this important point numerically, we plot in Fig. 5.1 results of the CG method (1.3) for the matrix $A = Q\Lambda Q^T$, where Q is the orthogonal matrix obtained from the Matlab QR-decomposition of the randomly generated matrix (computed by the Matlab command `randn(n)`), and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ is a diagonal matrix with the eigenvalues

$$\lambda_i = \lambda_1 + \frac{i-1}{n-1}(\lambda_n - \lambda_1)\,\rho^{n-i}, \quad i = 2, \ldots, n-1, \tag{5.1}$$

see [43]. We have used $n = 48$, $\lambda_1 = 0.1$, $\lambda_n = 1000$, $\rho = 0.9$, $x = (1, \ldots, 1)^T$, $b = Ax$, and $x_0 = (0, \ldots, 0)^T$. We have simulated the exact arithmetic values by double reorthogonalization of the residual vectors (see [22]). The quantities obtained from the CG implementation with the double reorthogonalized residuals will be denoted by (E). Fig. 5.1 shows that when the double reorthogonalization is applied, the corresponding A-norm of the error (dash-dotted line) can be very different from the A-norm of the error of the ordinary finite precision (FP) CG implementation (solid line). Without reorthogonalization, the orthogonality among the (FP) Lanczos vectors, measured by the Frobenius norm $\|I - V_j^T V_j\|_F$ (dotted line), is

ETNA
Kent State University
etna@mcs.kent.edu

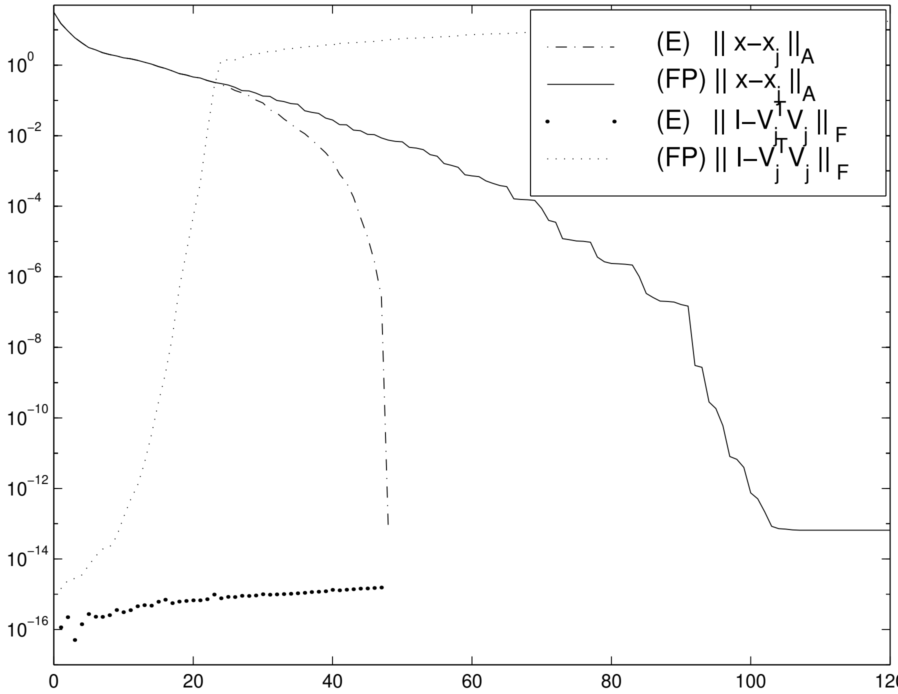

FIG. 5.1. *The A-norm of the error for the CG implementation with the double reorthogonalized residuals* (E) (*dashed-dotted line*) *is compared to the A-norm of the error of the ordinary finite precision CG implementation* (FP) (*solid line*). *The corresponding loss of orthogonality among the normalized residuals is plotted by the dots resp. the dotted line.*

lost after a few iterations. With double reorthogonalization the orthogonality is kept close to machine precision (dots). Experiments were performed using Matlab 5.1 on a personal computer with machine precision $\varepsilon \sim 10^{-16}$.

We see that the delay of convergence due to loss of orthogonality can be very substantial. Consider now application of the estimates (4.6)–(4.9) to finite precision computations. In derivation of all these estimates we assumed exact arithmetic. Consequently, in these derivations we did not count for any loss of orthogonality and delay of convergence. For the example presented above, the bounds can therefore be expected to give good results for the double reorthogonalized CG (dash-dotted convergence curve). Should they give anything reasonable also for the ordinary (FP) CG implementation (solid convergence curve)? If yes, then why? The following section explains that this question is of fundamental importance.

6. Examples. Indeed, without a proper rounding error analysis of the identities (4.1)–(4.4) there is no justification that the estimates derived assuming exact arithmetic will work in finite precision arithmetic. For example, when the significant loss of orthogonality occurs, the bound $\mu_{j,d}$ given by (4.7) does not work!

This fact is demonstrated in Fig. 6.1 which presents experimental results for the problem described in the previous section (see Fig. 5.1). It plots the computed estimate $|\mu_{j,d}|^{1/2}$ (dashed line) and demonstrates the importance of the correction term

$$c_{j,d} = -r_j^T(x_j - x_0) + r_{j+d}^T(x_{j+d} - x_0), \tag{6.1}$$

($|c_{j,d}|^{1/2}$ is plotted by dots). Fig. 6.1 shows clearly that when the global orthogonality (measured by $\|I - V_j^T V_j\|_F$ and plotted by a dotted line) grows greater than $\|x - x_j\|_A$ (solid

ETNA
Kent State University
etna@mcs.kent.edu

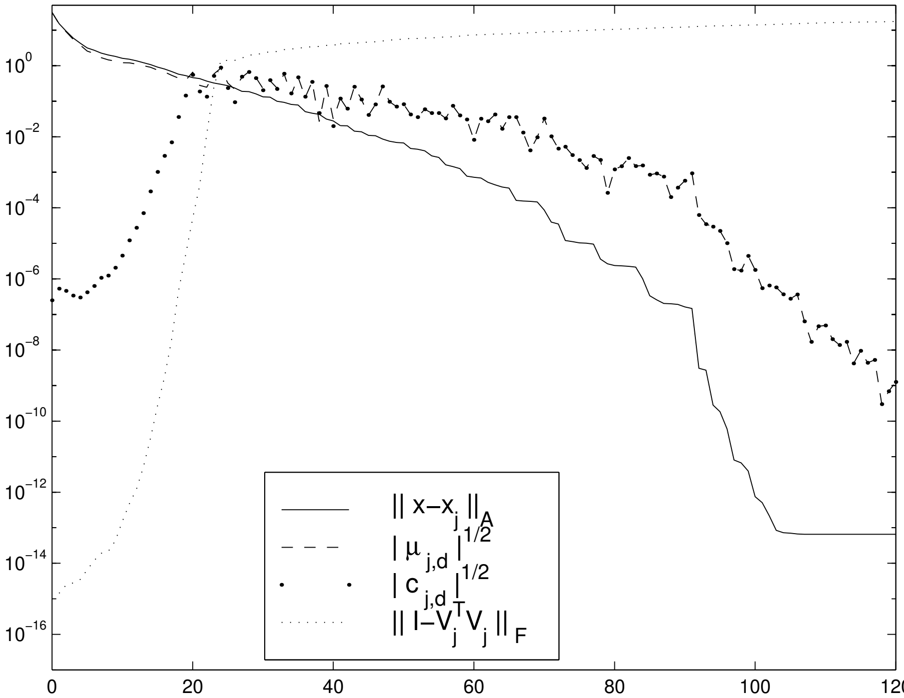

FIG. 6.1. *Error estimate* $\mu_{j,d}^{1/2}$ *can fail. The computed estimate* $|\mu_{j,d}|^{1/2}$ (*dashed line*) *for the* A*-norm of the error* (*solid line*) *gives useful information about convergence only until the loss of orthogonality* (*dotted line*) *crosses the convergence curve. After that point* $\mu_{j,d}$ *can even become negative, and must be modified by adding the correction term* $c_{j,d}$ ($|c_{j,d}|^{1/2}$ *is plotted by dots*). *We used* $d = 4$.

line), the bound ${\mu_{j,d}}^{1/2}$, which is based on global orthogonality, ceases to give any useful information about convergence ($\mu_{j,d}$ may even become negative, therefore we plot the second root of its absolute value). Adding the correction term $c_{j,d}$ to $\mu_{j,d}$ gives $\vartheta_{j,d}$, see (4.8), which gives estimates comparable to $\eta_{j,d}$ and $\nu_{j,d}$ (see Section 11). In this experiment we used $d = 4$.

It is important to understand that the additional rounding errors in computing $\eta_{j,d}$ $\mu_{j,d}$, $\vartheta_{j,d}$ and $\nu_{j,d}$ from the given formulas (the algorithm CGQL and (4.7)–(4.9)) do not affect significantly the values of the computed bounds and do not represent a problem. The problem is in the fact, that when the orthogonality is significantly lost, the input quantities used in the algorithm CGQL and in the formulas (4.7)–(4.9) are significantly different from their exact precision counterparts. These quantities affected by the loss of orthogonality are plugged into the formulas which assume, in their derivation, exact orthogonality.

In order to stress the previous point and to underline the necessity of rounding error analysis of the identities (4.7)–(4.9), we present the following analogous example. In the Lanczos method the eigenvalues $\theta_1^{(j)} < \theta_2^{(j)} < \ldots < \theta_j^{(j)}$ of T_j (Ritz values) are considered approximations to the eigenvalues of the matrix A (see Section 2). Let $\theta_l^{(j)}$, $z_l^{(j)} = V_j s_l^{(j)}$ (where $s_l^{(j)}$ is the normalized eigenvector of T_j corresponding to $\theta_l^{(j)}$) represents an approximate eigenpair of A. In exact arithmetic we have the following bound for the distance of $\theta_l^{(j)}$ to the nearest eigenvalue of A

$$\min_i |\lambda_i - \theta_l^{(j)}| \leq \frac{\|Az_l^{(j)} - \theta_l^{(j)} z_l^{(j)}\|}{\|z_l^{(j)}\|} = \|Az_l^{(j)} - \theta_l^{(j)} z_l^{(j)}\|, \tag{6.2}$$

where $\|z_l^{(j)}\| = 1$ due to the orthonormality of the Lanczos vectors $v_1, \ldots, v_j$. Using (2.3), $\|Az_l^{(j)} - \theta_l^{(j)} z_l^{(j)}\| = \beta_{j+1}(e_j, s_l^{(j)})$, which gives

$$\min_i |\lambda_i - \theta_l^{(j)}| \;\le\; \beta_{j+1}(e_j, s_l^{(j)}) \equiv \delta_{lj}, \tag{6.3}$$

see, e.g., [38], [36]. Consequently, in exact arithmetic, if δ_{lj} is small, then $\theta_l^{(j)}$ must be close to some λ_i. In finite precision arithmetic loss of orthogonality has, among the others, a very unpleasant effect: we cannot guarantee, in general, that $z_l^{(j)}$, which is a linear combination of $v_1, \ldots, v_j$ has a nonvanishing norm. We can still compute δ_{lj} from β_{j+1} and T_j; the effect of rounding errors in this additional computation is negligible. We can therefore say, similarly to the analogous statements published about computation of the convergence estimates in the CG method, that δ_{lj} is in the presence of rounding errors computed "accurately". Does δ_{lj} computed in finite precision arithmetic tell anything about convergence of $\theta_l^{(j)}$ to some λ_i? Yes, it does! But this affirmative answer is based neither on the exact precision formulas (6.2) and (6.3), nor on the fact that δ_{lj} is computed "accurately". It is based on an ingenious analysis due to Paige, who have shown that the orthogonality can be lost in the directions of the well approximated eigenvectors only. For the complicated details of this difficult result we refer to [33], [37] and to the summary given in [44, Theorem 2]. We see that even in finite precision computations small δ_{lj} guarantees that $\theta_l^{(j)}$ approximates some λ_i to high accuracy. It is very clear, however, that this conclusion is the result of the rounding error analysis of the Lanczos method given by Paige, and no similar statement could be made without this analysis.

In the following three sections we present rounding error analysis of the bound $\nu_{j,d}$ given by (4.4) and (4.9). We concentrate on $\nu_{j,d}$ because it is the simplest of all the others. If $\nu_{j,d}$ is proved numerically stable, then there is a small reason for using the other bounds $\eta_{j,d}$ or $\vartheta_{j,d}$ in practical computations.

7. Finite precision CG computations.

In the analysis we assume the standard model of floating point arithmetic with machine precision ε, see, e.g. [25, (2.4)],

$$\mathrm{fl}[a \circ b] = (a \circ b)(1+\delta), \qquad |\delta| \le \varepsilon, \tag{7.1}$$

where a and b stands for floating-point numbers and the symbol $\circ$ stands for the operations addition, subtraction, multiplication and division. We assume that this model holds also for the square root operation. Under this model, we have for operations involving vectors v, w, a scalar α and the matrix A the following standard results [18], see also [20], [35]

$$\|\alpha\, v - \mathrm{fl}[\alpha\, v]\| \le \varepsilon\, \|\alpha\, v\|, \tag{7.2}$$

$$\|v + w - \mathrm{fl}[v+w]\| \le \varepsilon\, (\|v\| + \|w\|), \tag{7.3}$$

$$|(v,w) - \mathrm{fl}[(v,w)]| \le \varepsilon\, n\, (1 + O(\varepsilon))\, \|v\|\, \|w\|, \tag{7.4}$$

$$\|Av - \mathrm{fl}[Av]\| \le \varepsilon\, c\, \|A\| \|v\|. \tag{7.5}$$

When A is a matrix with at most h nonzeros in any row and if the matrix-vector product is computed in the standard way, $c = hn^{1/2}$. In the following analysis we count only for the terms linear in the machine precision epsilon ε and express the higher order terms as $O(\varepsilon^2)$. By $O(const)$ where $const$ is different from ε^2 we denote $const$ multiplied by a bounded positive term of an insignificant size which is independent of the $const$ and of any other variables present in the bounds.

Numerically, the CG iterates satisfy

$$x_{j+1} = x_j + \gamma_j p_j + \varepsilon z_j^x, \tag{7.6}$$

ETNA
Kent State University
etna@mcs.kent.edu

$$r_{j+1} = r_j - \gamma_j A p_j + \varepsilon z_j^r, \tag{7.7}$$

$$p_{j+1} = r_{j+1} + \delta_{j+1} p_j + \varepsilon z_j^p, \tag{7.8}$$

where εz_j^x, εz_j^r and εz_j^p account for the local roundoff ($r_0 = b - Ax_0 - \varepsilon f_0$, $\varepsilon \|f_0\| \le \varepsilon\{\|b\| + \|Ax_0\| + c\|A\|\|x_0\|\} + O(\varepsilon^2)$). The local roundoff can be bounded according to the standard results (7.2)–(7.5) in the following way

$$\varepsilon \, \|z_j^x\| \le \varepsilon \, \{\|x_j\| + 2 \, \|\gamma_j p_j\|\} + O(\varepsilon^2) \le \varepsilon \, \{3\|x_j\| + 2\|x_{j+1}\|\} + O(\varepsilon^2), \tag{7.9}$$

$$\varepsilon \, \|z_j^r\| \le \varepsilon \, \{\|r_j\| + 2 \, \|\gamma_j A p_j\| + c \, \|A\| \|\gamma_j p_j\|\} + O(\varepsilon^2), \tag{7.10}$$

$$\varepsilon \, \|z_j^p\| \le \varepsilon \, \{\|r_{j+1}\| + 2 \, \|\delta_{j+1} p_j\|\} + O(\varepsilon^2) \le \varepsilon \, \{3\|r_{j+1}\| + 2\|p_{j+1}\|\} + O(\varepsilon^2). \tag{7.11}$$

Similarly, the computed coefficients γ_j and δ_j satisfy

$$\gamma_j = \frac{\|r_j\|^2}{p_j^T A p_j} + \varepsilon \zeta_j^\gamma, \quad \delta_j = \frac{\|r_j\|^2}{\|r_{j-1}\|^2} + \varepsilon \zeta_j^\delta. \tag{7.12}$$

Assuming $n\varepsilon \ll 1$, the local roundoff $\varepsilon \zeta_j^\delta$ is bounded, according to (7.1) and (7.4), by

$$\varepsilon |\zeta_j^\delta| \le \varepsilon \, \frac{\|r_j\|^2}{\|r_{j-1}\|^2} \, O(n) + O(\varepsilon^2). \tag{7.13}$$

Using (7.2)–(7.5) and $\|A\|\|p_j\|^2/(p_j, Ap_j) \le \kappa(A)$,

$$\begin{aligned} \mathrm{fl}[(p_j, Ap_j)] &= (p_j, Ap_j) + \varepsilon \, \|Ap_j\|\|p_j\| O(n) + \varepsilon \, \|A\|\|p_j\|^2 O(c) + O(\varepsilon^2) \\ &= (p_j, Ap_j)\big(1 + \varepsilon \, \kappa(A) O(n + c)\big) + O(\varepsilon^2). \end{aligned}$$

Assuming $\varepsilon(n + c)\, \kappa(A) \ll 1$, the local roundoff $\varepsilon \zeta_j^\gamma$ is bounded by

$$\varepsilon |\zeta_j^\gamma| \le \varepsilon \, \kappa(A) \frac{\|r_j\|^2}{(p_j, Ap_j)} \, O(n + c) + O(\varepsilon^2). \tag{7.14}$$

It is well-known that in finite precision arithmetic the true residual $b - Ax_j$ differs from the recursively updated residual vector r_j,

$$r_j = b - Ax_j - \varepsilon f_j. \tag{7.15}$$

This topic was studied in [42] and [20]. The results can be written in the following form

$$\|\varepsilon f_j\| \le \varepsilon \, \|A\| \, (\|x\| + \max_{0 \le i \le j} \|x_i\|) \, O(jc), \tag{7.16}$$

$$\|r_j\| = \|b - Ax_j\| \, (1 + \varepsilon F_j), \tag{7.17}$$

where εF_j is bounded by

$$|\varepsilon F_j| = \frac{|\|r_j\| - \|b - Ax_j\||}{\|b - Ax_j\|} \le \frac{\|r_j - (b - Ax_j)\|}{\|b - Ax_j\|} = \frac{\varepsilon \|f_j\|}{\|b - Ax_j\|}. \tag{7.18}$$

Rounding errors affect results of CG computations in two main ways: they delay convergence (see Section 5) and limit the ultimate attainable accuracy. Here we are primarily interested in estimating the convergence rate. We therefore assume that the final accuracy level has not been reached yet and εf_j is, in comparison to the size of the true and iterative residuals, small. In the subsequent text we will relate the numerical inaccuracies to the

ETNA
Kent State University
etna@mcs.kent.edu

A-norm of the error $\|x - x_j\|_A$. The following inequalities derived from (7.18) will prove useful,

$$\lambda_1^{1/2}\, \|x - x_j\|_A\, (1 + \varepsilon\, F_j) \leq \|r_j\| \leq \lambda_n^{1/2}\, \|x - x_j\|_A\, (1 + \varepsilon\, F_j). \tag{7.19}$$

The monotonicity of the A-norm and of the Euclidean norm of the error is in CG preserved (with small additional inaccuracy) also in finite precision computations (see [19], [22]). Using this fact we get for $j \geq i$

$$\varepsilon\, \frac{\|r_j\|}{\|r_i\|} \leq \varepsilon\, \frac{\lambda_n^{1/2}}{\lambda_1^{1/2}} \cdot \frac{\|x - x_j\|_A}{\|x - x_i\|_A} \cdot \frac{(1 + \varepsilon\, F_j)}{(1 + \varepsilon\, F_i)} \leq \varepsilon\, \kappa(A)^{1/2} + O(\varepsilon^2). \tag{7.20}$$

This bound will be used later.

8. Finite precision analysis – basic identity. The bounds (4.6)–(4.9) are mathematically equivalent. We will concentrate on the simplest one given by $\nu_{j,d}$ (4.9) and prove that it gives (up to a small term) correct estimates also in finite precision computations. In particular, we prove that the ideal (exact precision) identity (4.4) changes numerically to

$$\|x - x_j\|_A^2 = \nu_{j,d} + \|x - x_{j+d}\|_A^2 + \widetilde{\nu}_{j,d}, \tag{8.1}$$

where $\widetilde{\nu}_{j,d}$ is as small as it can be (the analysis here will lead to much stronger results than the analysis of the finite precision counterpart of (4.1) given in [17]). Please note that the difference between (4.4) and (8.1) *is not trivial*. The ideal and numerical counterparts of each individual term in these identities may be orders of magnitude different! Due to the facts that rounding errors in computing $\nu_{j,d}$ numerically from the quantities γ_i, r_i are negligible and that $\widetilde{\nu}_{j,d}$ will be related to $\varepsilon\, \|x - x_j\|_A$, (8.1) will justify the estimate $\nu_{j,d}$ in finite precision computations.

From the identity for the numerically computed approximate solution

$$\begin{aligned} \|x - x_j\|_A^2 &= \|x - x_{j+1} + x_{j+1} - x_j\|_A^2 \\ &= \|x - x_{j+1}\|_A^2 + 2\,(x - x_{j+1})^T A (x_{j+1} - x_j) + \|x_{j+1} - x_j\|_A^2, \end{aligned}$$

we obtain easily

$$\|x - x_j\|_A^2 - \|x - x_{j+1}\|_A^2 = \|x_{j+1} - x_j\|_A^2 + 2\,(x - x_{j+1})^T A (x_{j+1} - x_j). \tag{8.2}$$

Please note that (8.2) represents an identity for the computed quantities. In order to get the desired form leading to (8.1), we will develop the right hand side of (8.2). In this derivation we will rely on local properties of the finite precision CG recurrences (7.6)–(7.8) and (7.12).

Using (7.6), the first term on the right hand side of (8.2) can be written as

$$\begin{aligned} \|x_{j+1} - x_j\|_A^2 &= (\gamma_j p_j + \varepsilon\, z_j^x)^T A (\gamma_j p_j + \varepsilon\, z_j^x) \\ &= \gamma_j^2\, p_j^T A p_j + 2\varepsilon\, \gamma_j p_j^T A z_j^x + O(\varepsilon^2) \\ &= \gamma_j^2\, p_j^T A p_j + 2\varepsilon\, (x_{j+1} - x_j)^T A z_j^x + O(\varepsilon^2). \end{aligned} \tag{8.3}$$

Similarly, the second term on the right hand side of (8.2) transforms, using (7.15), to the form

$$\begin{aligned} 2\,(x - x_{j+1})^T A (x_{j+1} - x_j) &= 2\,(r_{j+1} + \varepsilon\, f_{j+1})^T (x_{j+1} - x_j) \\ &= 2\, r_{j+1}^T (x_{j+1} - x_j) + 2\varepsilon\, f_{j+1}^T (x_{j+1} - x_j). \end{aligned} \tag{8.4}$$

ETNA
Kent State University
etna@mcs.kent.edu

Combining (8.2), (8.3) and (8.4),

$$
\begin{aligned}
\|x-x_j\|_A^2 - \|x-x_{j+1}\|_A^2 = \gamma_j^2\, p_j^T A p_j + 2\, r_{j+1}^T (x_{j+1}-x_j) \\
+ 2\varepsilon\,(f_{j+1} + A z_j^x)^T (x_{j+1}-x_j) + O(\varepsilon^2).
\end{aligned} \tag{8.5}
$$

Substituting for γ_j from (7.12), the first term in (8.5) can be written as

$$
\gamma_j^2\, p_j^T A p_j = \gamma_j \|r_j\|^2 + \varepsilon\, \gamma_j\, p_j^T A p_j\, \zeta_j^\gamma = \gamma_j \|r_j\|^2 + \varepsilon\, \gamma_j \|r_j\|^2 \left\{ \zeta_j^\gamma \frac{p_j^T A p_j}{\|r_j\|^2} \right\}.
$$

Consequently, the difference between the squared A-norms of the error in the consecutive steps can be written in the form convenient for the further analysis

$$
\begin{aligned}
\|x-x_j\|_A^2 - \|x-x_{j+1}\|_A^2 = \gamma_j \|r_j\|^2 + \varepsilon\, \gamma_j \|r_j\|^2 \left\{ \zeta_j^\gamma \frac{p_j^T A p_j}{\|r_j\|^2} \right\} \\
+ 2\, r_{j+1}^T (x_{j+1}-x_j) \\
+ 2\varepsilon\,(f_{j+1} + A z_j^x)^T (x_{j+1}-x_j) + O(\varepsilon^2).
\end{aligned} \tag{8.6}
$$

The goal of the following analysis is to show that until $\|x-x_j\|_A$ reaches its ultimate attainable accuracy level, the terms on the right hand side of (8.6) are, except for $\gamma_j\|r_j\|^2$, insignificant. Bounding the second term will not represent a problem. The norm of the difference $x_{j+1}-x_j = (x-x_j)-(x-x_{j+1})$ is bounded by $2\|x-x_j\|_A/\lambda_1^{1/2}$. Therefore the size of the fourth term is proportional to $\varepsilon\,\|x-x_j\|_A$. The third term is related to the line-search principle. Ideally (in exact arithmetic), the $(j+1)$-th residual is orthogonal to the difference between the $(j+1)$-th and j-th approximation (which is a multiple of the j-th direction vector). This is equivalent to the line-search: ideally the $(j+1)$-th CG approximation minimizes the A-norm of the error along the line determined by the j-th approximation and the j-th direction vector. Here the term $r_{j+1}^T(x_{j+1}-x_j)$, with r_{j+1}, x_j and x_{j+1} computed numerically, examines how closely the line-search holds in finite precision arithmetic. In fact, bounding the local orthogonality $r_{j+1}^T(x_{j+1}-x_j)$ represents the technically most difficult part of the remaining analysis.

9. Local orthogonality in the Hestenes and Stiefel implementation. Since the classical work of Paige it is well-known that in the three-term Lanczos recurrence local orthogonality is preserved close to the machine epsilon (see [35]). We will derive an analogy of this for the CG algorithm, and state it as an independent result.

The local orthogonality term $r_{j+1}^T(x_{j+1}-x_j)$ can be written in the form

$$
r_{j+1}^T (x_{j+1}-x_j) = r_{j+1}^T (\gamma_j p_j + \varepsilon\, z_j^x) = \gamma_j r_{j+1}^T p_j + \varepsilon\, r_{j+1}^T z_j^x. \tag{9.1}
$$

Using the bound $\|r_{j+1}\| \le \lambda_n^{1/2} \|x-x_{j+1}\|_A (1+\varepsilon\, F_{j+1}) \le \lambda_n^{1/2}\|x-x_j\|_A(1+\varepsilon\, F_{j+1})$, see (7.19), the size of the second term in (9.1) is proportional to $\varepsilon\,\|x-x_j\|_A$. The main step consist of showing that the term $r_{j+1}^T p_j$ is sufficiently small. Multiplying the recurrence (7.7) for r_{j+1} by the column vector p_j^T gives (using (7.8) and (7.12))

$$
\begin{aligned}
p_j^T r_{j+1} &= p_j^T r_j - \gamma_j p_j^T A p_j + \varepsilon\, p_j^T z_j^r \\
&= (r_j + \delta_j p_{j-1} + \varepsilon\, z_{j-1}^p)^T r_j - \left(\frac{\|r_j\|^2}{p_j^T A p_j} + \varepsilon\, \zeta_j^\gamma \right) p_j^T A p_j + \varepsilon\, p_j^T z_j^r \\
&= \delta_j\, p_{j-1}^T r_j + \varepsilon\, \{ r_j^T z_{j-1}^p - \zeta_j^\gamma p_j^T A p_j + p_j^T z_j^r \}.
\end{aligned} \tag{9.2}
$$

Denoting

$$M_j \equiv r_j^T z_{j-1}^p - \zeta_j^\gamma p_j^T A p_j + p_j^T z_j^r, \tag{9.3}$$

the identity (9.2) is

$$p_j^T r_{j+1} = \delta_j \, p_{j-1}^T r_j + \varepsilon \, M_j. \tag{9.4}$$

Recursive application of (9.4) for $p_{j-1}^T r_j, \dots, p_1^T r_2$ with $p_0^T r_1 = \|r_0\|^2 - \gamma_0 \, p_0^T A p_0 + \varepsilon \, p_0^T z_0^r = \varepsilon \, \{-\zeta_0^\gamma r_0^T A r_0 + p_0^T z_0^r\} \equiv \varepsilon \, M_0$, gives

$$p_j^T r_{j+1} = \varepsilon \, M_j + \varepsilon \sum_{i=1}^{j} \left(\prod_{k=i}^{j} \delta_k \right) M_{i-1}. \tag{9.5}$$

Since

$$\varepsilon \prod_{k=i}^{j} \delta_k = \varepsilon \prod_{k=i}^{j} \frac{\|r_k\|^2}{\|r_{k-1}\|^2} + O(\varepsilon^2) = \varepsilon \, \frac{\|r_j\|^2}{\|r_{i-1}\|^2} + O(\varepsilon^2),$$

we can express (9.5) as

$$p_j^T r_{j+1} = \varepsilon \, \|r_j\|^2 \sum_{i=0}^{j} \frac{M_i}{\|r_i\|^2} + O(\varepsilon^2). \tag{9.6}$$

Using (9.3),

$$\frac{|M_i|}{\|r_i\|^2} \leq \frac{\|z_{i-1}^p\|}{\|r_i\|} + |\zeta_i^\gamma| \frac{p_i^T A p_i}{\|r_i\|^2} + \frac{\|p_i\| \|z_i^r\|}{\|r_i\|^2}. \tag{9.7}$$

¿From (7.11) it follows

$$\varepsilon \, \frac{\|z_{i-1}^p\|}{\|r_i\|} \leq \varepsilon \left\{ 3 + 2 \frac{\|p_i\|}{\|r_i\|} \right\} + O(\varepsilon^2). \tag{9.8}$$

Using (7.14),

$$\varepsilon \, |\zeta_i^\gamma| \frac{p_i^T A p_i}{\|r_i\|^2} \leq \varepsilon \, \kappa(A) \, O(n + c) + O(\varepsilon^2). \tag{9.9}$$

The last part of (9.7) is bounded using (7.10) and (7.12)

$$\begin{aligned} \varepsilon \, \frac{\|p_i\| \|z_i^r\|}{\|r_i\|^2} &\leq \varepsilon \left\{ \frac{\|p_i\| \|r_i\|}{\|r_i\|^2} + 2 \, \gamma_i \frac{\|p_i\| \|A p_i\|}{\|r_i\|^2} + c \, \gamma_i \frac{\|p_i\| \|A\| \|p_i\|}{\|r_i\|^2} \right\} + O(\varepsilon^2) \\ &= \varepsilon \left\{ \frac{\|p_i\|}{\|r_i\|} + 2 \, \frac{\|p_i\| \|A p_i\|}{p_i^T A p_i} + c \, \frac{\|A\| \|p_i\|^2}{p_i^T A p_i} \right\} + O(\varepsilon^2) \\ &\leq \varepsilon \left\{ \frac{\|p_i\|}{\|r_i\|} + (2 + c) \, \kappa(A) \right\} + O(\varepsilon^2), \end{aligned} \tag{9.10}$$

where

$$\varepsilon \, \frac{\|p_i\|}{\|r_i\|} \leq \varepsilon \, \frac{\|r_i\| + \delta_i \|p_{i-1}\|}{\|r_i\|} + O(\varepsilon^2) \leq \varepsilon \left\{ 1 + \frac{\|r_i\|}{\|r_{i-1}\|} \, \frac{\|p_{i-1}\|}{\|r_{i-1}\|} \right\} + O(\varepsilon^2). \tag{9.11}$$

ETNA
Kent State University
etna@mcs.kent.edu

Recursive application of (9.11) for $\|p_{i-1}\|/\|r_{i-1}\|$, $\|p_{i-2}\|/\|r_{i-2}\|$, …, $\|p_1\|/\|r_1\|$ with $\|p_0\|/\|r_0\| = 1$ gives

$$\varepsilon \frac{\|p_i\|}{\|r_i\|} \le \varepsilon \left\{ 1 + \frac{\|r_i\|}{\|r_{i-1}\|} + \frac{\|r_i\|}{\|r_{i-2}\|} + \ldots + \frac{\|r_i\|}{\|r_0\|} \right\} + O(\varepsilon^2). \tag{9.12}$$

The size of $\varepsilon \, \|r_i\|/\|r_k\|$, $i \ge k$ is, according to (7.20), less or equal than $\varepsilon\, \kappa(A)^{1/2} + O(\varepsilon^2)$. Consequently,

$$\varepsilon \frac{\|p_i\|}{\|r_i\|} \le \varepsilon \, \{1 + i\, \kappa(A)^{1/2}\} + O(\varepsilon^2). \tag{9.13}$$

Summarizing (9.8), (9.9), (9.10) and (9.13), the ratio $\varepsilon \, |M_i|/\|r_i\|^2$ is bounded as

$$\varepsilon \frac{|M_i|}{\|r_i\|^2} \le \varepsilon \, \kappa(A)\, O(8 + 2c + n + 3i) + O(\varepsilon^2). \tag{9.14}$$

Combining this result with (9.6) proves the following theorem.

THEOREM 9.1. *Using the previous notation, let $\varepsilon\,(n + c)\,\kappa(A) \ll 1$. Then the local orthogonality between the direction vectors and the iteratively computed residuals is in the finite precision implementation of the conjugate gradient method (7.6)–(7.8) and (7.12) bounded by*

$$|p_j^T r_{j+1}| \le \varepsilon \, \|r_j\|^2 \kappa(A)\, O((j + 1)(8 + 2c + n + 3j)) + O(\varepsilon^2). \tag{9.15}$$

10. Final precision analysis – conclusions.

We now return to (8.6) and finalize our discussion. Using (9.1) and (9.6),

$$\begin{aligned} \|x - x_j\|_A^2 - \|x - x_{j+1}\|_A^2 \;=\; & \gamma_j \|r_j\|^2 \\ & + \varepsilon\, \gamma_j \|r_j\|^2 \left\{ \zeta_j^{\gamma} \frac{p_j^T A p_j}{\|r_j\|^2} + 2 \sum_{i=0}^{j} \frac{M_i}{\|r_i\|^2} \right\} \\ & + 2\varepsilon \, \{(f_{j+1} + A z_j^x)^T (x_{j+1} - x_j) + r_{j+1}^T z_j^x \} \\ & + O(\varepsilon^2). \end{aligned} \tag{10.1}$$

The term

$$E_j^{(1)} \equiv \varepsilon \left\{ \zeta_j^{\gamma} \frac{p_j^T A p_j}{\|r_j\|^2} + 2 \sum_{i=0}^{j} \frac{M_i}{\|r_i\|^2} \right\}$$

is bounded using (7.14) and (9.14),

$$|E_j^{(1)}| \le \varepsilon \, \kappa(A)\, O(n + c + 2(j + 1)(8 + 2c + n + 3j)) + O(\varepsilon^2). \tag{10.2}$$

We write the remaining term on the right hand side of (10.1) proportional to ε as

$$2\varepsilon \, \{(f_{j+1} + A z_j^x)^T (x_{j+1} - x_j) + r_{j+1}^T z_j^x\} \equiv \|x - x_j\|_A \, E_j^{(2)}, \tag{10.3}$$

where

$$\begin{aligned} |E_j^{(2)}| &= 2\varepsilon \left| (f_{j+1} + A z_j^x)^T \left(\frac{x_{j+1} - x + x - x_j}{\|x - x_j\|_A} \right) + \frac{r_{j+1}^T}{\|x - x_j\|_A} z_j^x \right| \\ &\le 2\varepsilon \, \{ 2\, (\|f_{j+1}\| \lambda_1^{-1/2} + \|A\|^{1/2} \|z_j^x\|) + \|A\|^{1/2} \|z_j^x\| \}. \end{aligned} \tag{10.4}$$

ETNA
Kent State University
etna@mcs.kent.edu

With (7.16) and (7.9),

$$
\begin{aligned}
|E_j^{(2)}| &\leq 4\varepsilon \|A\|^{1/2}\kappa(A)^{1/2}(\|x\| + \max_{0\leq i\leq j+1}\|x_i\|)\, O(jc) \\
&\quad + 5\|A\|^{1/2}\varepsilon(3\|x_j\| + 2\|x_{j+1}\|) + O(\varepsilon^2) \\
&\leq \varepsilon \|A\|^{1/2}\kappa(A)^{1/2}(\|x\| + \max_{0\leq i\leq j+1}\|x_i\|)\, O(4jc+25) + O(\varepsilon^2). \qquad (10.5)
\end{aligned}
$$

Finally, using the fact that the monotonicity of the A-norm and the Euclidean norm of the error is preserved also in finite precision CG computations (with small additional inaccuracy, see [19], [22]), we obtain the finite precision analogy of (4.4), which is formulated as a theorem.

THEOREM 10.1. *With the notation defined above, let $\varepsilon\,(n+c)\,\kappa(A) \ll 1$. Then the CG approximate solutions computed in finite precision arithmetic satisfy*

$$
\|x - x_j\|_A^2 - \|x - x_{j+d}\|_A^2 \;=\; \nu_{j,d} + \nu_{j,d}\, E_{j,d}^{(1)} \;+\; \|x - x_j\|_A\, E_{j,d}^{(2)} + O(\varepsilon^2), \qquad (10.6)
$$

where

$$
\nu_{j,d} = \sum_{i=j}^{j+d-1} \gamma_i \|r_i\|^2 \qquad (10.7)
$$

and the terms due to rounding errors are bounded by

$$
\begin{aligned}
&|E_{j,d}^{(1)}| \leq O(d) \max_{j\leq i\leq j+d-1} |E_i^{(1)}|, \qquad (10.8)\\
&\qquad |E_i^{(1)}| \leq \varepsilon\, \kappa(A)\, O(t^{(1)}(n)) + O(\varepsilon^2),\\
&|E_{j,d}^{(2)}| \leq O(d) \max_{j\leq i\leq j+d-1} |E_i^{(2)}|, \qquad (10.9)\\
&\qquad |E_i^{(2)}| \leq \varepsilon\, \|A\|^{1/2}\kappa(A)^{1/2}(\|x\| + \max_{0\leq i\leq j+1}\|x_i\|)\, O(t^{(2)}(n)) + O(\varepsilon^2).
\end{aligned}
$$

$O(t^{(1)}(n))$ and $O(t^{(2)}(n))$ represent terms bounded by a small degree polynomial in n independent of any other variables.

Please note that the value $\nu_{j,d}$ is in Theorem 10.1 computed *exactly* using (10.7). Errors in computing $\nu_{j,d}$ *numerically* (i.e. in computing $\mathrm{fl}(\sum_{i=j}^{j+d-1}\gamma_i\|r_i\|^2)$) are negligible in comparison to $\nu_{j,d}$ multiplied by the bound for the term $|E_i^{(1)}|$ and need not be considered here. Theorem 10.1 therefore says that for the numerically computed approximate solutions

$$
\|x - x_j\|_A^2 - \|x - x_{j+d}\|_A^2 \;= \mathrm{fl}(\nu_{j,d}) + \widetilde{\nu}_{j,d}, \qquad (10.10)
$$

where the term $\widetilde{\nu}_{j,d}$ "perturbes" the ideal identity (4.4) in the finite precision case. Here $\widetilde{\nu}_{j,d}$ denotes quantity insignificantly different from $\widetilde{\nu}_{j,d}$ in (8.1). Consequently, the numerically computed value $\nu_{j,d}$ can be trusted until it reaches the level of $\widetilde{\nu}_{j,d}$. Based on the assumption $\varepsilon(n+c)\,\kappa(A) \ll 1$ and (10.8) we consider $|E_i^{(1)}| \ll 1$. Then, assuming (4.5), the numerically computed value $\nu_{j,d}$ gives a good estimate for the A-norm of the error $\|x - x_j\|_A^2$ until

$$
\|x - x_j\|_A\, |E_{j,d}^{(2)}| \ll \|x - x_j\|_A^2,
$$

which is equivalent to

$$
\|x - x_j\|_A \gg |E_{j,d}^{(2)}|. \qquad (10.11)
$$

ETNA
Kent State University
etna@mcs.kent.edu

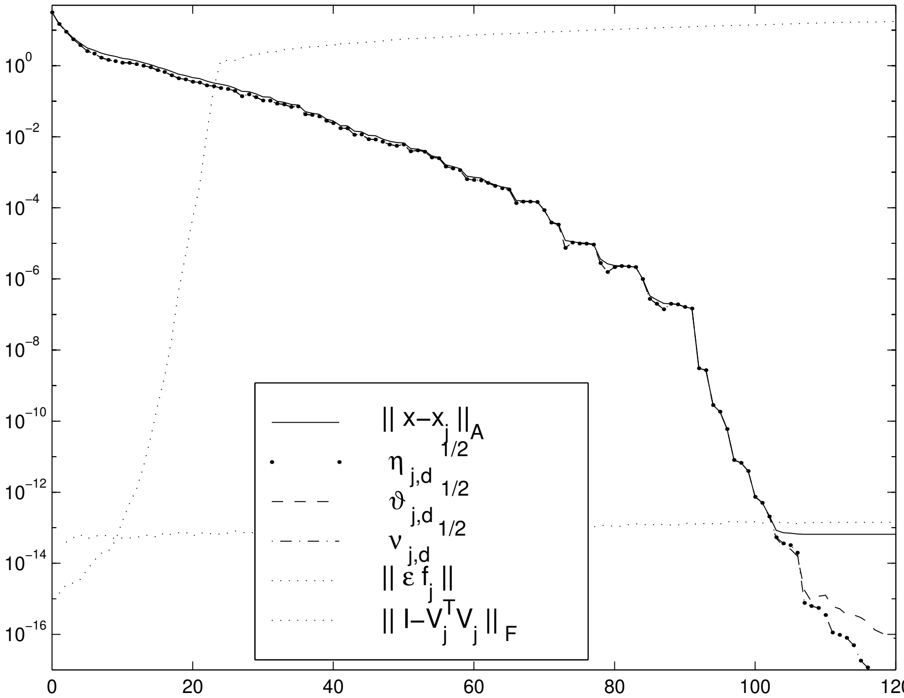

FIG. 11.1. *Error estimates $\eta_{j,d}^{1/2}$ (dots), $\vartheta_{j,d}^{1/2}$ (dashed-line) and $\nu_{j,d}^{1/2}$ (dash-doted line). They essentially coincide until $\|x - x_j\|_A$ (solid line) reaches its ultimate attainable accuracy. The loss of orthogonality is plotted by the dotted line. We used $d = 4$.*

The value $E_{j,d}^{(2)}$ represents various terms. Its upper bound is, apart from $\kappa(A)^{1/2}$, which comes into play as an effect of the worst-case rounding error analysis, linearly dependent on an upper bound for $\|x - x_0\|_A$. The value of $E_{j,d}^{(2)}$ is (as similar terms or constants in any other rounding error analysis) not important. What is important is the following possible interpretation of (10.11): until $\|x - x_j\|_A$ reaches a level close to $\varepsilon\|x - x_0\|_A$, the computed estimate $\nu_{j,d}$ must work.

11. Numerical Experiments. We present illustrative experimental results for the system $Ax = b$ described in Section 5. We set $d = 4$.

Fig. 11.1 demonstrates, that the estimates $\eta_{j,d}^{1/2}$ (computed by the algorithm CGQL [16], dotted line), $\vartheta_{j,d}^{1/2}$ (dashed line) and $\nu_{j,d}^{1/2}$ (dash-dotted line) give in the presence of rounding errors similar results; all the lines essentially coincide until $\|x - x_j\|_A$ (solid line) reaches its ultimate attainable accuracy level. Loss of orthogonality, measured by $\|I - V_j^T V_j\|_F$, is plotted by the strictly increasing dotted line. We see that the orthogonality of the computed Lanczos basis is completely lost at $j \sim 22$. The term $\|\varepsilon f_j\|$ measuring the difference between the directly and iteratively computed residuals (horizontal dotted line) remains close to machine precision $\varepsilon \sim 10^{-16}$ throughout the whole computation.

Fig. 11.2 shows, in addition to the loss of orthogonality (dotted line) and the Euclidean norm of the error $\|x - x_j\|$, the bound for the last one derived in the following way from (1.7). Using the identity

$$\|x - x_j\|^2 = \sum_{i=j}^{j+d-1} \frac{\|p_i\|^2}{(p_i, Ap_i)} \left(\|x - x_i\|_A^2 + \|x - x_{i+1}\|_A^2\right) + \|x - x_{j+d}\|^2, \tag{11.1}$$

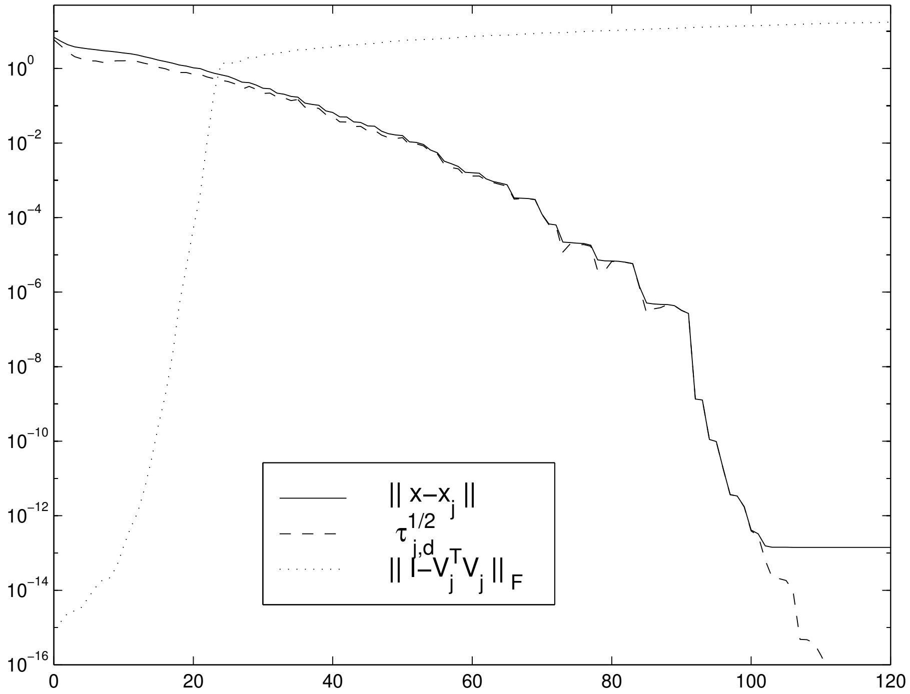

FIG. 11.2. *Lower bound* $\tau_{j,d}^{1/2}$ (*dashed line*) *for the Euclidean norm of the error* (*solid line*). *The bound* $\tau_{j,d}^{1/2}$ (*with* $d = 4$) *gives, despite the loss of orthogonality* (*dotted line*), *very good approximation to* $\|x - x_j\|$.

and replacing the unknown squares of the A-norms of the errors

$$\|x - x_j\|_A^2,\ \|x - x_{j+1}\|_A^2,\ \ldots,\ \|x - x_{j+d}\|_A^2$$

by their estimates

$$\sum_{i=j}^{j+2d-1} \gamma_i \|r_i\|^2,\ \sum_{i=j+1}^{j+2d-1} \gamma_i \|r_i\|^2,\ \ldots,\ \sum_{i=j+d}^{j+2d-1} \gamma_i \|r_i\|^2$$

gives ideally

$$\|x - x_j\|^2 \geq \sum_{i=j}^{j+d-1} \frac{\|p_i\|^2}{(p_i, Ap_i)} \left(\gamma_i \|r_i\|^2 + 2 \sum_{k=i+1}^{j+2d-1} \gamma_k \|r_k\|^2 \right) + \|x - x_{j+d}\|^2. \tag{11.2}$$

Similarly as above, if d is chosen such that

$$\|x - x_j\|^2 \gg \|x - x_{j+d}\|^2 \quad \text{and} \quad \|x - x_{j+d}\|_A^2 \gg \|x - x_{j+2d}\|_A^2,$$

then

$$\tau_{j,d} \equiv \sum_{i=j}^{j+d-1} \frac{\|p_i\|^2}{(p_i, Ap_i)} \left(\gamma_i \|r_i\|^2 + 2 \sum_{k=i+1}^{j+2d-1} \gamma_k \|r_k\|^2 \right) \tag{11.3}$$

represents ideally a tight lower bound for the squared Euclidean norm of the CG error $\|x - x_j\|^2$. Please note that evaluating (11.3) requires $2d$ extra steps.

In experiments shown in Fig. 11.1 and Fig. 6.1 we used a fixed value $d = 4$. It would be interesting to design an adaptive error estimator, which would use some heuristics for adjusting d according to the desired accuracy of the estimate and the convergence behaviour. A similar approach can be used for eliminating the disadvantage of $2d$ extra steps related to (11.3). We hope to report results of our work on that subject elsewhere.

12. Conclusions. Based on the results presented above we believe that the estimate for the A-norm of the error $\nu_{j,d}^{1/2}$ should be incorporated into any software realization of the CG method. It is simple and numerically stable. It is worth to consider the estimate $\tau_{j,d}^{1/2}$ for the Euclidean norm of the error, and compare it (including complexity and numerical stability) with other existing approaches not discussed here (e.g. [6], [31]). The choice of d remains a subject of further work.

By this paper we wish to pay a tribute to the truly seminal paper of Hestenes and Stiefel [24] and to the work of Golub who shaped the whole field.

Acknowledgments. Many people have contributed to the presentation of this paper by their advice, helpful objections, and remarks. We wish to especially thank Martin Gutknecht, Gerard Meurant, Chris Paige, Beresford Parlett, Lothar Reichel, Miroslav Rozložník, and the anonymous referee for their help in revising the text.

REFERENCES

[1] M. ARIOLI, *Stopping criterion for the Conjugate Gradient algorithm in a Finite Element method framework*, submitted to Numer. Math., (2001).

[2] M. ARIOLI AND L. BALDINI, *Backward error analysis of a null space algorithm in sparse quadratic programming*, SIAM J. Matrix Anal. Appl., 23 (2001), pp. 425–442.

[3] O. AXELSSON AND I. KAPORIN, *Error norm estimation and stopping criteria in preconditioned Conjugate Gradient iterations*, Numer. Linear Algebra Appl., 8 (2001), pp. 265–286.

[4] D. CALVETTI, G. H. GOLUB, AND L. REICHEL, *Estimation of the L-curve via Lanczos bidiagonalization*, BIT, 39 (1999), pp. 603–609.

[5] D. CALVETTI, S. MORIGI, L. REICHEL, AND F. SGALLARI, *Computable error bounds and estimates for the Conjugate Gradient Method*, Numer. Algorithms, 25 (2000), pp. 79–88.

[6] ———, *An iterative method with error estimators*, J. Comput. Appl. Math., 127 (2001), pp. 93–119.

[7] G. DAHLQUIST, S. EISENSTAT, AND G. H. GOLUB, *Bounds for the error of linear systems of equations using the theory of moments*, J. Math. Anal. Appl., 37 (1972), pp. 151–166.

[8] G. DAHLQUIST, G. H. GOLUB, AND S. G. NASH, *Bounds for the error in linear systems*, in Proc. Workshop on Semi-Infinite Programming, R. Hettich, ed., Springer, Berlin, 1978, pp. 154–172.

[9] P. DEUFLHARD, *Cascadic conjugate gradient methods for elliptic partial differential equations I: Algorithm and numerical results*, preprint SC 93-23, Konrad-Zuse-Zentrum für Informationstechnik Berlin, Heilbronnen Str., D-10711 Derlin, October 1993.

[10] ———, *Cascadic conjugate gradient methods for elliptic partial differential equations: Algorithm and numerical results*, Contemp. Math., 180 (1994), pp. 29–42.

[11] B. FISCHER, *Polynomial Based Iteration Methods for Symmetric Linear Systems*, Wiley Teubner Advances in Numerical Mathematics, Wiley Teubner, 1996.

[12] B. FISCHER AND G. H. GOLUB, *On the error computation for polynomial based iteration methods*, in Recent Advances in Iterative Methods, G. H. Golub, A. Greenbaum, and M. Luskin, eds., Springer, N.Y., 1994, pp. 59–67.

[13] A. FROMMER AND A. WEINBERG, *Verified error bounds for linear systems through the Lanczos process*, Reliable Computing, 5 (1999), pp. 255–267.

[14] W. GAUTSCHI, *A survey of Gauss-Christoffel quadrature formulae*, in E.B. Christoffel. The Influence of His Work on Mathematics and the Physical Sciences, P. Bultzer and F. Fehér, eds., Birkhauser, Boston, 1981, pp. 73–157.

[15] G. H. GOLUB AND G. MEURANT, *Matrices, moments and quadrature*, in Numerical Analysis 1993, vol 303, Pitman research notes in mathematics series, D. Griffiths and G. Watson, eds., Longman Sci. Tech. Publ., 1994, pp. 105–156.

[16] ———, *Matrices, moments and quadrature II: How to compute the norm of the error in iterative methods*, BIT, 37 (1997), pp. 687–705.

[17] G. H. GOLUB AND Z. STRAKOŠ, *Estimates in quadratic formulas*, Numer. Algorithms, 8 (1994), pp. 241–268.

[18] G. H. GOLUB AND C. VAN LOAN, *Matrix Computation*, The Johns Hopkins University Press, Baltimore MD, third ed., 1996.

[19] A. GREENBAUM, *Behavior of slightly perturbed Lanczos and Conjugate Gradient recurrences*, Linear Algebra Appl., 113 (1989), pp. 7–63.

[20] ———, *Estimating the attainable accuracy of recursively computed Residual methods*, SIAM J. Matrix Anal. Appl., 18 (3) (1997), pp. 535–551.

[21] ———, *Iterative Methods for Solving Linear Systems*, SIAM, Philadelphia, 1997.

[22] A. GREENBAUM AND Z. STRAKOŠ, *Predicting the behavior of finite precision Lanczos and Conjugate Gradient computations*, SIAM J. Matrix Anal. Appl., 18 (1992), pp. 121–137.

[23] M. H. GUTKNECHT AND Z. STRAKOŠ, *Accuracy of two three-term and three two-term recurrences for Krylov space solvers*, SIAM J. Matrix Anal. Appl., 22 (2001), pp. 213–229.

[24] M. R. HESTENES AND E. STIEFEL, *Methods of conjugate gradients for solving linear systems*, J. Research Nat. Bur. Standards, 49 (1952), pp. 409–435.

[25] N. J. HIGHAM, *Accuracy and stability of numerical algorithms*, SIAM, Philadelphia, PA, 1996.

[26] S. KARLIN AND L. S. SHAPLEY, *Geometry of moment spaces*, Memoirs of the Americam Mathematical Society 12, Providence, (1953).

[27] C. LANCZOS, *Solution of systems of linear equations by minimized iterations*, J. Research Nat. Bur. Standards, 49 (1952), pp. 33–53.

[28] G. MEURANT, *The computation of bounds for the norm of the error in the Conjugate Gradient algorithm*, Numer. Algorithms, 16 (1997), pp. 77–87.

[29] ———, *Computer Solution of Large Linear Systems*, vol. 28 of Studies in Mathematics and Its Applications, Elsevier, 1999.

[30] ———, *Numerical experiments in computing bounds for the norm of the error in the preconditioned Conjugate Gradient algorithm*, Numer. Algorithms 22, 3-4 (1999), pp. 353–365.

[31] ———, *Towards a reliable implementation of the conjugate gradient method*. Invited plenary lecture at the Latsis Symposium: Iterative Solvers for Large Linear Systems, Zurich, February 2002.

[32] Y. NOTAY, *On the convergence rate of the Conjugate Gradients in the presence of rounding errors*, Numer. Math., 65 (1993), pp. 301–317.

[33] C. C. PAIGE, *The Computation of Eigenvalues and Eigenvectors of Very Large Sparse Matrices*, PhD thesis, Intitute of Computer Science, University of London, London, U.K., 1971.

[34] ———, *Computational variants of the Lanczos method for the eigenproblem*, J. Inst. Math. Appl., 10 (1972), pp. 373–381.

[35] ———, *Error analysis of the Lanczos algorithm for tridiagonalizing a symmetric matrix*, J. Inst. Math. Appl., 18 (1976), pp. 341–349.

[36] ———, *Accuracy and effectiveness of the Lanczos algorithm for the symmetric eigenproblem*, Linear Algebra Appl., 34 (1980), pp. 235–258.

[37] C. C. PAIGE AND Z. STRAKOŠ, *Correspondence between exact arithmetic and finite precision behaviour of Krylov space methods*, in XIV. Householder Symposium, J. Varah, ed., University of British Columbia, 1999, pp. 250–253.

[38] B. N. PARLETT, *The Symmetric Eigenvalue Problem*, Prentice Hall, Englewood Cliffs, 1980.

[39] ———, *Do we fully understand the symmetric Lanczos algorithm yet?*, in Proceedins of the Lanczos Centenary Conference, Philadelphia, 1994, SIAM, pp. 93–107.

[40] P. E. SAYLOR AND D. C. SMOLARSKI, *Addendum to: Why Gaussian quadrature in the complex plane?*, Numer. Algorithms, 27 (2001), pp. 215–217.

[41] ———, *Why Gaussian quadrature in the complex plane?*, Numer. Algorithms, 26 (2001), pp. 251–280.

[42] G. L. SLEIJPEN, H. A. VAN DER VORST, AND D. R. FOKKEMA, *BiCGstab(l) and other hybrid BiCG methods*, Numer. Algorithms, 7 (1994), pp. 75–109.

[43] Z. STRAKOŠ, *On the real convergence rate of the Conjugate Gradient method*, Linear Algebra Appl., 154-156 (1991), pp. 535–549.

[44] ———, *Convergence and numerical behaviour of the Krylov space methods*, in Algorithms for Large Sparse Linear Algebraic Systems: The State of the Art and Applications in Science and Engineering, G. W. Althaus and E. Spedicato, eds., NATO ASI Institute, Kluwer Academic, 1998, pp. 175–197.

[45] Z. STRAKOŠ AND A. GREENBAUM, *Open questions in the convergence analysis of the Lanczos process for the real symmetric eigenvalue problem*, IMA preprint series 934, University of Minnesota, 1992.

[46] G. SZEGÖ, *Orthogonal Polynomials*, AMS Colloq. Publ. 23, AMS, Providence, 1939.

[47] K. F. WARNICK, *Nonincreasing error bound for the Biconjugate Gradient method*, report, University of Illinois, 2000.