Roundoff-Error Analysis of a New Class of Conjugate-Gradient Algorithms*

H. Woźniakowski†
Department of Computer Science
Carnegie-Mellon University
Pittsburgh, Pennsylvania 15213

Dedicated to Alston S. Householder
on the occasion of his seventy-fifth birthday.

Submitted by G. W. Stewart

ABSTRACT

We perform the rounding-error analysis of the conjugate-gradient algorithms for the solution of a large system of linear equations $A\mathbf{x}=\mathbf{b}$ where A is an hermitian and positive definite matrix. We propose a new class of conjugate-gradient algorithms and prove that in the spectral norm the relative error of the computed sequence $\{\mathbf{x}_k\}$ (in floating-point arithmetic) depends at worst on $\zeta\kappa^{3/2}$, where ζ is the relative computer precision and κ is the condition number of A. We show that the residual vectors $\mathbf{r}_k=A\mathbf{x}_k-\mathbf{b}$ are at worst of order $\zeta\kappa\|A\|\,\|\mathbf{x}_k\|$. We point out that with iterative refinement these algorithms are numerically stable. If $\zeta\kappa^2$ is at most of order unity, then they are also well behaved.

1. INTRODUCTION

We study conjugate-gradient algorithms (for brevity, cg algorithms) for the solution of a large linear system $A\mathbf{x}=\mathbf{b}$ where A is an $n\times n$ hermitian and positive definite matrix. By a cg algorithm we mean an implementation of the cg iteration in floating-point arithmetic. We are primarily interested in the roundoff-error analysis of these algorithms.

*Most of the material in this paper was first presented by the author at the Gatlinburg VII meeting at Asilomar, December 1977. This research was supported in part by the National Science Foundation under Grant MCS 75-222-55 and the Office of Naval Research under Contract N00014-76-C-0370, NR 044-422.

†On leave from the University of Warsaw.

0024-3795/80/010507+23$01.75

It is well known that the cg iteration enjoys optimal complexity in a sense to be made precise in Sec. 2. In exact arithmetic it generates a sequence of orthogonal residual vectors $\mathbf{r}_k = A\mathbf{x}_k - \mathbf{b}$, and the solution $\boldsymbol{\alpha} = A^{-1}\mathbf{b}$ is obtained after at most n steps. (See, among others, [3, 6, 2].) Many of these theoretical properties do not hold in the presence of rounding errors. It is no longer true that the computed residual vectors are orthogonal (or even nearly orthogonal) and that the nth computed vector $\mathbf{x}_n$ is a reasonable approximation to $\boldsymbol{\alpha}$.

The aim of this paper is to understand the behavior of some cg algorithms in the presence of rounding errors. We are primarily interested in studying how the matrix condition number $\kappa = \|A\| \|A^{-1}\|$, where $\|A\|$ denotes the spectral norm of A, influences the relative error of the computed sequence $\{\mathbf{x}_k\}$.

We know that direct algorithms of practical interest as well as many iterative algorithms with iterative refinement are well behaved, i.e., they assure the computation of an approximation $\mathbf{y}$ in floating point arithmetic (f1) such that $\mathbf{y}$ is the exact solution of a slightly perturbed system, i.e., $(A + \delta A)\mathbf{y} = \mathbf{b}$, where $\|\delta A\|$ is of order $\zeta \|A\|$ and ζ is the relative computer precision. Equivalently, the residual vector $\mathbf{r} = A\mathbf{y} - \mathbf{b}$ has a norm of order $\zeta \|A\| \|\mathbf{y}\|$. When it cannot be established that an algorithm is well behaved, it is sometimes possible to prove a weaker property, namely that an algorithm is *numerically stable*, i.e., the relative error of a computed $\mathbf{y}$ is of order $\zeta\kappa$. See [9, 10] for direct algorithms, and [4, 11, 12] for iterative algorithms.

To help the reader get an idea of what results can be expected for cg algorithms we report numerical tests. We tested Concus, Golub and O'Leary's [1] conjugate-gradient algorithm. We performed $j (j \gg n)$ iterative steps finding the best possible approximation $\mathbf{x}_k$, $k \leqslant j$, among all computed vectors. Next we computed the relative error of $\mathbf{x}_k$ and its residual vector. Define the number s such that

$$\frac{\|\mathbf{x}_k - \alpha\|}{\|\mathbf{x}_k\|} = \zeta\kappa^s. \tag{1.1}$$

We would like $s = 1$, which would imply the numerical stability of the algorithm. However, for most cases s was about $\frac{3}{2}$ and the residual vector had a spectral norm of order $\zeta\kappa\|A\| \|\mathbf{x}_k\|$. Therefore, the algorithm is neither well behaved nor numerically stable. A natural problem is to understand why this is so and to seek a cg algorithm which is numerically stable and (perhaps) well behaved.

To understand why $s = \frac{3}{2}$, recall that the cg algorithms minimize the error in the A-norm, $\|A^{1/2}(\mathbf{x}_k - \boldsymbol{\alpha})\|$ (not in the spectral norm $\|\mathbf{x}_k - \boldsymbol{\alpha}\|$). Therefore it seems natural to measure the error by $\|A^{1/2}(\mathbf{x}_k - \boldsymbol{\alpha})\|$ instead of by

$\|\mathbf{x}_k-\boldsymbol{\alpha}\|$. Suppose there exists a numerically stable cg algorithm in the A-norm, i.e.,

$$\|A^{1/2}(\mathbf{x}_k-\boldsymbol{\alpha})\| = O\big(\zeta\kappa\|A^{1/2}\mathbf{x}_k\|\big). \tag{1.2}$$

Note that the condition number of A in the A-norm coincides with the condition number of A in the spectral norm. Since $\|\mathbf{x}_k-\boldsymbol{\alpha}\| \leqslant \|A^{-1/2}\|\,\|A^{1/2}(\mathbf{x}_k-\boldsymbol{\alpha})\|$, (1.2) yields

$$\|\mathbf{x}_k-\boldsymbol{\alpha}\| = O\big(\zeta\kappa^{3/2}\|\mathbf{x}_k\|\big).$$

This explains why $s=\frac{3}{2}$ might be expected in (1.1). See the Appendix, where a detailed discussion of numerical tests is reported.

We have not succeeded in analyzing classical cg algorithms, including that proposed by Concus, Golub and O'Leary [1]. In this paper we propose a new class of cg algorithms and prove that for these algorithms there exists a computed vector $\mathbf{x}_k$ such that

$$\|A^{1/2}(\mathbf{x}_k-\boldsymbol{\alpha})\| \leqslant C\zeta\kappa\|A^{1/2}\|\,\|\mathbf{x}_k\|, \tag{1.3}$$

where C is a constant of order at most n. We shall denote this class of algorithms by Φ. Note that κ occurs linearly in (1.3). In general, we cannot say that (1.3) means numerical stability of the cg algorithms in its "own" norm, since we have $\|A^{1/2}\|\,\|\mathbf{x}_k\|$ instead of $\|A^{1/2}\mathbf{x}_k\|$. However, if $\|A^{1/2}\|\,\|\mathbf{x}_k\|$ is of order $\|A^{1/2}\mathbf{x}_k\|$, then these cg algorithms are *numerically stable in the A-norm.*

For the residual vectors we are only able to prove that

$$\|\mathbf{r}_k\| \leqslant C\zeta\kappa\|A\|\,\|\mathbf{x}_k\|. \tag{1.4}$$

We tested one algorithm φ from Φ. For most cases φ looked like a well-behaved algorithm, i.e., $\|\mathbf{r}_k\|$ was of order $\zeta\|A\|\,\|\mathbf{x}_k\|$. However, for a few cases, $\|\mathbf{r}_k\|$ was of order $\zeta\kappa\|A\|\,\|\mathbf{x}_k\|$. This proves that (1.4) is sharp and some cg algorithms from Φ are *not well behaved.*

Many iterative algorithms have this property, i.e., they are numerically stable but not well behaved. Examples include the Chebyshev, SOR, Richardson and Jacobi iterative algorithms (see [11, 12]). However, it was shown by Jankowski and Woźniakowski [4] that any algorithm (direct or iterative) which computes an approximation y such that $\|\mathbf{y}-\boldsymbol{\alpha}\| \leqslant q\|\boldsymbol{\alpha}\|$ with $q<1$ followed by iterative refinement in single precision becomes numerically stable, and if $\zeta\kappa^2$ is of order unity then it is also well behaved. Hence,

any cg algorithm from Φ with iterative refinement is numerically stable in the spectral norm whenever $\zeta\kappa^{3/2}$ is bounded away from unity, and it is well behaved whenever $\zeta\kappa^2$ is of order unity.

We summarize the contents of the paper. In Sec. 2 we briefly state the basic theoretical properties of the cg iteration and derive a three-term recurrence formula for the vectors $\{\mathbf{x}_k\}$ which explains the connection between the cg iteration and the steepest-descent iteration.

Section 3 deals with the rounding error analysis of a steepest-descent algorithm. We prove that the inequality (1.3) holds for this algorithm.

Section 4 deals with the roundoff-error analysis of a new class of cg algorithms. Based on the results of Sec. 3, we prove (1.3) and (1.4), which are the main results of this paper.

In the final section we pose a conjecture on the speed of convergence of sequences computed by cg algorithms.

2. GRADIENT AND CONJUGATE GRADIENT ITERATIONS

In this section we briefly derive some basic properties of the gradient and conjugate-gradient iterations. We consider the solution of a large linear system

$$A\mathbf{x}=\mathbf{b}, \tag{2.1}$$

where $A=A^*>0$ is an $n\times n$ hermitian and positive definite matrix and $\mathbf{b}$ is a $n\times 1$ given vector. Suppose that the information about the matrix A is given by a procedure which computes $\mathbf{y}=A\mathbf{x}$ for a given $\mathbf{x}$. For large systems A is usually sparse, which permits the evaluation of $\mathbf{y}$ in time and storage proportional to n.

We solve (2.1) iteratively by constructing a sequence $\{\mathbf{x}_k\}$ converging to the solution $\boldsymbol{\alpha}=A^{-1}\mathbf{b}$. Let $B=B^*>0$ be a matrix which commutes with A: $BA=AB$. For instance one can set $B=A^p$ for a real p. Let $\|\mathbf{x}\|_B=\sqrt{(B\mathbf{x},\mathbf{x})}$ $=\|B^{1/2}\mathbf{x}\|$, where $\|\mathbf{x}\|=\sqrt{(\mathbf{x},\mathbf{x})}$ is the spectral norm.

We recall the definition of the gradient iteration which constructs the sequence $\{\mathbf{x}_k\}$ as follows. Let $\mathbf{x}_0$ be a given initial approximation and

$$\mathbf{x}_{k+1}=\mathbf{x}_k-c_k\mathbf{r}_k, \qquad \mathbf{r}_k=A\mathbf{x}_k-\mathbf{b}, \tag{2.2}$$

where c_k is chosen in such a way that the error $e_{k+1}=\|\mathbf{x}_{k+1}-\boldsymbol{\alpha}\|_B$ is minimized, i.e., $\|\mathbf{x}_{k+1}-\boldsymbol{\alpha}\|_B=\inf_c \|\mathbf{x}_k-c\mathbf{r}_k\|_B$. This yields

$$c_k=\frac{(\mathbf{r}_k,B(\mathbf{x}_k-\boldsymbol{\alpha}))}{(\mathbf{r}_k,B\mathbf{r}_k)}. \tag{2.3}$$

Note that $(B(\mathbf{x}_{k+1}-\boldsymbol{\alpha}),\mathbf{r}_k)=0$. The coefficient c_k is computable only for certain matrices B. Suppose that $B=A^p$, where p is an integer. Then $c_k(\mathbf{r}_k,A^{p-1}\mathbf{r}_k)/(\mathbf{r}_k,A^p\mathbf{r}_k)$ is computable. For $B=I$, we cannot, in general, compute the numerator of (2.3), $(\mathbf{r}_k,\mathbf{x}_k-\boldsymbol{\alpha})$. However, if one considers the system $M\mathbf{x}=\mathbf{g}$ with a nonsingular M which is nonhermitian or nonpositive definite, and if one agrees to multiply this system by M^*, then $A=M^*M$, $\mathbf{b}=M^*\mathbf{g}$ and $c_k=(M\mathbf{x}_k-\mathbf{g},M\mathbf{x}_k-\mathbf{g})/(\mathbf{r}_k,\mathbf{r}_k)$ is computable.

It is well known that $\{\mathbf{x}_k\}$ converges to $\boldsymbol{\alpha}$ and

$$e_{k+1}=\sqrt{e_k^2-c_k(\mathbf{e}_k,B\mathbf{r}_k)}\leqslant\left(\frac{\kappa-1}{\kappa+1}\right)e_k\leqslant\left(\frac{\kappa-1}{\kappa+1}\right)^{k+1}e_0, \tag{2.4}$$

where $\mathbf{e}_k=\mathbf{x}_k-\boldsymbol{\alpha}$, $e_k=\|\mathbf{e}_k\|_B$ and $\kappa=\|A\|\,\|A^{-1}\|$ is the condition number of the matrix A. (Note that $\|A\|_B=\|A\|$ and $\|A^{-1}\|_B=\|A^{-1}\|$.)

Recall that for $B=A$, the iteration (2.2), (2.3) is called the steepest-descent iteration. It has, in general, very slow convergence and therefore is not recommended in numerical practice. The conjugate-gradient iteration is much more efficient. The following derivation of the cg iteration focuses on its complexity optimality.

Consider a class of iterations for which the error formula satisfies the relation

$$\mathbf{x}_k-\boldsymbol{\alpha}=W_k(A)(\mathbf{x}_0-\boldsymbol{\alpha}), \tag{2.5}$$

where W_k is a polynomial of degree at most k and $W_k(0)=1$. A natural complexity question is how to choose the polynomials W_k. Since we want to minimize the computational complexity (cost), we seek W_k such that the error $e_k=\|\mathbf{x}_k-\boldsymbol{\alpha}\|_B$ is minimized. This means that the polynomials W_k are the solution of the following problem:

$$\|W_k(A)(\mathbf{x}_0-\boldsymbol{\alpha})\|_B=\inf_{P\in W_k(0,1)}\|P(A)(\mathbf{x}_0-\boldsymbol{\alpha})\|_B, \tag{2.6}$$

where $W_k(0,1)$ is the class of polynomials of degree at most k normalized to unity at the origin. The solution of (2.6) is given by the orthogonal polynomials defined as follows (see e.g., [6]). Let

$$\mathbf{x}_0-\boldsymbol{\alpha}=\sum_{j=1}^{m}c_j\boldsymbol{\xi}_j, \tag{2.7}$$

where $\boldsymbol{\xi}_j$ is an eigenvector of A associated with the eigenvalue λ_j: $A\boldsymbol{\xi}_j=\lambda_j\boldsymbol{\xi}_j$, $\|\boldsymbol{\xi}_j\|=1$, $0<\lambda_1<\lambda_2<\cdots<\lambda_m$, with $m\leqslant n$ and $c_j\neq 0$ for $j=1,2,\ldots,m$. Note

that ξ_j is also an eigenvector of B: $B\xi_j = \beta_j \xi_j$ for $\beta_j > 0$, $j = 1,2,\ldots,m$. Define the inner product

$$(f,g) = \sum_{j=1}^{m} |c_j|^2 \beta_j \lambda_j f(\lambda_j)\, \overline{g(\lambda_j)}\,, \tag{2.8}$$

where f and g are functions defined on the interval $[\lambda_1, \lambda_m]$. The polynomials W_k, $W_k(0) = 1$, which minimize (2.6) are the orthogonal polynomials with respect to the inner product (2.8), i.e.,

$$(W_k, W_i) = \sum_{j=1}^{m} |c_j|^2 \beta_j \lambda_j W_k(\lambda_j)\, \overline{W_i(\lambda_j)} = 0 \tag{2.9}$$

for $k \neq i$. From the orthogonality of W_k it follows that they satisfy a three-term recurrence formula. We choose a different form of the three-term recurrence formula than usual in order to emphasize the connection between the cg iteration and the gradient one. This form is defined as follows:

$$\begin{aligned} &W_0(\lambda) \equiv 1, \\ &W_1(\lambda) = 1 - c_0 \lambda, \\ &W_{k+1}(\lambda) = W_k(\lambda) - c_k \lambda W_k(\lambda) - u_k \{ W_{k-1}(\lambda) - W_k(\lambda) + c_k \lambda W_k(\lambda) \}, \qquad k \geqslant 1, \end{aligned} \tag{2.10}$$

where

$$c_k = \frac{(W_k, W_k)}{(\lambda W_k, W_k)}, \tag{2.11}$$

$$u_0 = 0, \qquad u_k = \frac{\left(W_k - c_k \lambda W_k, \frac{1}{\lambda}(W_{k-1} - W_k) + c_k W_k \right)}{\left(W_{k-1} - W_k + c_k \lambda W_k, \frac{1}{\lambda}(W_{k-1} - W_k) + c_k W_k \right)}, \quad k \geqslant 1. \tag{2.12}$$

From this we get the three-term recurrence formula for the sequence $\{\mathbf{x}_k\}$,

$$\begin{aligned} \mathbf{z}_k &= \mathbf{x}_k - c_k \mathbf{r}_k, \qquad \mathbf{r}_k = A\mathbf{x}_k - \mathbf{b} \\ \mathbf{x}_{k+1} &= \mathbf{z}_k - u_k \mathbf{y}_k, \qquad \mathbf{y}_k = \mathbf{x}_{k-1} - \mathbf{z}_k. \end{aligned} \tag{2.13}$$

From (2.11), (2.12) and (2.9) we get

$$c_k = \frac{(\mathbf{r}_k, B(\mathbf{x}_k - \boldsymbol{\alpha}))}{(\mathbf{r}_k, B\mathbf{r}_k)}, \qquad u_0 = 0, \qquad u_k = \frac{(\mathbf{y}_k, B(\mathbf{z}_k - \boldsymbol{\alpha}))}{(\mathbf{y}_k, B\mathbf{y}_k)}, \quad k \geqslant 1. \tag{2.14}$$

The conjugate-gradient iteration (2.13) consists of computing $\mathbf{z}_k$ and $\mathbf{x}_{k+1}$. The vector $\mathbf{z}_k$ is obtained by one step of the gradient iteration (2.2) and is the best approximation of $\boldsymbol{\alpha}$ along the line $\mathbf{r}_k$. The vector $\mathbf{x}_{k+1}$ is the best approximation of $\boldsymbol{\alpha}$ along the line $\mathbf{y}_k$, $\|\mathbf{x}_{k+1} - \boldsymbol{\alpha}\|_B = \inf_u \|\mathbf{z}_k - \boldsymbol{\alpha} - u\mathbf{y}_k\|_B$.

From the orthogonality of W_k it follows that

$$\begin{aligned} (B(x_k - \alpha), r_j) &= 0 \quad \text{for} \quad j \neq k, \\ (B(x_k - \alpha), x_j - x_{j-1}) &= 0 \quad \text{for} \quad j \leqslant k. \end{aligned} \tag{2.15}$$

This yields $u_k = c_k(\mathbf{r}_k, B(\mathbf{x}_k - \boldsymbol{\alpha}))/(\mathbf{y}_k, B\mathbf{y}_k)$, which can be computed for $B = I$ whenever $A = M^*M$.

The conjugate-grandient method (2.13) converges in exactly m steps, i.e.,

$$\mathbf{x}_k = \boldsymbol{\alpha} \qquad \text{for} \quad k \geqslant m. \tag{2.16}$$

From (2.6) one can estimate the speed of convergence for initial approximations $\mathbf{x}_k$, $k \leqslant m$. Setting $P(\lambda) = T_k(f(\lambda))/T_k(f(0))$ in (2.6), where T_k is the kth Chebyshev polynomial of degree k and $f(\lambda) = (\lambda_m + \lambda_1 - 2\lambda)/(\lambda_m - \lambda_1)$, we get

$$\|\mathbf{x}_k - \alpha\|_B \leqslant \|P(A)(\mathbf{x}_0 - \boldsymbol{\alpha})\|_B \leqslant 2\left(\frac{\sqrt{a} - 1}{\sqrt{a} + 1}\right)^k \|\mathbf{x}_0 - \boldsymbol{\alpha}\|_B, \tag{2.17}$$

where $a = \lambda_m/\lambda_1$. [Compare with (2.4).] For the spectral norm we have

$$\|\mathbf{x}_k - \boldsymbol{\alpha}\| \leqslant 2\sqrt{\frac{\beta_{\max}}{\beta_{\max}}} \left(\frac{\sqrt{a} - 1}{\sqrt{a} + 1}\right)^k \|\mathbf{x}_0 - \boldsymbol{\alpha}\|, \tag{2.18}$$

where

$$\beta_{\min} = \min_{1 \leqslant j \leqslant m} \beta_j, \qquad \beta_{\max} = \max_{1 \leqslant j \leqslant m} \beta_j.$$

(For instance, if $B = A^p$, then $\beta_{\max}/\beta_{\min} = (\lambda_m/\lambda_1)^p \leqslant \kappa^p$.)

It seems to us that the choice $B=A^p$ for $p=0$, 1 or 2 covers all cases of practical interest.

For $B=A^0=I$ we minimize $\|\mathbf{x}_k-\boldsymbol{\alpha}\|$. To compute the coefficients c_k and u_k in (2.14) we assume in this case that $A=M^*M$ and $\mathbf{b}=M^*\mathbf{g}$ for a nonsingular M (M and $\mathbf{g}$ being given as data). This variant of the cg iteration is called the *minimum-error iteration.*

For $B=A^1$ we minimize $\|A^{1/2}(\mathbf{x}_k-\boldsymbol{\alpha})\|$. This corresponds to the *classical conjugate-gradient iteration.*

For $B=A^2$ we minimize the residual vectors $\mathbf{r}_k$, since $\|\mathbf{x}_k-\boldsymbol{\alpha}\|_B=\|A(\mathbf{x}_k-\boldsymbol{\alpha}\|=\|\mathbf{r}_k\|$. This variant is called the *minimal-residual iteration.*

3. ROUNDOFF-ERROR ANALYSIS OF GRADIENT ALGORITHMS

We shall show that the roundoff-error analysis of cg algorithms belonging to Φ can be primarily based on the roundoff-error analysis of the gradient algorithms to be studied in this section. Therefore in this section we analyze gradient algorithms in the presence of rounding errors. We focus our attention on a steepest-descent algorithm ($B=A$) and mention the corresponding results for the gradient algorithms with $B=I$ or $B=A^2$.

We consider a steepest descent algorithm in floating-point binary arithmetic (fl) with the relative computer precision $\zeta=2^{-t}$, where t is the number of mantissa bits. To simplify further estimates we shall use the relation $\underset{1}{=}$, which is defined as follows. Let f and h be two scalar functions defined on $[0,\zeta_0]$. By

$$f(\zeta) \underset{1}{=} h(\zeta)$$

we mean that there exists a constant c such that $f(\zeta)=h(\zeta)[1+\epsilon(\zeta)]$, where $|\epsilon(\zeta)|\leqslant c\zeta$ for $0\leqslant\zeta\leqslant\zeta_0$. By

$$f(\zeta) \underset{1}{\leqslant} h(\zeta)$$

we mean

$$f(\zeta)\leqslant h(\zeta) \quad \text{or} \quad f(\zeta) \underset{1}{=} h(\zeta).$$

The relation $\underset{1}{\leqslant}$ enables us to ignore the terms of order ζ^2 in the presence of the term of order ζ.

Let $\mathbf{x}_k$ and $\mathbf{r}_k$ denote the vectors computed in fl by an algorithm. We assume that

$$\mathbf{r}_k=\mathrm{fl}(A\mathbf{x}_k-\mathbf{b}). \tag{3.1}$$

Let $\mathbf{r}_k^* = A\mathbf{x}_k - \mathbf{b}$ denote the exact value of the residual vector, and let

$$\mathbf{e}_k = A^{1/2}(\mathbf{x}_k - \boldsymbol{\alpha}), \qquad e_k = \|\mathbf{e}_k\|. \tag{3.2}$$

We analyze (3.1). Assume that the algorithm for evaluation of $\mathbf{r}_k$ satisfies the relation

$$\mathbf{r}_k = (I + D_k^1)\left[(A + E_k^1)\mathbf{x}_k - \mathbf{b}\right] = \mathbf{r}_k^* + \delta\mathbf{r}_k, \tag{3.3}$$

where D_k^1 is a diagonal matrix such that $\|D_k^1\| \leqslant \zeta$ and $\|E_k^1\| \leqslant \zeta\|A\|C_1$ with the constant C_1 depending only on the size of the problem: $C_1 = C_1(n)$. Hence

$$\delta\mathbf{r}_k = E_k^1\mathbf{x}_k + D_k^1(\mathbf{r}_k^* + E_k^1\mathbf{x}_k),$$

$$\|\delta\mathbf{r}_k\| \underset{1}{\leqslant} \zeta\|A\|\,\|\mathbf{x}_k\|C_1 + \zeta\|\mathbf{r}_k^*\|. \tag{3.4}$$

To assure that $\mathbf{r}_k$ is a reasonable approximation to $\mathbf{r}_k^*$ we assume that $\|\mathbf{r}_k\| > \zeta\|A\|\,\|\mathbf{x}_k\|C_1$. Note that the opposite inequality $\|\mathbf{r}_k\| \leqslant \zeta\|A\|\,\|\mathbf{x}_k\|C_1$ means that $\mathbf{x}_k$ is the exact solution of the system $(A - \delta A)\mathbf{x}_k = \mathbf{b}$, where

$$\delta A = \frac{(\mathbf{r}_k - \delta\mathbf{r}_k)\mathbf{x}_k^T}{\|\mathbf{x}_k\|^2} \quad \text{and} \quad \|\delta A\| \underset{1}{\leqslant} \zeta\|A\|2C_1.$$

This means that the algorithm is well behaved and the iteration should be terminated. Therefore we shall analyze the following algorithm of the steepest-descent iteration.

ALGORITHM 3.1. Let $\mathbf{x}_0$ be given and let $k = 0$.

(i) Compute $\mathbf{r}_k = \mathrm{fl}(A\mathbf{x}_k - \mathbf{b})$.
(ii) If $\|\mathbf{r}_k\| \leqslant \zeta\|A\|\,\|\mathbf{x}_k\|C_1$ then formally set $\mathbf{x}_i = \mathbf{x}_k$. $\forall i \geqslant k+1$. STOP.
(iii) If $\|\mathbf{r}_k\| > \zeta\|A\|\,\|\mathbf{x}_k\|C_1$ then compute

$$c_k = \mathrm{fl}\left(\frac{(\mathbf{r}_k, \mathbf{r}_k)}{(\mathbf{r}_k, A\mathbf{r}_k)}\right), \tag{3.5}$$

$$\mathbf{x}_{k+1} = \mathrm{fl}(\mathbf{x}_k - c_k\mathbf{r}_k). \tag{3.6}$$

$$k := k+1. \text{ GO TO (i).}$$

We do not define a termination criterion for Algorithm 3.1 unless it is well behaved. We want to verify its numerical stability by $\overline{\lim}\|\mathbf{x}_k - \boldsymbol{\alpha}\|$, and

$\mathbf{x}_k$ has to be defined for all k. Therefore we formally set $\mathbf{x}_i = \mathbf{x}_k$ $\forall i \geq k+1$ in (ii).

Recall that the computed inner product $(\mathbf{a}, \mathbf{b})$ in fl satisfies the relation

$$\mathrm{fl}((\mathbf{a}, \mathbf{b})) = ((I+D)\mathbf{a}, \mathbf{b}), \tag{3.7}$$

where D is a diagonal matrix such that $\|D\| \leq \zeta C_2$. The constant C_2 depends on a particular algorithm used for the summation of n numbers. For the standard algorithm $C_2 = n$, whereas when the Moller algorithm is employed $C_2 = 3$. See [9] for the first and [5] for the second result.

We are ready to prove

LEMMA 3.1. *Suppose that* $\|\mathbf{r}_k\| > \zeta \|A\| \|\mathbf{x}_k\| C_1$ $\forall k$. *Then the sequence* $\{\mathbf{x}_k\}$ *computed by Algorithm 3.1 satisfies the following error formula:*

$$e_{k+1} \underset{1}{\leq} \sqrt{e_k^2 - c_k^* \|\mathbf{r}_k^*\|^2} + \zeta \|A^{1/2}\| \|\mathbf{x}_k\| + \zeta c_k^* \{\|A^{3/2}\| \|\mathbf{x}_k\| 5C_1 + \|A\| e_k (C_1 + 2C_2 + 8)\}, \tag{3.8}$$

where $\mathbf{r}_k^* = A\mathbf{x}_k - \mathbf{b}$, $c_k^* = (\mathbf{r}_k^*, \mathbf{r}_k^*)/(\mathbf{r}_k^*, A\mathbf{r}_k^*)$.

Proof. We analyze the computation of c_k. Due to the assumption that $\|\mathbf{r}_k\| > \zeta \|A\| \|\mathbf{x}_k\| C_1$ $\forall k$, (3.3) and (3.4) imply that $\mathbf{r}_k^* \neq \mathbf{0}$ and c_k^* is well defined. We have

$$c_k = \frac{\left((I+D_k^2)\mathbf{r}_k, \mathbf{r}_k\right)}{\left((I+D_k^3)(A+E_k^2)\mathbf{r}_k, \mathbf{r}_k\right)} (1+\epsilon_k^1) = \frac{(\mathbf{r}_k, \mathbf{r}_k)}{(\mathbf{r}_k, A\mathbf{r}_k)} (1+\epsilon_k^2), \tag{3.9}$$

where $\|D_k^i\| \leq \zeta C_2$ for $i = 2$ and 3, $\|E_k^2\| \leq \zeta \|A\| C_1$, ϵ_k^1 is the relative error of division, $|\epsilon_k^1| \leq \zeta$, and

$$1+\epsilon_k^2 = \left(1 + \frac{(D_k^2 \mathbf{r}_k, \mathbf{r}_k)}{(\mathbf{r}_k, \mathbf{r}_k)}\right)\left(1 + \frac{\left(\left[E_k^2 + D_k^3(A+E_k^2)\right]\mathbf{r}_k, \mathbf{r}_k\right)}{(\mathbf{r}_k, A\mathbf{r}_k)}\right)^{-1} (1+\epsilon_k^1),$$

$$|\epsilon_k^2| \underset{1}{\leq} \zeta \left(C_2 + 1 + \|A\| (C_1 + C_2) \frac{(\mathbf{r}_k, \mathbf{r}_k)}{(\mathbf{r}_k, A\mathbf{r}_k)}\right). \tag{3.10}$$

From (3.3) we get

$$\frac{(\mathbf{r}_k,\mathbf{r}_k)}{(\mathbf{r}_k,A\mathbf{r}_k)} = c_k^*(1+\epsilon_k^3),$$

$$|\epsilon_k^3| \underset{1}{\leqslant} \frac{2\|\delta\mathbf{r}_k\|}{\|\mathbf{r}_k^*\|} + \frac{2\|A^{1/2}\|\,\|\delta\mathbf{r}_k\|}{\|A^{1/2}\mathbf{r}_k^*\|}$$

$$\underset{1}{\leqslant} 2\zeta(\|A\|\,\|\mathbf{x}_k\|C_1 + \|\mathbf{r}_k^*\|)\left(\frac{1}{\|\mathbf{r}_k^*\|} + \frac{\|A^{1/2}\|}{\|A^{1/2}\mathbf{r}_k^*\|}\right)$$

$$\underset{1}{\leqslant} 4\zeta\|A^{1/2}\|\frac{\|A\|\,\|\mathbf{x}_k\|C_1 + \|\mathbf{r}_k^*\|}{\|A^{1/2}\mathbf{r}_k^*\|}. \tag{3.11}$$

From (3.9) and (3.11) we have

$$c_k = c_k^*(1+\delta c_k), \quad |\delta c_k| \underset{1}{\leqslant} |\epsilon_k^2| + |\epsilon_k^3|$$

$$\underset{1}{\leqslant} \zeta\left[C_2 + 1 + c_k^*\|A\|(C_1+C_2) + 4\|A^{1/2}\|\frac{\|A\|\,\|\mathbf{x}_k\|C_1 + \|\mathbf{r}_k^*\|}{\|A^{1/2}\mathbf{r}_k^*\|}\right]. \tag{3.12}$$

We now analyze (3.6). We have

$$\mathbf{x}_{k+1} = (I+D_k^5)\left[\mathbf{x}_k - (I+D_k^4)c_k\mathbf{r}_k\right] = \mathbf{x}_k - c_k^*\mathbf{r}_k^* + \delta\mathbf{x}_{k+1}, \tag{3.13}$$

where $\|D_k^i\| \leqslant \zeta$ for $i=4$ and 5, and

$$\delta\mathbf{x}_{k+1} = -c_k^*\,\delta\mathbf{r}_k - c_k^*\,\delta c_k\mathbf{r}_k^* - D_k^4 c_k^*\mathbf{r}_k^* + D_k^5(x_k - c_k^*\mathbf{r}_k^*) + O(\zeta^2). \tag{3.14}$$

From (3.4), (3.12) and (3.14) we get

$$\mathbf{e}_{k+1} = \mathbf{e}_k - c_k^*A^{1/2}\mathbf{r}_k^* + A^{1/2}\delta\mathbf{x}_{k+1},$$

$$\|A^{1/2}\delta\mathbf{x}_{k+1}\| \underset{1}{\leqslant} \zeta\|A^{1/2}\|\,\|\mathbf{x}_k\| + \zeta c_k^*\left[\|A^{3/2}\|\,\|\mathbf{x}_k\|5C_1\right. \tag{3.15}$$

$$\left. + \|A\|e_k(C_2+8)\right] + \zeta(c_k^*)^2\|A\|\,\|A^{1/2}\mathbf{r}_k^*\|(C_1+C_2).$$

Note that

$$c_k^* \|A^{1/2}\mathbf{r}_k^*\| = \frac{(A^{1/2}\mathbf{r}_k^*, \mathbf{e}_k)}{\|A^{1/2}\mathbf{r}_k^*\|} \leqslant e_k.$$

From this and (2.4) we get (3.8), which completes the proof of Lemma 3.1. ■

Lemma 3.1 shows how the error e_{k+1} depends on the theoretical and rounding errors. It is interesting to notice that the bound on the rounding error increases with c_k^*, whereas the bound on the theoretical error decreases with increasing c_k^*.

We want to find the limiting properties of the sequence $\{e_k\}$ which satifies (3.8). To achieve this we use the following lemma.

LEMMA 3.2. *Let*

$$e_{k+1} \leqslant \sqrt{e_k^2 - c_k^* \|\mathbf{r}_k^*\|^2} + a_k + c_k^* b_k + c_k^* e_k d$$

for given nonnegative sequences $\{a_k\}$, $\{b_k\}$ *and a constant* d *such that* $2d\|A^{-1}\| < 1$. *Then*

$$\overline{\lim_k}\, e_k \leqslant 3\kappa \frac{\overline{\lim_k} \dfrac{b_k}{\|A\|} + \overline{\lim_k}\, a_k}{1 - \dfrac{2d\kappa}{\|A\|}}. \tag{3.16}$$

Proof. Let ϵ be any positive number. Choose k_0 such that $a_k - \epsilon \leqslant a = \overline{\lim}_k a_k$, $b_k - \epsilon \leqslant b = \overline{\lim}_k b_k$ for $k \geqslant k_0$. Then $e_{k+1} \leqslant f(c_k^*)$, where

$$f(c) = \sqrt{e_k^2 - c\|\mathbf{r}_k^*\|^2} + (a+\epsilon) + c(b+\epsilon) + ce_k d$$

for $c \in [\|A\|^{-1}, \|A^{-1}\|]$. Consider two cases.

Case (i). Assume that $e_k \geqslant 2\|A^{-1}\|(b+\epsilon)/(1-2\|A^{-1}\|d)$ $\forall k \geqslant k_0$. We show that f is decreasing for $c \geqslant 0$. Indeed,

$$f'(c) = \frac{-\|\mathbf{r}_k^*\|^2}{2\sqrt{e_k^2 - c\|\mathbf{r}_k^*\|^2}} + b + \epsilon + de_k$$

$$\leqslant f'(0) = \frac{-\|\mathbf{r}_k^*\|^2}{2e_k} + b + \epsilon + de_k.$$

Since

$$de_k - \frac{\|r_k^*\|^2}{2e_k} \leqslant \left(d - \frac{1}{2\|A^{-1}\|}\right)e_k = \frac{-1}{2\|A^{-1}\|}(1-2d\|A^{-1}\|)e_k$$

$$\leqslant -(b+\epsilon),$$

we get $f'(c) \leqslant 0$. Thus

$$f(c) \leqslant f(\|A\|^{-1}) = \sqrt{e_k^2 - \frac{\|r_k^*\|^2}{\|A\|}} + a + \epsilon + \frac{b+\epsilon}{\|A\|} + \frac{e_k d}{\|A\|}.$$

Since $\sqrt{e_k^2 - \|r_k^*\|^2/\|A\|} \leqslant \sqrt{1-\kappa^{-1}}\, e_k$, we get

$$e_{k+1} \leqslant \sqrt{1-\kappa^{-1}}\, e_k + (a+\epsilon) + \frac{b+\epsilon}{\|A\|} + e_k d\|A\|. \tag{3.17}$$

Note that $\{e_k\}$ is a nonincreasing sequence and

$$\overline{\lim_k}\, e_k \leqslant \frac{a+\epsilon+(b+\epsilon)/\|A\|}{1-\sqrt{1-\kappa^{-1}} - d/\|A\|}.$$

Since $1-\sqrt{1-\kappa^{-1}} \geqslant 1/(2\kappa)$, we finally get

$$\overline{\lim_k}\, e_k \leqslant \beta_0 \overset{\text{df}}{=} 2\kappa \frac{a+\epsilon+(b+\epsilon)/\|A\|}{1-2d\kappa/\|A\|}. \tag{3.18}$$

Case (ii). Assume there exists k_1 such that $e_{k_1} < 2\|A^{-1}\|(b+\epsilon)/(1-2\|A^{-1}\|d) \leqslant \beta_0$. We prove inductively that

$$e_k \leqslant 1.5\beta_0, \qquad k \geqslant k_1. \tag{3.19}$$

This holds for $k = k_1$. Suppose first that e_k satisfies (3.19) and additionally $e_k \geqslant \beta_0$. From (3.17) we get

$$e_{k+1} \leqslant \sqrt{1-\kappa^{-1}}\, e_k + a + \epsilon + \frac{b+\epsilon}{\|A\|} + \frac{e_k d}{\|A\|} < e_k \leqslant 1.5\beta_0.$$

If $e_k < \beta_0$, then

$$\begin{aligned} e_{k+1} &\leq e_k + a + \epsilon + (b+\epsilon)\|A^{-1}\| + \|A^{-1}\| e_k d \\ &\leq \left(1 + \frac{d\kappa}{\|A\|}\right) + \left(a + \epsilon + \frac{b+\epsilon}{\|A\|}\right)\kappa \\ &= \beta_0\left[1 + \frac{d\kappa}{\|A\|} + 0.5\left(1 - \frac{2d\kappa}{\|A\|}\right)\right] \\ &= 1.5\beta_0. \end{aligned}$$

Hence, in all cases we proved that $\overline{\lim}_k e_k \leq 1.5\beta_0$. Letting ϵ tend to zero, we get (3.16). This completes the proof. ■

From Lemmas 3.1 and 3.2 we immediately conclude the asymptotic behavior of the sequence $\{\mathbf{x}_n\}$ computed by Algorithm 3.1.

THEOREM 3.1. *If* $\beta \overset{\text{df}}{=} 2\zeta\kappa(C_1 + 2C_2 + 8) < 1$, *then Algorithm* 3.1 *computes the sequence* $\{\mathbf{x}_k\}$ *such that*

$$\overline{\lim_k} \|A^{1/2}(\mathbf{x}_k - \boldsymbol{\alpha})\| \underset{1}{\leq} \zeta\kappa \frac{3(5C_1+1)}{1-\beta} \|A^{1/2}\| \overline{\lim_k} \|\mathbf{x}_k\|. \tag{3.20}$$

Proof. Suppose first that there exists k_0 such that $\|\mathbf{r}_{k_0}\| \leq \zeta \|A\| \|\mathbf{x}_{k_0}\| C_1$. Then Algorithm 3.1 yields $\mathbf{x}_i = x_{k_0}$ for $i \geq k_0$. From (3.3) and (3.4) we get

$$\begin{aligned} \overline{\lim_k} \|A^{1/2}(\mathbf{x}_k - \boldsymbol{\alpha})\| &= \|A^{1/2}(\mathbf{x}_{k_0} - \boldsymbol{\alpha})\| \leq \|A^{-1/2}\|\left(\|\mathbf{r}_{k_0}\| + \|\delta\mathbf{r}_{k_0}\|\right) \\ &\underset{1}{\leq} \zeta\kappa^{1/2} 2C_1 \|A^{1/2}\| \overline{\lim_k} \|\mathbf{x}_k\|, \end{aligned}$$

which obviously proves (3.20).

Hence we can now assume that $\|\mathbf{r}_k\| > \zeta \|A\| \|\mathbf{x}_k\| C_1 \quad \forall k$. Applying Lemma 3.2 with $a_k = \zeta \|A^{1/2}\| \|\mathbf{x}_k\|$, $b_k = \zeta \|A^{3/2}\| \|\mathbf{x}_k\| 5C_1$ and $d = \zeta \|A\|(C_1 + 2C_2 + 8)$, we get (3.20) from Lemma 3.1. ■

Theorem 3.1 states that if k is large and $\|\mathbf{x}_k\| \cong \overline{\lim_k} \|\mathbf{x}_k\|$, then the computed $\mathbf{x}_k$ approximates $\boldsymbol{\alpha}$ with the error

$$\|A^{1/2}(\mathbf{x}_k - \boldsymbol{\alpha})\| \leq \zeta\kappa \|A^{1/2}\| \|\mathbf{x}_k\| C, \tag{3.21}$$

where $C=3(5C_1+1)+O(\zeta)$. Note that (3.21) does not imply the numerical stability of Algorithm 3.1, since we have $\|A^{1/2}\|\,\|\mathbf{x}_k\|$ instead of $\|A^{1/2}\mathbf{x}_k\|$. From (3.21) we get

$$\frac{\|\mathbf{x}_k-\boldsymbol{\alpha}\|_A}{\|\mathbf{x}_k\|_A}\leqslant\zeta\kappa\left(\frac{\|A^{1/2}\|\,\|\mathbf{x}_k\|}{\|A^{1/2}\mathbf{x}_k\|}\right)C\leqslant\zeta\kappa^{3/2}C. \tag{3.22}$$

This means that the relative error of $\mathbf{x}_k$ in the A-norm depends at worst on $\zeta\kappa^{3/2}$. However if $\|A^{1/2}\|\,\|\mathbf{x}_k\|\cong\|A^{1/2}\mathbf{x}_k\|$, then *Algorithm 3.1 is numerically stable in the A-norm.*

We pass to results for the spectral norm. From (3.21) we have

$$\frac{\|\mathbf{x}_k-\boldsymbol{\alpha}\|}{\|\mathbf{x}_k\|}\leqslant\frac{\|A^{-1/2}\|\,\|A^{1/2}(\mathbf{x}_k-\boldsymbol{\alpha})\|}{\|\mathbf{x}_k\|}\leqslant\zeta\kappa^{3/2}C. \tag{3.23}$$

However, if

$$\|A^{1/2}(\mathbf{x}_k-\boldsymbol{\alpha})\|\cong\|A^{1/2}\|\,\|\mathbf{x}_k-\boldsymbol{\alpha}\|, \tag{3.24}$$

then we get *numerical stability in the spectral norm.* Note that (3.24) will often hold. For instance, let $\mathbf{x}_k-\boldsymbol{\alpha}=\sum_{i=1}^{n}c_i\boldsymbol{\xi}_i$, where $\boldsymbol{\xi}_i$ are the eigenvectors of A. Suppose that $c_i=c$ (or $c_i\cong c$) for all i. Then $\|A^{1/2}(\mathbf{x}_k-\boldsymbol{\alpha})\|=\|A^{1/2}\|\,|c|(\sum_{i=1}^{n}(\lambda_i/\lambda_{\max})^2)^{1/2}$ and $\|A^{1/2}\|\,\|x_k-\alpha\|=\|A^{1/2}\|\,|c|\sqrt{n}$, and these two quantities differ at most by a factor of $\sqrt{n}$. Thus, (3.24) holds.

For the residual vector $\mathbf{r}_k^*=A\mathbf{x}_k-\mathbf{b}$ we get

$$\|\mathbf{r}_k^*\|\leqslant\zeta\kappa\left(\frac{\|A\mathbf{x}_k-\mathbf{b}\|}{\|A^{1/2}\|\,\|A^{1/2}(\mathbf{x}_k-\boldsymbol{\alpha})\|}\right)\|A\|\,\|\mathbf{x}_k\|C$$

$$\leqslant\zeta\kappa\|A\|\,\|\mathbf{x}_k\|C. \tag{3.25}$$

Numerical tests confirm that the residual vectors sometimes depends on $\zeta\kappa$. This means that Algorithm 3.1 is *not well behaved.* However, if $\|A\mathbf{x}_k-\mathbf{b}\|\cong\|A^{1/2}(\mathbf{x}_k-\boldsymbol{\alpha})\|/\|A^{-1/2}\|$, then the residual vector $\mathbf{r}_k^*$ depends at worst on $\zeta\kappa^{1/2}$.

Numerical stability and/or the well-behaved property may be achieved by the use of iterative refinement even if the residuals are computed in single precision. From Theorems 3.1 and 4.3 in [4] it follows that *Algorithm 3.1 with iterative refinement in single precision is numerically stable* whenever $\zeta\kappa^{3/2}C<1$, and it is *well behaved* whenever $\zeta\kappa^2$ is at most of order unity.

We summarize the properties of Algorithm 3.1.

COROLLARY 3.1. *Algorithm 3.1 constructs an approximation* $\mathbf{x}_k$ *such that*

$$\|A^{1/2}(\mathbf{x}_k-\boldsymbol{\alpha})\| \leqslant \zeta\kappa\|A^{1/2}\|\,\|\mathbf{x}_k\|C,$$

$$\|\mathbf{x}_k-\boldsymbol{\alpha}\| \leqslant \zeta\kappa^{3/2}\|\mathbf{x}_k\|C,$$

$$\|A\mathbf{x}_k-\mathbf{b}\| \leqslant \zeta\kappa\|A\|\,\|\mathbf{x}_k\|C,$$

where $C=3(5C_1+1)+O(\zeta)$. *Furthermore, if* $\|A^{1/2}\|\,\|\mathbf{x}_k\| \cong \|A^{1/2}\mathbf{x}_k\|$, *then the algorithm is numerically stable in the A-norm, and if* $\|A^{1/2}\|\|\mathbf{x}_k-\boldsymbol{\alpha}\| \cong \|A^{1/2}(\mathbf{x}_k-\boldsymbol{\alpha})\|$, *then the algorithm is numerically stable in the spectral norm.*

Corollary 3.1 summarizes the numerical properties of the steepest-descent algorithm. It shows that the algorithm may be neither well behaved nor numerically stable. However, the algorithm is guaranteed to compute an approximation with a realtive error of order at most $\zeta\kappa^{3/2}$. If the problem is not too ill conditioned, this is a satisfactory result.

We end this section by a remark concerning the gradient algorithms for $B=I$ or $B=A^2$. They differ from Algorithm 3.1 by different computations of c_k in (3.5). Based on proof techniques similar to those used here, it is possible to show that there exists an index k such that

$$\|B^{1/2}(\mathbf{x}_k-\boldsymbol{\alpha})\| \leqslant \zeta\kappa\|B^{1/2}\|\,\|\mathbf{x}_k\|C,$$

$$\|\mathbf{x}_k-\boldsymbol{\alpha}\| \leqslant \zeta\kappa\|B^{1/2}\|\,\|B^{-1/2}\|\,\|\mathbf{x}_k\|C,$$

$$\|A\mathbf{x}_k-\mathbf{b}\| \leqslant \zeta\kappa\|A\|\,\|\mathbf{x}_k\|C$$

for a certain $C=C(n)$. This shows that the best estimates are obtained in the "natural" norm of the algorithm (i.e., in the B-norm) and that the residual vectors may depend on $\zeta\kappa$ for every choice of B.

4. ROUNDOFF-ERROR ANALYSIS OF A CLASS OF CONJUGATE-GRADIENT ALGORITHMS

We deal with the conjugate-gradient iteration for $B=A$ which generates the sequence $\{\mathbf{x}_k\}$ as follows:

$$\begin{aligned} \mathbf{z}_k &= \mathbf{x}_k - c_k\mathbf{r}_k, \qquad \mathbf{r}_k = A\mathbf{x}_k-\mathbf{b}, \\ \mathbf{x}_{k+1} &= \mathbf{z}_k - u_k\mathbf{y}_k, \qquad \mathbf{y}_k = \mathbf{x}_{k-1}-\mathbf{z}_k, \end{aligned} \tag{4.1}$$

where

$$c_k = \frac{(\mathbf{r}_k, \mathbf{r}_k)}{(\mathbf{r}_k, A\mathbf{r}_k)}, \qquad u_0 = 0, \qquad u_k = \frac{(\mathbf{y}_k, A(\mathbf{z}_k - \alpha))}{(\mathbf{y}_k, A\mathbf{y}_k)}, \quad k \geqslant 1. \tag{4.2}$$

See (2.13) and (2.14). It was pointed out to the author by Wieladek [7] that (4.1) has an interesting local property. Namely, no matter how the vectors $\mathbf{z}_k$ and $\mathbf{y}_k$ are computed, the coefficient u_k is chosen in such a way that the error $\|\mathbf{x}_{k+1} - \alpha\|_A$ is minimized along the line $\mathbf{y}_k$. Note that the cost of one step of the cg iteration depends on how one computes the residual vectors and the coefficients c_k and u_k. The number of matrix-vector multiplications needed to perform one step may vary from one to four.

We define a new class of cg algorithms Φ by the following properties. We assume that any algorithm φ from the class Φ computes the vector $\mathbf{z}_k$ by Algorithm 3.1. That is,

$$\mathbf{r}_k = \mathrm{fl}(A\mathbf{x}_k - \mathbf{b}), \tag{4.3}$$

and if $\|\mathbf{r}_k\| > \zeta \|A\| \, \|\mathbf{x}_k\| C_1$, then

$$c_k = \mathrm{fl}\left(\frac{(\mathbf{r}_k, \mathbf{r}_k)}{(\mathbf{r}_k, A\mathbf{r}_k)} \right), \tag{4.4}$$

$$\mathbf{z}_k = \mathrm{fl}(\mathbf{x}_k - c_k \mathbf{r}_k). \tag{4.5}$$

Thus, the computation of $\mathbf{z}_k$ may require two matrix-vector multiplications.

There are many different ways of computing the coefficient u_k. One may use theoretical orthogonality relations (2.15) as well as the direct substitutions for $\mathbf{y}_k$ and $\mathbf{z}_k$ from (4.1). For instance, it follows from (2.15) and (4.1) that in theory $u_k > 0$. For the sake of generality we do not specify an algorithm for the computation of u_k. We only assume that an algorithm φ computes $\tilde{u}_k$ such that

$$\tilde{u}_k = u_k(1 + \delta u_k), \qquad |\delta u_k| \leqslant 1, \tag{4.6}$$

where $u_k = (\mathbf{y}_k, A(\mathbf{z}_k - \alpha))/(\mathbf{y}_k, A\mathbf{y}_k)$ for the computed vectors $\mathbf{z}_k$ and $\mathbf{y}_k = \mathrm{fl}(\mathbf{x}_{k-1} - \mathbf{z}_k)$. Note that (4.6) means that $\tilde{u}_k$ can be a very crude approximation of u_k. A particular algorithm φ for which (4.6) holds is given in Example 4.1. Knowing $\tilde{u}_k$, we finally compute

$$\mathbf{x}_{k+1} = \mathrm{fl}(\mathbf{z}_1 - \tilde{u}_k \mathbf{y}_k). \tag{4.7}$$

Thus the class Φ contains algorithms which differ by the computation of u_k. We are ready to prove

LEMMA 4.1. *Let φ be a cg algorithm defined by* (4.3) *to* (4.7). *Then*

$$\|\mathbf{x}_{k+1}-\boldsymbol{\alpha}\|_A \leqslant (1+2\zeta)\|\mathbf{z}_k-\boldsymbol{\alpha}\|_A + \frac{\zeta\|A^{1/2}\|\,\|\mathbf{x}_{k+1}\|}{1-\zeta}. \tag{4.8}$$

Proof. From (4.7) we have

$$\mathbf{x}_{k+1}=\left(I+D_k^7\right)\left[\mathbf{z}_k-\tilde{u}_k\left(I+D_k^6\right)\mathbf{y}_k\right],$$

where D_k^6 and D_k^7 are diagonal matrices and $\|D_k^i\| \leqslant \zeta$ for $i=6$ and 7. Thus

$$\begin{aligned}\mathbf{x}_{k+1}&=\mathbf{z}_k-\tilde{u}_k\mathbf{y}_k+\delta\mathbf{x}_{k+1},\\ \delta\mathbf{x}_{k+1}&=-\tilde{u}_kD_k^6\mathbf{y}_k+\left[I-\left(I+D_k^7\right)^{-1}\right]\mathbf{x}_{k+1}.\end{aligned} \tag{4.9}$$

From (4.6) we get

$$\begin{aligned}\|\delta\mathbf{x}_{k+1}\|_A &\leqslant 2\zeta|u_k|\,\|\mathbf{y}_k\|\|\mathbf{y}_k\|_A+\frac{\zeta\|A^{1/2}\|\,\|\mathbf{x}_{k+1}\|}{1-\zeta}\\ &\leqslant 2\zeta\|\mathbf{z}_k-\boldsymbol{\alpha}\|_A+\frac{\zeta\|A^{1/2}\|\,\|\mathbf{x}_{k+1}\|}{1-\zeta}.\end{aligned} \tag{4.10}$$

Let $\mathbf{x}(c)=\mathbf{z}_k-c\mathbf{y}_k$. Consider $f(c)=\|\mathbf{x}(c)-\boldsymbol{\alpha}\|_A$. It is easy to verify that $f(c)\leqslant\|\mathbf{z}_k-\boldsymbol{\alpha}\|_A$ for $|c|\leqslant 2|u_k|$ and $\operatorname{sign}(c)=\operatorname{sign}(u_k)$. Since the computed coefficient $\tilde{u}_k$ satisfies these conditions, (4.9) and (4.10) yield

$$\|\mathbf{x}_{k+1}-\boldsymbol{\alpha}\|_A \leqslant (1+2\zeta)\|\mathbf{z}_k-\boldsymbol{\alpha}\|_A + \frac{\zeta\|A^{1/2}\|\,\|\mathbf{x}_{k+1}\|}{1-\zeta}.$$

This proves (4.8). ∎

Lemma 4.1 expresses the error of $\mathbf{x}_{k+1}$ in terms of $\mathbf{z}_k$. Since $\mathbf{z}_k$ is obtained by one step of the steepest-descent algorithm, the error $\|\mathbf{z}_k-\boldsymbol{\alpha}\|_A$ satisfies (3.8). From Lemma 3.2 we immediately get the basic result of this paper.

THEOREM 4.1. *Let* $\beta \overset{\text{df}}{=} 2\zeta\kappa(C_1+2C_2+8)<1$. *Any cg algorithm* φ *from the class* Φ *computes the sequence* $\{\mathbf{x}_k\}$ *such that*

$$\overline{\lim_k} \|A^{1/2}(\mathbf{x}_k-\boldsymbol{\alpha})\| \leqslant \zeta\kappa\frac{3(5C_1+2)}{1-\beta} \overline{\lim_k} \|\mathbf{x}_k\|. \tag{4.11}$$

Theorem 4.1 states the numerical properties of the cg algorithms from the class Φ. Since (4.11) is essentially equivalent to (3.20), the discussion of numerical properties of the steepest-gradient algorithm is also valid for the cg algorithms from Φ. In particular we can estimate the error $\mathbf{x}_k-\boldsymbol{\alpha}$ and the residual vector $\mathbf{r}_k$ in the spectral norm, as in (3.22) to (3.25). This is summarized in the following corollary.

COROLLARY 4.1. *Any cg algorithm* φ *from the class* Φ *computes an approximation* $\mathbf{x}_k$ *such that*

$$\|A^{1/2}(\mathbf{x}_k-\boldsymbol{\alpha})\| \leqslant \zeta\kappa\|A^{1/2}\|\,\|\mathbf{x}_k\|C,$$

$$\|\mathbf{x}_k-\boldsymbol{\alpha}\| \leqslant \zeta\kappa^{3/2}\|\mathbf{x}_k\|C,$$

$$\|A\mathbf{x}_k-\mathbf{b}\| \leqslant \zeta\kappa\|A\|\,\|\mathbf{x}_k\|C,$$

where $C=3(5C_1+2)+O(\zeta)$. *Furthermore, if* $\|A^{1/2}\|\,\|\boldsymbol{\alpha}\| \cong \|A^{1/2}\boldsymbol{\alpha}\|$, *then the algorithm* φ *is numerically stable in the A-norm, and if* $\|A^{1/2}\|\,\|\mathbf{x}_k-\boldsymbol{\alpha}\| \cong \|A^{1/2}(\mathbf{x}_k-\boldsymbol{\alpha})\|$, *then the algorithm* φ *is stable in the spectral norm.*

Corollary 4.1 assures that the algorithm φ computes $\mathbf{x}_k$ with the relative error in the spectral norm depending at worst on $\zeta\kappa^{3/2}$. The residual vector has the spectral norm of order at most $\zeta\kappa$. We repeat that the algorithm φ with iterative refinement in single precision is numerically stable whenever $\zeta\kappa^{3/2}C<1$ and well behaved whenever $\zeta\kappa^2$ is at most of order unity.

We now give an example of an algorithm φ which satisfies (4.6).

EXAMPLE 4.1. Let $\mathbf{x}_k$ and $\mathbf{x}_{k-1}$ be the computed vectors and $\mathbf{r}_k, \mathbf{r}_{k-1}$ the corresponding residual vectors. Let $\mathbf{v}_k=\text{fl}(A\mathbf{r}_k)$ be the computed vector which is used for the computation of c_k.

We propose the following algorithm for the computation of u_k. Let

$$w_1 = \mathrm{fl}((\mathbf{y}_k, \mathbf{r}_k - c_k\mathbf{v}_k)),$$

$$w_2 = \mathrm{fl}((\mathbf{y}_k, \mathbf{r}_{k-1} - \mathbf{r}_k + c_k\mathbf{v}_k)).$$

Thus the computation of w_1 and w_2 does not require further matrix-vector multiplications. Repeating a part of the analysis of Sec. 3, it is possible to show that

$$w_1 = (\mathbf{y}_k, A(\mathbf{z}_k - \boldsymbol{\alpha})) + \delta w_1, \qquad |\delta w_1| \leqslant \zeta \|A\| \|\mathbf{y}_k\| \|\mathbf{x}_k\| C_3,$$

$$w_2 = (\mathbf{y}_k, A\mathbf{y}_k) + \delta w_2, \qquad |\delta w_2| \leqslant \zeta \|A\| \|\mathbf{y}_k\| \|\mathbf{x}_k\| C_4,$$

where $C_3 \cong C_1 + 1$ and $C_4 \cong 2C_1 + 1$. From this we get

$$\frac{w_1}{w_2} = u_k(1 + \delta u_k), \qquad |\delta u_k| \leqslant \zeta \|A\| \|\mathbf{y}_k\| \|\mathbf{x}_k\| \left(\frac{C_3}{|w_1|} + \frac{C_4}{|w_2|} \right).$$

This suggests the following algorithm for the computation of $\tilde{u}_k$,

$$\tilde{u}_k = \begin{cases} \dfrac{w_1}{w_2} & \text{if} \quad \zeta \|A\| \|\mathbf{y}_k\| \|\mathbf{x}_k\| \left(\dfrac{C_3}{|w_1|} + \dfrac{C_4}{|w_2|} \right) < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Hence, (4.6) is satisfied. Note that $\tilde{u}_k = 0$ means that $\mathbf{x}_{k+1} = \mathbf{z}_k = \mathrm{fl}(\mathbf{x}_k - c_k\mathbf{r}_k)$ is obtained by one step of the steepest-descent algorithm. This can be interpreted as the initialization of the cg algorithm from the vector $\mathbf{x}_k$.

It may also be observed that vectors $\mathbf{z}_k$ and $\mathbf{y}_k$ need not be stored. One step of the algorithm can be performed having five vectors $\mathbf{x}_k$, $\mathbf{x}_{k-1}$, $\mathbf{r}_k$, $\mathbf{r}_{k-1}$ and $\mathbf{v}_k = A\mathbf{r}_k$ in storage and using two matrix-vector multiplications.

We have performed many numerical tests using this algorithm. In most cases the algorithm was well behaved in the spectral norm. However, in a few cases (about 5 percent) numerical tests experimentally confirmed the sharpness of the error bounds in Corollary 4.1.

We end this section by a remark on the cg algorithms for $B = I$ and $B = A^2$. Based on the results of Sec. 3 and assuming that the computed coefficient $\tilde{u}_k = u_k(1 + \delta u_k)$, where $|\delta u_k| \leqslant 1$ and $u_k = (\mathbf{y}_k, B(\mathbf{z}_k - \boldsymbol{\alpha}))/(\mathbf{y}_k, B\mathbf{y}_k)$ for the computed vectors $\mathbf{y}_k$ and $\mathbf{z}_k$, it is possible to prove that there exists an

index k such that the computed $\mathbf{x}_k$ satisfies

$$\|B^{1/2}(\mathbf{x}_k-\boldsymbol{\alpha})\| \leqslant \zeta\kappa\|B^{1/2}\|\,\|\mathbf{x}_k\|C,$$

$$\|\mathbf{x}_k-\boldsymbol{\alpha}\| \leqslant \zeta\kappa\|B^{1/2}\|\,\|B^{-1/2}\|\,\|\mathbf{x}_k\|C,$$

$$\|A\mathbf{x}_k-\mathbf{b}\| \leqslant \zeta\kappa\|A\|\,\|\mathbf{x}_k\|C$$

for a certain constant $C=C(n)$. Note that for $B=I$ we conclude the numerical stability of the minimum-error algorithm. A detailed analysis for the minimal-residual method, $B=A^2$, may be found in [8].

5. FINAL COMMENTS

We have shown that the relative error of the computed vector $\mathbf{x}_k$ by a cg algorithm from Φ depends at worst on $\zeta\kappa^{3/2}$. Since for many practical cases the required accuracy is larger than $\zeta\kappa^{3/2}$, this is a quite satisfactory result.

As we mentioned before, we have not succeeded in analyzing classical cg algorithms. However, we believe that at least some of them have similar numerical properties.

We want to pose another problem of practical interest connected with the numerical properties of cg algorithms. We know that in theory the sequence $\{\mathbf{x}_k\}$ approximates the solution $\boldsymbol{\alpha}$ with the best possible speed of convergence in the class (2.5). Is this still true in the presence of rounding errors? It is important to know the speed of convergence of the computed sequence $\{\mathbf{x}_k\}$ and to see how much of the theoretical optimality continues to hold in fl. We observe experimentally that the computed sequence initially approximates $\boldsymbol{\alpha}$ at least as fast as the Chebyshev iteration, i.e., $\|\mathbf{x}_k-\boldsymbol{\alpha}\|_A \leqslant 2[(\sqrt{\kappa}-1)/(\sqrt{\kappa}+1)]^k\|\mathbf{x}_0-\boldsymbol{\alpha}\|_A$. Furthermore, in many cases the error $\|\mathbf{x}_k-\boldsymbol{\alpha}\|_A$ is significantly less than the above bound. Therefore we propose the following conjecture.

CONJECTURE 5.1. There exists a cg algorithm which computes the sequence $\{\mathbf{x}_k\}$ such that

$$\|\mathbf{x}_k-\boldsymbol{\alpha}\|_A \leqslant \max\left\{\zeta\kappa\|\boldsymbol{\alpha}\|_A C, 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k\|\mathbf{x}_0-\boldsymbol{\alpha}\|_A\right\} \tag{5.1}$$

for all k where $C=C(n)$.

Thus, (5.1) means that as long as $\|\mathbf{x}_k-\boldsymbol{\alpha}\|_A$ is greater than $\zeta\kappa\|\boldsymbol{\alpha}\|_A C$, we have at least Chebyshev speed of convergence. For large k, (5.1) means numerical stability in the A-norm.

APPENDIX

We describe numerical tests of Concus, Golub and O'Leary's [1] conjugate-gradient algorithm defined as follows:

$$\mathbf{x}_0 = \text{a given approximation},$$

$$\mathbf{r}_k = A\mathbf{x}_k - \mathbf{b},$$

$$c_k = \frac{(\mathbf{r}_k,\mathbf{r}_k)}{(\mathbf{r}_k,A\mathbf{r}_k)},$$

$$w_{k+1} = \left(1-\frac{(\mathbf{r}_k,\mathbf{r}_k)}{(\mathbf{r}_{k-1},\mathbf{r}_{k-1})}\frac{c_k}{c_{k-1}}\frac{1}{w_k}\right)^{-1}, \qquad w_1 = 1,$$

$$\mathbf{x}_{k+1} = \mathbf{x}_{k-1} + w_{k+1}(c_k\mathbf{r}_k + \mathbf{x}_k - \mathbf{x}_{k-1}), \qquad \mathbf{x}_{-1} = \mathbf{0},$$

for $k=0,1,\cdots$

We tested this algorithm for the matrix $A=(I-2\mathbf{w}\mathbf{w}^T)D(I-2\mathbf{w}\mathbf{w}^T)$, where $\mathbf{w}$ ($\|\mathbf{w}\|=1$) was a vector produced by a subroutine which generates random numbers and $D=\operatorname{diag}(\lambda_1,\lambda_2,\ldots,\lambda_n)$ was a diagonal matrix with $\lambda_i>0$. We chose $n\in[50,200]$ for different distributions of λ_i varying the condition number from 10^2 to 10^8. We defined λ_i as

(i) $\lambda_i = a+(1-a)(i-1)/(n-1)$ for a positive small a,
(ii) $\lambda_1 = q$, $\lambda_i = a+(1-a)(i-2)/(n-2)$, $i=2,\ldots,n$, where $0<q\ll a\leqslant 1$,
(iii) $\lambda_i = q^{n-i}$ for some $q\leqslant 1$.

We computed j iterative steps until the limiting accuracy was achieved. For ill-conditioned problems with $n=50$, j was several thousand. As we mentioned in the Introduction, the best possible computed approximation had the relative error of order $\zeta\kappa^{3/2}$ (where ζ equals 10^{-14} for the CDC 3600 computer used in the experiments) and the residual vector had norm of order $\zeta\kappa\|\boldsymbol{\alpha}\|$.

We also tested the algorithm φ described in Example 4.1 for the above examples. In most cases the algorithm φ produced residual vectors of order $\zeta\|\boldsymbol{\alpha}\|$. Thus it behaved better than the Concus-Golub-O'Leary algorithm. For a few cases with eigenvalues distributed as in (iii) and with initial error

$\mathbf{x}_0 - \boldsymbol{\alpha} = \sum_{j=1}^{n} c_j \boldsymbol{\xi}_j$ such that $|c_1| \gg |c_2| \gg \cdots \gg |c_n|$, the algorithm φ produced residual vectors of order $\zeta\kappa\|\alpha\|$.

I wish to express my gratitude to J. F. Traub and Å. Björk for many valuable comments on the manuscript. I also thank my colleagues M. Jankowski, Z. Kacewicz, A. Kielbasinski, A. Smoktunowicz and C. Wasilkowski from the University of Warsaw for stimulating discussions. My special thanks are to R. Wieladek, Institute of Geophysics of the Polish Academy of Sciences, for pointing out a property of a cg algorithm which allowed me to analyze the cg iteration.

REFERENCES

1 P. Concus, G. H. Golub and D. P. O'Leary, A generalized conjugate gradient method for the numerical solution of elliptic partial differential equations, in *Sparse Matrix Computations* (J. R. Bunch and D. J. Rose, Eds.), Academic, New York, 1976, pp. 309–332.

2 M. Engeli, Th. Ginsburg, H. Rutishauser and E. Stiefel, *Refined Iterative Methods for Computation of the Solution and the Eigenvalues of Self-Adjoint Boundary Value Problems*, Birkhäuser, Stuttgart, 1959.

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

4 M. Janowski and H. Woźniakowski, Iterative refinement implies numerical stability, *Nordisk Tidskr. Informationsbehandling (BIT)* 17:303–311 (1977).

5 A. Kielbasiński, An addition algorithm with corrections and some of its applications (in Polish), *Mat. Stosowana* 1:23–41 (1973).

6 E. Stiefel, Kernel polynomials in linear algebra and their numerical applications, *NBS Appl. Math.* 49:1–22 (1958).

7 R. Wieladek, private communication.

8 R. Wieladek, Round-off error analysis of the minimal residual method, in *Publications of the Institute of Geophysics*, Polish Academy of Sciences, PWN, Warsaw, 1978, to appear.

9 J. M. Wilkinson, *Rounding Errors in Algebraic Processes*, Prentice-Hall, Englewood Cliffs, N.J., 1963.

10 J. M. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon, Oxford, 1965.

11 H. Woźniakowski, Numerical stability of the Chebyshev method for the solution of large linear systems, *Numer. Math.* 28:191–209 (1977).

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

Received 7 May 1979