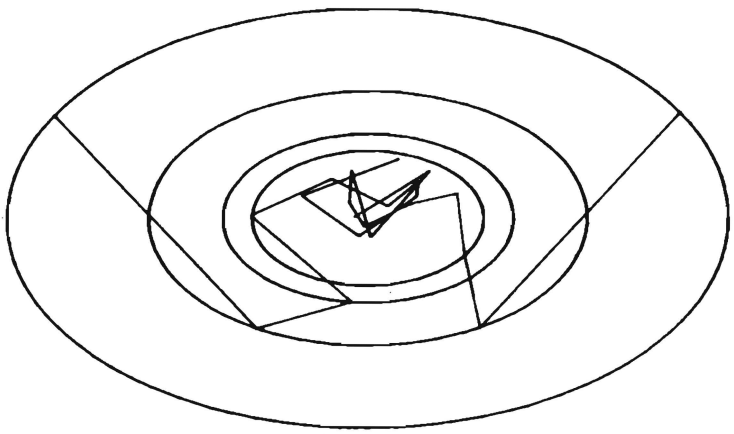
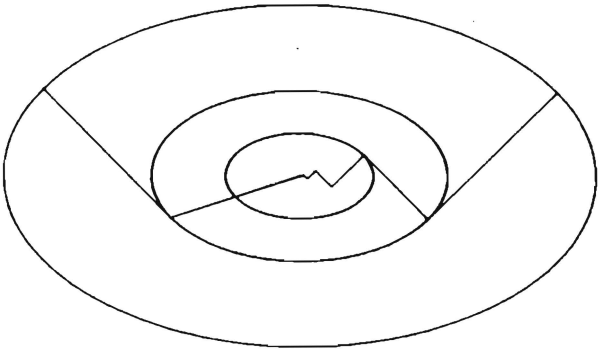
Round-off error analysis of descent methods for solving linear equations

ROUND-OFF ERROR ANALYSIS OF DESCENT METHODS FOR SOLVING LINEAR EQUATIONS



JO BOLLEN



The diagram on the front page is drawn from a test for a two dimensional linear system with eigenvalues 1/3 and 1, where the eigenvector components of the solution $\hat{x}$ are equal and where the (absolute values of the) eigenvector components of the initial error vector $\hat{x} - x_0$ are in the ratio of 3 to 1. Furtermore, $\|\hat{x}\| = 1$ and $\|\hat{x} - x_0\| = {}_{10}{-}1$. The test is performed using artificial floating point arithmetic with artificial relative precision ${}_{10}{-}2$ (see section 1.6). The left-hand part represents the iterands computed by the conjugate gradient method whereas the right-hand part represents the iterands computed by the gradient method. The ellipses correspond to the level lines of the objective function $F(x) = \frac{1}{2}\|A^{\frac{1}{2}}(\hat{x} - x)\|^2$.

The diagram above is drawn from a test with the same parameter setting but the computations are performed with (almost) exact accuracy (artificial relative precision ${}_{10}{-}10$). It shows the step-wise-linear convergence of the gradient method and the termination after two steps of the conjugate gradient method. The front page diagram illustrates the influence of round-off on the numerical behavior of both methods.

It is my colleague Herman Willemsen who not only performed the tests described above but also did all the programming and testing needed to obtain the numerical results presented in this thesis. I like to express my gratitude for the patience he showed in working with slowly convergent processes.

ROUND-OFF ERROR ANALYSIS OF DESCENT METHODS FOR SOLVING LINEAR EQUATIONS

ROUND-OFF ERROR ANALYSIS OF DESCENT METHODS FOR SOLVING LINEAR EQUATIONS

PROEFSCHRIFT

TER VERKRIJGING VAN DE GRAAD VAN DOCTOR IN DE TECHNISCHE WETENSCHAPPEN AAN DE TECHNISCHE HOGESCHOOL EINDHOVEN, OP GEZAG VAN DE RECTOR MAGNIFICUS, PROF. IR. J. ERKELENS, VOOR EEN COMMISSIE AANGEWEZEN DOOR HET COLLEGE VAN DEKANEN IN HET OPENBAAR TE VERDEDIGEN OP DINSDAG 2 DECEMBER 1980 TE 16.00 UUR

DOOR

JOSEPH ANTONIUS MARIA BOLLEN

GEBOREN TE GELEEN

DRUK WIBRO HELMOND

Dit proefschrift is goedgekeurd
door de promotoren

Prof.dr. G.W. Veltkamp
en
Prof.dr.ir. M.L.J. Hautus

Aan Irma,
aan mijn vader en
ter nagedachtenis aan mijn moeder

CONTENTS

CHAPTER 1

INTRODUCTION

1.1. Introduction and summary

There are two classes of numerical methods for solving linear systems $Ax = b$, viz. direct methods and iterative methods. Direct methods decompose the original matrix A in order to obtain an equivalent linear system that is easy to solve numerically. Some commonly used methods are Gaussian elimination, QR-decomposition by Householder's method, modified Gram-Schmidt and Cholesky decomposition. Iterative methods compute successive approximations of the solution, without making any changes to the original matrix. Some commonly used iterative methods are Jacobi-, Richardson-, Gauss-Seidel-, Chebyshev- and Lanczos-iterations, systematic overrelaxation, alternating direction iterations, gradient method and conjugate gradient method.

A basic distinction between direct and iterative methods is that a direct method yields the solution $\hat{x} := A^{-1}b$ exactly in a finite number of arithmetical operations (if the latter are performed without round-off), whereas an iterative method in general produces an infinite sequence $\{x_i\}$ whose limit is the solution $\hat{x}$. Each approximation is obtained from its predecessor(s) by a finite number of arithmetical operations. As a consequence, for a direct method the number of arithmetical operations is known in advance, whereas for an iterative method the number of arithmetical operations depends on the required accuracy of the computed solution.

Another distinction between direct and iterative methods is the difference in storage requirements. In most direct methods, where all entries of the matrix A are stored in a two dimensional array, the matrices resulting from the decomposition of A can be stored by overwriting (parts of) A. Iterative methods can be applied without storing A explicitly. One only needs a black box for the execution of matrix by vector product operations. For fairly small systems this distinction

is minor. However, for large sparse systems where the matrix A has a relatively small number of nonzero entries, the decomposition matrices generated by direct methods generally are less sparse than A and this may give rise to rather excessive storage requirements. On the other hand, an iterative method can take full advantage of the sparsity of A, because in the matrix by vector product operations the zero-entries can be skipped.

An additional aspect of importance in the discussion of the solution of linear systems is the influence on the computed solution of round-off, this round-off being due to the floating point computations with finite relative precision ε. As far as direct methods are concerned Wilkinson [65] proved that most commonly used direct methods are well-behaved and numerically stable. Well-behaved methods compute an approximation x which is the exact solution of a linear system with a slightly perturbed A, i.e., $(A+E)x = b$, where E is of order $\varepsilon\|A\|$. Consequently, a well-behaved methods computes an approximation x whose relative error $\|\hat{x} - x\| / \|x\|$ does not exceed a quantity of order $\varepsilon\|A\|\|A^{-1}\| =: \varepsilon\kappa$. A method that computes an approximation with a relative error at most of order $\varepsilon\kappa$ is called a numerically stable method. Hence, a well-behaved method is a numerically stable method but not necessarily vice versa. As far as iterative methods are concerned there is up to now only very little literature presenting results on the influence of round-off. This partly is due to the fact that iterative methods seemed to be more self-correcting than direct methods, so that one expected iterative methods to be well-behaved spontaneously. Another reason is that the users of iterative methods generally are more interested in how many iterations are needed to obtain an approximation with a reasonable accuracy than in the maximally attainable accuracy after maybe many iterations. Nevertheless it is somewhat remarkable that round-off error analyses of iterative methods hardly exist. Wózniakowski is one of the first authors who in very recent years published results on good-behavior and numerical stability for iterative methods such as Chebyshev iterations (cf. Wózniakowski [77]) and SOR, Jacobi-, Gauss-Seidel- and Richardson-iterations (cf. Wózniakowski [78]; see also Jankowski and Wózniakowski [77] and Wózniakowski [80]).

This thesis is intended to be a contribution to this rather new field of research, concerning the round-off error analysis of iterative

methods for solving linear systems. We will study the class of descent methods which is a large sub-class of iterative methods. Descent methods can be characterized as follows. Given an objective function $F(x)$, one starts at an initial point, determines, according to a fixed rule, a direction of movement and then moves in that direction to the local minimum of the objective function. At the new point a new direction is determined and the process is repeated. The objective function F must satisfy the following three important properties: $F(\hat{x}) = 0$, $F(x) > 0$ if $x \neq \hat{x}$, and F is convex. We consider descent methods where $F(x)$ is taken to be the quadratic function $(\hat{x} - x, A(\hat{x} - x))$, expressed in terms of the Euclidean inner product. A is supposed to be a positive definite matrix. In addition to the choice of the objective function, the main difference between the various descent methods rests with the rule by which successive directions are chosen.

We pay special attention to the gradient method and the conjugate gradient method. The gradient method (often referred to as steepest descent method) is a descent method that is especially important from a theoretical point of view, since it is one of the simplest methods for which a satisfactory analysis of the convergence behavior exists (in the case of exact computations). The method is characterized by the rule that at each iterand x_i the residual vector $b - Ax_i$ is chosen as the direction of movement. The conjugate gradient method, developed independently by Hestenes and Stiefel [52], is an iterative method as well as a direct method. It is an iterative (descent) method in the sense that at each step a better approximation to the solution is obtained. At each iterand x_i an A-orthogonal version of the residual vector $b - Ax_i$ is chosen as the direction of movement. It is a direct method in the sense that it yields the solution after at most n steps, where n is the dimension of the linear system (in the case of exact computations). The early enthusiasm on this finite termination property soon diminished after it turned out that in the presence of round-off for some ill-conditioned linear systems the n-th computed iterand x_i is not even a reasonable approximation to the solution. The method became of mainly academic interest, at least as a solver for linear equations. The paper of Reid [71], in which the iterative character of the method was emphasized, reactivated the interest in the conjugate gradient method and nowadays the method is known as an iterative method with very strong convergence properties for large

sparse linear systems with moderate condition number. For these systems the method often yields acceptable approximations after much less than n steps.
This thesis contains a number of results on the good-behavior and numerical stability of the gradient method and the conjugate gradient method for both well- and ill-conditioned systems. These results mainly deal with the ultimate numerical convergence behavior. We have carried out a number of computational tests in order to verify the analytical results of our round-off error analysis.

We now summarize the contents of this thesis.

In this introductory Chapter 1 some basic notions required in the sequel are presented. After the introduction of some notational conventions in section 2 we discuss in section 3 the preliminaries on rounding errors in floating point computations. Section 4 deals with the concepts of good-behavior and (A-) numerical stability which serve as a qualification for the attainable accuracy of computed solutions in the presence of round-off. We also briefly recall some definitions concerning the speed of convergence of iterative processes. The reason why and the way in which we use the Bachmann-Landau O-symbol in our round-off error analysis is explained in section 5. The final section of Chapter 1 describes the construction of our test problems and the implementation of the gradient method and the conjugate gradient method computations. We also give a description of what we call artificial floating point arithmetic.
In Chapter 2 a general theory for the algebraic and numerical behavior of descent methods (DM's) is presented. In section 1 we discuss the fundamental idea behind DM's and we point out that the methods can be based either on recursive or on true residual vectors. In section 2 the definitions of recursive residual descent methods (RRDM's) and true residual descent methods (TRDM's) are given and we deduce some well-known algebraic properties that are fundamental for studying the properties of DM's in the presence of round-off (henceforth denoted by numerical properties). We also briefly review some well-known DM's like the Gauss-Seidel method, the Gauss-Southwell method, the gradient method and the conjugate gradient method. Numerical properties of RRDM's are derived in section 3. The numerical behavior of the recursive residuals is treated in subsection 3.1 and then the numerical

behavior of the approximations x_i is treated in subsection 3.2. Subsection 3.1 contains the main theorem (theorem 2.3.1.4) for the numerical performance of RRDM's. In section 4 we derive numerical properties for TRDM's and the main result is stated in theorem 2.4.6. The usability of this main result is demonstrated by applying it to the Gauss-Southwell method.

In Chapter 3 the general theory of chapter 2 is applied to the gradient method (GM). The definition of the GM is given in section 1, where we also review some of its well-known algebraic properties. In section 2 numerical analogues of these algebraic properties are derived for the RRGM, whereas in section 3 this is done for the TRGM. The main result for the RRGM is the step-wise linear convergence to zero (cf. section 1.4) of the recursive residual vectors. The proof of good-behavior and numerical stability is the main result for the TRGM. Section 4 reports on numerical results obtained by tests with the RRGM and the TRGM.

Chapter 4 is devoted to the conjugate gradient method (CGM). In section 1 we give the definition of what we call the most natural version of the CGM, which is one of the (algebraically equivalent) versions contained in the paper of Hestenes and Stiefel [52] (cf. also Reid [71]). We also deduce some of its numerous elegant algebraic properties. In section 2 numerical analogues of some of these algebraic properties are derived for the RRCGM, whereas in section 3 this is done for the TRCGM. The main result for the RRCGM is the bi-step-wise linear convergence to zero (see section 4.2) of the recursive residuals. For the TRCGM our main result is that this method computes at least one approximation x_i for which the residual is at most of the order $\varepsilon\kappa^{\frac{1}{2}}\|A\|\|x_1\|$, which is a factor $\kappa^{\frac{1}{2}}$ worse than good-behavior. We also point out how it can be understood that in many actual executions of the process good-behavior is observed. In section 4 we report on numerical results obtained by tests with the RRCGM and the TRCGM.

In Chapter 5 we discuss some variants of the CGM as defined in Chapter 4. In section 1 we introduce four so-called independent start conjugate gradient methods (ISCGM's); their algebraic properties are derived in section 2. In section 3 we demonstrate the numerical implications of these properties for one particular version. The main result is that the natural version of the conjugate gradient method as

considered in Chapter 4 seems to be more robust than the other versions as far as the influence of round-off errors is concerned.

1.2. *Notations and conventions*

In this section we describe our notational conventions and we give a list of the general symbols, which we shall use throughout this monograph.

Vectors

All vectors are supposed to be (real) column vectors. The vector x_i indicates the i-th approximation to the solution $\hat{x}$ of the linear system, determined by the descent method on hand. The vector p_i indicates the i-th direction vector of the descent method (see section 1.2). The residual vector $b - Ax_i$ is denoted by r_i of $\hat{r}_i$ (for the difference see below). By (x,y) we mean the *Euclidean inner product* of the vectors x and y and by $\|x\|$ we mean the *Euclidean norm* of the vector x. Thus

$$(1) \qquad (x,y) := \sum_{j=1}^{n} x_j y_j , \quad \|x\| := (x,x)^{\frac{1}{2}} .$$

The indices j or ℓ in connection with a vector indicate the j-th or ℓ-th component of the vector.

Matrices

The descent methods under consideration are basically designed for solving a system of linear equations, denoted by $Ax = b$, with a (real) positive definite matrix A. We briefly call such a system a *definite system*. The order of the (square) matrix A is called the *dimension* of the linear system and it is denoted by n. The spectral decomposition of the $m \times n$ matrix A is denoted by

$$(2) \qquad A = U\Lambda U^T ,$$

where

U is an orthogonal $n \times n$ matrix, whose columns u_i are a complete set of *orthonormal eigenvectors* of A,

Λ is an $n \times n$ diagonal matrix, whose diagonal entries λ_i are the n (positive) *eigenvalues* of A. Without loss of generality we always assume that the eigenvalues are ordered according to $0 < \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$.

The positive definite matrices A^α ($\alpha = -1/2, 1/2, 3/2$) are defined by

$$A^\alpha = U\Lambda^\alpha U^T . \tag{3}$$

Hence $A^{\frac{1}{2}}A^{\frac{1}{2}} = A$, $A^{-\frac{1}{2}}A^{\frac{1}{2}} = I$, $A^{-\frac{1}{2}}A^{\frac{3}{2}} = A$, etc.
The norm of a matrix A, denoted by $\|A\|$, is always meant to be the spectral norm, defined by

$$\|A\| := \max_{x \neq 0} \frac{\|Ax\|}{\|x\|} = \lambda_n . \tag{4}$$

The rate of change of the solution of a linear system with respect to a change in the coefficients as well as the influence of round-off on the computed solution and on the rate of convergence is expressed in terms of the (spectral) *condition number* of A, denoted by $\kappa(A)$ and defined by

$$\kappa(A) := \|A\|\|A^{-1}\| = \lambda_n/\lambda_1 . \tag{5}$$

Since we only consider the condition number of the matrix A corresponding to a specific definite linear system, there is no confusion if we write simply κ instead of $\kappa(A)$.
An important inequality in connection with the condition number which will frequently be used in our convergence considerations is the *Kantorovich inequality*, which states that if A is a positive definite matrix, then for any vector x one has

$$\frac{(x,x)}{(x,Ax)\,(x,A^{-1}x)} \geq \frac{4\kappa}{(\kappa+1)^2} . \tag{6}$$

In (6), equality holds iff x is a multiple of the vector $v_1 + v_n$, where v_1 is an eigenvector corresponding to the smallest eigenvalue and v_n is an eigenvector corresponding to the largest eigenvalue. The quotient at the left-hand side of (6) will be called the *Kantorovich quotient* of the vector x with respect to A. It can be written in terms of norms as $\|x\|^4 / (\|A^{\frac{1}{2}}x\|\|A^{-\frac{1}{2}}x\|)^2$.

Round-off

Our round-off error analysis is based on round-off due to the use of floating point numbers and floating point arithmetic. Vectors and numbers that actually have been computed and stored by the floating point machine are called *machine vectors* and *machine numbers*. If an expression S involving machine vectors and machine numbers is evaluated using normal floating point arithmetic, then this is denoted by fl(S); if the expressions is computed using *artificial floating point arithmetic* (defined in section 1.6), then this is denoted by fla(S). Round-off occurring at basic arithmetical operations is expressed in terms of *round-off matrices*, denoted by the capital characters F, G, D and E, each referring to specific operations like, e.g., vector addition, scalar by vector products, etc., as described in section 1.3. If we want to indicate the difference between an exact vector or number and a computed vector or number, then we use the symbol δ. For instance, we write

$$(8) \qquad z := fl(x+y) = x+y+\delta z \ , \quad k := fl(\ell * m) = \ell * m + \delta k \ ,$$

where x, y, z and δz are machine vectors and ℓ, m, k and δk are machine numbers. The vector δz is called a *round-off vector* and the scalar δk is called a *round-off scalar*. Intermediate round-off scalars are denoted by Greek letters like ζ, η, μ, ν, ξ, ρ, σ, τ, ω (see e.g. formula (2.3.1.16)).

For the vectors x_i and p_i, mentioned before, we do not have a different notation to indicate whether they stand for the computed vectors or the exact vectors. With respect to the *residual vector* $b - Ax_i$ we make the following distinction. In considerations on numerical behavior the vector $\hat{r}_i$ stands for the (exact) vector $b - Ax_i$, whereas r_i stands for some computed vector that would be equal to $\hat{r}_i$ when using exact arithmetic.

A property holding if an algorithm is performed using exact arithmetic is called an *algebraic property*; a property holding if the descent method is performed using floating point arithmetic is called a *numerical property*. The term *analytical results* refers to algebraic as well as to numerical properties and is used in contrast with the term *numerical results*, pointing to numerical experiments.

In any (sub)section theorems, propositions, lemmas, definitions and remarks are numbered 1, 2, ..., and formulas are numbered (1), (2), ... If, in some (sub)section, we refer to theorem 2 (say), then we mean theorem 2 of the (sub)section on hand. If we refer to theorem 1.2.3 (say), then we mean theorem 3 of section 1.2. Opposite to each number for pagination there is a number indicating the (sub)section in question.

We conclude this section with a list of general symbols.

$\mathbb{R}$	set of real numbers
$\mathbb{R}^n$	set of column n-vectors over $\mathbb{R}$
$\mathbb{N}$	set $\{1,2,\ldots\}$ of natural numbers
$\mathbb{N}_0$	set $\{0,1,\ldots\}$ of nonnegative integers
A^T	transpose of the matrix A
A^{-1}	inverse of the matrix A
A_{ij}	(i,j)-th entry of the matrix A
I	identity matrix
$\|A\|$	matrix with entries $\|A\|_{ij} := \|A_{ij}\|$
$A < B$	for all entries there holds $A_{ij} < B_{ij}$
$\mathrm{diag}(a_1,\ldots,a_n)$	diagonal matrix with diagonal entries $a_1,\ldots,a_n$
x^T	transpose of vector x
$\hat{x}$	solution $A^{-1}b$ of the linear system $Ax = b$
$\{x_i\}$	sequence of vectors $x_1, x_2, \ldots$
$\overline{\lim}\, x_i$	limes superior of the sequence $\{x_i\}$
span $\{x_1,\ldots,x_i\}$	subspace spanned by the vectors $x_1,\ldots,x_i$
ψ_i	$\|A^{\frac{1}{2}}\|\|x_i\| \,/\, \|A^{\frac{1}{2}}(\hat{x} - x_i)\|$
φ_i	$\|A\|\|x_i\| \,/\, \|A(\hat{x} - x_i)\|$
χ_i	$\|A^{3/2}\|\|x_i\| \,/\, \|A^{3/2}(\hat{x} - x_i)\|$
$a \sim b$	a approximately equals b, a is of the order of b
$\|a\|$	absolute value of a

$a \gg b$	a is much greater than b
a mod b	the remainder when dividing a by b
ent(a)	entier of a; largest integer not exceeding a
$a_i \to 0 \ (i \to \infty)$	$\lim_{i\to\infty} a_i = 0$
$\forall$	universal quantifier
$\exists$	existential quantifier
$\square$	end of a (proof of a) theorem, lemma, proposition or remark
$\Rightarrow$	implication sign
*	product sign (only used occasionally to avoid confusion)
$\oplus$	any basic dyadic arithmetical operation +, -, *, /
$\nabla F(x)$	gradient of the vector function F
O	Bachmann-Landau symbol
B	base of the floating point numbers
t	length of the mantissa of the floating point numbers
ε	relative machine precision; $\varepsilon := \frac{1}{2}B^{1-t}$
C_1	constant depending on n and ε, denoting the upper bound for the norm of the round-off matrix E, representing round-off at matrix by vector product computations
C_2	constant depending on n and ε, denoting the upper bound for the norm of the round-off matrix D, representing round-off at inner product computations

1.3. *Preliminaries on rounding errors and floating point arithmetic*

Throughout this thesis we assume that the algorithms based on descent methods are performed in floating point arithmetic. The floating point numbers will be assumed to have *base* B and a *mantissa length* of t digits ($B \geq 2$, $t \geq 1$). Then every real number in the floating point range of the machine can be represented with a relative error which does not exceed *the relative machine precision* ε, which is defined by $\varepsilon = \frac{1}{2}B^{1-t}$. Furthermore we assume that we have a machine with *proper rounding arithmetic* in the sense of Dekker [79]. This means that the execution of any dyadic arithmetical operation $\oplus$ (this can be +, −, *, /) on two machine numbers a and b gives a machine number $fl(a \oplus b)$ such that there is no other machine number closer to the exact result of $a \oplus b$. Consequently, the following relations hold

$$(1) \qquad fl(a \oplus b) = (a \oplus b)(1+\xi) ,$$

$$(2) \qquad (1+\eta)fl(a \oplus b) = a \oplus b ,$$

where both $|\xi| \leq \varepsilon$, $|\eta| \leq \varepsilon$.

We do not put a restriction on the range of the exponent of the machine numbers. Hence we neglect the possibility of underflow or overflow.

From (1) and (2) it follows that adding or subtracting two machine vectors x and y and multiplying a machine vector x by a machine number a (implemented in the obvious way) gives computed vectors $fl(x \pm y)$ and $fl(ax)$ satisfying

$$(3) \qquad fl(x \pm y) = (I + F_1)(x \pm y) ,$$

$$(4) \qquad (I + G_1)fl(x \pm y) = x \pm y ,$$

$$(5) \qquad fl(ax) = (I + F_2)ax ,$$

$$(6) \qquad (I + G_2)fl(ax) = ax ,$$

where F_1, F_2, G_1 and G_2 are diagonal matrices, satisfying

$$(7) \qquad |F_1| \leq \varepsilon I , \quad |F_2| \leq \varepsilon I , \quad |G_1| \leq \varepsilon I , \quad |G_2| \leq \varepsilon I ,$$

and consequently

$$(8) \qquad \|F_1\| \leq \varepsilon , \quad \|F_2\| \leq \varepsilon , \quad \|G_1\| \leq \varepsilon , \quad \|G_2\| \leq \varepsilon .$$

We assume that the algorithm for the calculation of the inner product of two machine vectors x and y satisfies

(9) $\quad fl((x,y)) = ((I+D)x,y)$,

where D is a diagonal matrix such that

(10) $\quad \|D\| \leq \varepsilon C_2$;

the constant C_2 depending only on n and ε. Throughout this monograph C_2 always stands for the bound of the round-off matrix D representing round-off errors at inner product computations.

REMARK 1. If the inner product calculation is performed in the obvious way, by multiplying corresponding successive components (in increasing order) and adding the result to the intermediate inner product, then the entries of the diagonal matrix D in (9) satisfy, under the restriction $n\varepsilon \to 0$ (cf. Wilkinson [65])

(11) $\quad |D_{11}| \leq n\varepsilon(1+o(1))$, $\quad |D_{ii}| \leq (n+2-i)\varepsilon(1+o(1))$,

$(i = 2,\ldots,n)$.

Consequently, the constant C_2 in (10) can be chosen $C_2 = n(1+o(1))$, $[n\varepsilon \to 0]$. (for the meaning of the o-symbol we refer to section 1.5.) □

We assume that the algorithm for matrix by vector product calculation is implemented in such a way that the computed vector fl(Ax), based on the machine matrix A and the machine vector x, satisfies

(12) $\quad fl(Ax) = (A+E)x$,

where E is a matrix such that

(13) $\quad \|E\| \leq \varepsilon C_1 \|A\|$;

the constant C_1 depending only on n and ε. Throughout this monograph C_1 always stands for the bound of the round-off matrix E representing round-off errors at matrix by vector product computations.

REMARK 2. In the real-world siutation, where the matrix by vector product calculation is performed in the obvious way, by computing inner products (in the way as described in remark 1) of rows of A and the vector x, it follows from (11) that fl(Ax) satisfies componentwise

$$(14) \qquad (\mathrm{fl}(Ax))_j = \sum_{\ell=1}^{n} A_{j\ell}\, x_\ell (1+\eta_{j\ell}) , \quad (j = 1,\ldots,n)$$

where, under the restriction $n\varepsilon \to 0$, for all $j = 1,\ldots,n$

$$(15) \qquad |\eta_{j1}| \le n\varepsilon(1+o(1)) , \quad |\eta_{j\ell}| \le (n+2-\ell)\varepsilon(1+o(1)) , \quad (\ell = 2,\ldots,n) .$$

Hence the matrix E in (12) has entries $E_{j\ell} = \eta_{j\ell} A_{j\ell}$ and consequently

$$(16) \qquad |E| \le n\varepsilon|A|(1+o(1)) , \quad \|E\| \le n^{3/2}\varepsilon\|A\|(1+o(1)) , \quad [n\varepsilon \to 0] .$$

According to the definition of C_1 we can choose $C_1 = n^{3/2}(1+o(1))$, $[n\varepsilon \to 0]$. Note that componentwise $(Ex)_j = \Sigma_{\ell=1}^{n} A_{j\ell}\, x_\ell\, \eta_{j\ell}$ and hence the round-off vector Ex generally is *randomly directed*. This is an important characteristic of the normal matrix by vector product computation, since for randomly directed vectors y, $\|Ay\| \sim \|A\|\|y\|$. Furthermore,

$$|(\mathrm{fl}(Ax))_j - (Ax)_j| \le n\varepsilon(|A||x|)_j$$

which indicates that the components are not necessarily computed with relative precision. □

If two vectors are added (or subtracted), then the rounding errors due to this operation can be expressed by (3) and (4). Another, rather unusual, way to express this rounding errors is given in the following lemma. It will be of special interest if the two vectors differ much in length. We shall meet this situation in Chapter 2.
From the assumption that we have proper rounding arithmetic it follows that if we add two machine numbers a and b for which $|b| < (\varepsilon/B)|a|$, then

$$(17) \qquad \mathrm{fl}(a+b) = a .$$

Using this relation we can prove the following lemma.

LEMMA 3. *If* x *and* y *are machine vectors, then*

$$(18) \qquad \mathrm{fl}(x+y) = x + (I+H)y ,$$

where H *is a diagonal matrix satisfying*

$$(19) \qquad |H| \le (B+\varepsilon)I , \quad \|H\| \le B+\varepsilon .$$

PROOF. Let $fl(x+y) = x+y+\delta$.
If $|y_j| < (\varepsilon/B)|x_j|$, then it follows from (17) that $\delta_j = -y_j$.
If $|y_j| \geq (\varepsilon/B)|x_j|$, then it follows with (1) that $|\delta_j| \leq \varepsilon|(x+y)_j| \leq$
$\leq (B+\varepsilon)|y_j|$. Hence in both cases $|\delta_j| \leq (B+\varepsilon)|y_j|$.
The proof of the lemma is completed by defining $H_{jj} := \delta_j/y_j$, $(y_j \neq 0)$,
$H_{jj} := 0$, $(y_j = 0)$, $H_{ji} := 0$, $(j \neq i)$. □

REMARK 4. Suppose we have a sequence $\{y_i\}$ of machine vectors that converges linearly on the average to zero with a convergence ratio no greater than L, i.e., $\|y_i\| \leq L^i\|y_0\|$, $L \in (0,1)$. Assume that $s := \Sigma_{\ell=0}^{\infty} y_\ell$ is computed by adding successive vectors (in increasing order of indices) to the intermediate sum vector $s_i := fl(\Sigma_{\ell=0}^{i} y_i)$. Then, in view of the foregoing lemma we obtain

$$s_{i+1} = fl(s_i + y_i) = s_i + y_i + \delta s_{i+1}\,, \quad \delta s_{i+1} = H_i\, y_i\,,$$
$$\|\delta s_{i+1}\| \leq (B+\varepsilon)\|y_i\|\,. \tag{20}$$

Hence $\|s_{i+1} - s_i\| \leq (1+B+\varepsilon)\|y_i\|$ and consequently $\|s_i\|$ is bounded for all $i \geq 0$. □

If the defining statements of a DM contain compound statements, then (unless stated differently) these statements are supposed to be performed in the obvious way, based on the elementary arithmetical operations described so far.

In order to investigate the influence of round-off due to a specific arithmetical operation, we sometimes assume in performing a round-off error analysis that all arithmetical operations are executed exactly, except for the one under consideration. This kind of analysis is called a *one-round-off error analysis*.

1.4. *Basic concepts of numerical stability, good-behavior and convergence rate*

To denote the quality of the approximate solution computed by an algorithm with floating point arithmetic, one generally uses the concepts of numerical stability and good-behavior. The speed of convergence to the solution is expressed in terms of order of convergence and convergence ratio.

We briefly recall (see Wózniakowski [77]) what we mean by numerical stability and good-behavior of an iterative method for solving a linear system $Ax = b$, where A is a nonsingular matrix and b is a (column) vector (solution vector $\hat{x}$). We assume that $\|\cdot\|$ denotes the Euclidean norm for vectors and the spectral norm for matrices (2-norm). Since our linear system is supposed to be definite and we only consider DM's that minimize the objective function $F(x) := (\hat{x} - x, A(\hat{x} - x)) = \|A^{\frac{1}{2}}(\hat{x} - x)\|^2$, it seems sensible to define also a stability concept connected with this function.

Suppose a DM is performed in floating point arithmetic (relative machine precision ε) with arbitrary initial point x_0 and a sequence $\{x_i\}$ is computed of approximations to the solution $\hat{x}$ of the linear system $Ax = b$. Then we have the following definitions (the constants g_1, g_2 and g_3 that appear are supposed to depend only on the dimension of the system).

DEFINITION 1. *The DM is said to be well-behaved (or, equivalently, has good-behavior) if for all initial points x_0 there exists an approximation x_i such that*

$$(1) \qquad (A+E)x_i = b \ ,$$

with some matrix E satisfying $\|E\| \le g_1 \varepsilon \|A\|$. □

In view of formula (1), good-behavior menas that the computed approximate solution is the exact solution of a slightly perturbed system. It is easily seen that (1) implies

$$(2) \qquad \|A(\hat{x} - x_i)\| \le g_1 \varepsilon \|A\| \|x_i\| \ .$$

On the other hand, if (2) is satisfied, then the matrix $E := (b - Ax_i)x_i^T / \|x_i\|^2$ satisfies equality (1) and the inequality of

definition 1. Hence a DM is well-behaved iff there exists an approximation x_i satisfying (2). The vector $A(\hat{x} - x_i) = b - Ax_i$ is called the *residual vector* and $\|A(\hat{x} - x_i)\|$ is called the *residual*. Since

$$\|\hat{x} - x_i\| \leq \|A^{-\frac{1}{2}}\| \|A^{\frac{1}{2}}(\hat{x} - x_i)\| \leq \|A^{-1}\| \|A(\hat{x} - x_i)\| , \tag{3}$$

inequality (2) implies

$$\|\hat{x} - x_i\| \leq g_1 \varepsilon\kappa \|x_i\| , \quad \|A^{\frac{1}{2}}(\hat{x} - x_i)\| \leq g_1 \varepsilon\kappa^{\frac{1}{2}} \|x_i\| . \tag{4}$$

This gives rise to the following definition.

DEFINITION 2. *The DM is said to be numerically stable if for all initial points* x_0 *there exists an approximation* x_i *satisfying*

$$\|\hat{x} - x_i\| \leq g_2 \varepsilon\kappa \|x_i\| . \tag{5}$$

The DM is said to be A*-numerically stable if for all initial points* x_0 *there exists an approximation* x_i *satisfying*

$$\|A^{\frac{1}{2}}(\hat{x} - x_i)\| \leq g_3 \varepsilon\kappa^{\frac{1}{2}} \|A^{\frac{1}{2}}\| \|x_i\| . \tag{6}$$ □

The vector $\hat{x} - x_i$ is called the *error vector*, $\|\hat{x} - x_i\|$ is called the *error*, $A^{\frac{1}{2}}(\hat{x} - x_i)$ is called the *natural error vector* and $\|A^{\frac{1}{2}}(\hat{x} - x_i)\|$ is called the *natural error*.

A numerically stable DM is of interest only if $g_2 \varepsilon\kappa$ is appreciably less than unity and for an A-numerically stable DM one wishes $g_3 \varepsilon\kappa^{\frac{1}{2}}$ to be appreciably less than unity. When using in the following chapters one of the concepts defined above, we shall always indicate the underlying restriction on n, ε, κ.

Note that (5) and (6) imply

$$\|\hat{x} - x_i\| \leq \frac{g_2 \varepsilon\kappa}{1 - g_2 \varepsilon\kappa} \|\hat{x}\| , \quad \|A^{\frac{1}{2}}(\hat{x} - x_i)\| \leq \frac{g_3 \varepsilon\kappa^{\frac{1}{2}}}{1 - g_3 \varepsilon\kappa} \|A^{\frac{1}{2}}\| \|\hat{x}\| , \tag{7}$$

hence we might as well (cf. Wózniakowski [77]) have used $\hat{x}$ instead of x_i in the right-hand sides of (5) and (6).

Good-behavior implies A-numerical stability and A-numerical stability implies numerical stability. On the other hand these implications do not, in general, hold vice versa.

Since

$$\|A(\hat{x} - x_i)\| \leq \|A^{\frac{1}{2}}\| \|A^{\frac{1}{2}}(\hat{x} - x_i)\| \leq \|A\| \|\hat{x} - x_i\| , \tag{8}$$

numerical stability only implies

$$\|A(\hat{x} - x_i)\| \leq g_2 \varepsilon \kappa \|A\| \|x_i\| , \quad \|A^{\frac{1}{2}}(\hat{x} - x_i)\| \leq g_2 \varepsilon \kappa \|A^{\frac{1}{2}}\| \|x_i\| , \tag{9}$$

(cf. formulas (2) and (6)) and A-numerical stability only implies

$$\|A(\hat{x} - x_i)\| \leq g_3 \varepsilon \kappa^{\frac{1}{2}} \|A^{\frac{1}{2}}\| \|x_i\| \tag{10}$$

(cf. formula (2)).

So, an (A-)numerically stable DM does not necessarily solve a nearby linear system. However, the (A-)numerically stability concept indicates that the solution x_i is satisfactory from the following point of view. If a vector x satisfies (1) and if $\|E\| \sim \varepsilon \|A\|$ and $\|A^{-1}Ex\| \sim \|A^{-1}\| \|E\| \|x\|$ (which generally is the case if E is random), then $\|\hat{x} - x\| = \|A^{-1}Ex\| \sim$ $\sim \varepsilon \kappa \|x\|$. Consequently, if a DM is numerically stable then there exists an iterand x_i whose error is of the order of magnitude of the error of the exact solution of a nearby system (nearby in the sense that $\|E\| / \|A\|$ is of the order of the machine precision). This last error is called the *inherent error* (cf. Stoer and Bulirsch [80]). A similar statement holds for an A-numerically stable DM, if it is formulated in terms of the natural error and the *inherent natural* error $\varepsilon \kappa^{\frac{1}{2}} \|x\|$.

REMARK 3. Wóźniakowski [80] defines good-behavior of a DM generating a sequence $\{x_i\}$ by the relation

$$\overline{\lim} \|A(\hat{x} - x_i)\| \leq g_1 \varepsilon \|A\| \overline{\lim} \|x_i\| , \tag{11}$$

and numerical stability by the relation

$$\overline{\lim} \|\hat{x} - x_i\| \leq g_2 \varepsilon \kappa \overline{\lim} \|x_i\| . \tag{12}$$

Both definitions are stronger than our corresponding definitions in the sense that for our case the inequalities (2) and (5) have to be satisfied for only one approximation whereas for Wóźniakowski's case

these inequalities have to be satisfied ultimately for all approximations. For practical implications this is a minor difference. □

It is not assumed explicitly that the DM generates a finite sequence $\{x_i\}$. Of course, if a (infinite) DM is well-behaved, then the iteration might be terminated as soon as the last computed approximation x_i satisfies (5), with an acceptable g_3 (for this purpose one needs an estimate for $\|A\|$, which is often easy to obtain). If a DM is not well-behaved but if one knows (e.g. due to (A-)numerical stability) that the method will compute an approximation x_i with residual satisfying $\|b - Ax_i\| \leq g_3 \varepsilon \kappa^{\ell} \|A\| \|x_i\|$, for some g_3 and $\ell > 0$, then this inequality can be used as a stopping criterion (one then needs not only an estimate for $\|A\|$, but also for $\|A^{-1}\|$, which is probably hard to obtain). What is more, if (say) $\ell = \frac{1}{2}$, then one can only guarantee that the error of the last approximation x_i is of order $\varepsilon \kappa^{3/2} \|x_i\|$ and its natural error is of order $\varepsilon \kappa \|x_i\|$, which might be unacceptably large.

Another important performance indicator of an iterative process is the rate of convergence. The concepts of numerical stability and good-behavior measure the ability of the method to arrive at a "correct" answer. The concept of convergence rate indicates how much effort (number of iteration steps) is necessary to obtain that answer. Although there exist numerous notions on convergence behavior we define only two notions, related to the speed of convergence of a sequence of vectors $\{x_i\}$, which are adequate for our purposes.

DEFINITION 4. *If for some* $L < 1$ *the sequence* $\{y_i\}$ *satisfies*

$$\|y_i\| \leq L\|y_{i-1}\| , \quad (i > 0) , \tag{13}$$

then the sequence is said to converge step-wise linearly to zero with a convergence ratio no greater than L. □

Most authors assume that $\lim_{i \to 0} (\|y - y_{i+1}\| / \|y - y_i\|) =: L_0 < 1$ exists and then the convergence to y is called linear with asymptotic convergence ratio L_0. But then next they use this definition also for cases where the limit does not necessarily exist and only an upper bound in the sense of (13) can be given (like, e.g., for the gradient method (3.1.7)). An advantage of the latter definition is that if one wants to compare

two linearly convergent sequences with different convergence ratios, then definitely the sequence corresponding to the smaller convergence ratio is ultimately closer to the limit. Upper bounds in the sense of (13) are not very suitable for comparing two sequences, but they only give information about each separate sequence.

DEFINITION 5. *If for some* g *and* L < 1 *the sequence* $\{y_i\}$ *satisfies*

$$\|y_i\| \leq gL^i\|y_0\| \qquad (i > 0) \ , \tag{14}$$

then the sequence is said to converge linearly to zero on the average with an average convergence ratio no greater than L. □

It is obvious that step-wise linear convergence implies linear convergence on the average, but not, in general, vice versa.
In contradiction to the definitions of (A-)numerical stability and good-behavior the two definitions above do not depend on ε, but the definitions concern any (algebraic or computed) infinite sequences. In practice we never compute an infinite sequence, but still for a finite number of vectors $\{y_0,\ldots,y_k\}$ we use the definitions 3 and 4, indicating that the validity of (13) and (14) is restricted to the values of i satisfying $0 \leq i \leq k$.

1.5. *The use of the o-symbol in round-off error analysis*

In our round-off error analysis we meet equalities and inequalities involving the relative machine precision ε, the condition number κ and the constants C_1 and C_2 corresponding to round-off due to matrix by vector product computations and inner product computations, respectively (cf. section 1.3). In order to simplify the expressions we want to be able to neglect terms of order ε^2 in the presence of a term of order ε, with a minimal loss of relevant information. For this purpose we use the *Bachmann-Landau o-notation*. For instance, we write

$$\varepsilon(1 + C_1\kappa + \varepsilon C_2\kappa^{\frac{1}{2}}) = \varepsilon(1 + C_1\kappa + o(1)) \ , \quad [\varepsilon C_2\kappa^{\frac{1}{2}} \to 0] \ , \tag{1}$$

where the expression between square brackets indicates that $o(1)$ stands for a quantity that is small if $\varepsilon C_2\kappa^{\frac{1}{2}}$ is small. Of course, one

could also write down an explicit inequality, say

$$(2) \qquad \varepsilon(1 + C_1\kappa + \varepsilon C_2\kappa^{\frac{1}{2}}) \leq \varepsilon(1.06 + C_1\kappa) \ , \quad |\varepsilon C_2\kappa^{\frac{1}{2}}| \leq 0.06 \ ,$$

based on a rather arbitrary restriction. However, the use of the o-symbol has some advantages relative to the use of explicit constants, as will become clear from the following considerations.
We first give two formal definitions and some properties concerning the o-symbol.

DEFINITION 1. *Let* f, g, h *be three scalar functions defined on a set* $D \subseteq \mathbb{R}^{\ell}$ $(\ell \in \mathbb{N})$, *then*

$$(3) \qquad f(x) \leq o(g(x)) \ , \quad [h(x) \to 0] \ ,$$

means

$$\forall_{\eta>0} \exists_{\delta>0} \forall_{x\in D}: |h(x)| \leq \delta \Rightarrow f(x) \leq \eta|g(x)| \ . \qquad \square$$

Note that constant δ only depends on η and not on x; the implication holds uniformly with respect to $x \in D$. In fact, (1) only supplies information to those x for which h(x) is small. The expression between square brackets is referred to as the *restriction* under which (1) holds.
The following definition presents itself quite naturally.

DEFINITION 2. *Let* f, g, h *be three scalar functions defined on a set* $D \subseteq \mathbb{R}^{\ell}$ $(\ell \in \mathbb{N})$, *then*

$$(4) \qquad f(x) = o(g(x)) \ , \quad [h(x) \to 0]$$

means

$$f(x) \leq o(g(x)) \ , \quad [h(x) \to 0] \ ,$$

and

$$-f(x) \leq o(g(x)) \ , \quad [h(x) \to 0] \ . \qquad \square$$

The statement $f(x) = o(g(x))$, $[h(x) \to 0]$, thus means

$$\forall_{\eta>0} \exists_{\delta>0} \forall_{x\in D}: |h(x)| < \delta \Rightarrow |f(x)| \leq \eta|g(x)| \ .$$

Consequently, $|f(x)| \leq o(g(x))$ is equivalent with $f(x) = o(g(x))$,

$[h(x) \to 0]$. In our analysis we use both $\leq$ and $=$, one with another, in these situations.

In nearly all formulas containing the O-symbol, it appears in the form $O(1)$. We remark that we do not define the meaning of $O(g(x))$ itself; as it is often done in asymptotic analysis (cf. De Bruijn [61]), we only give the interpretation of some complete formulas.
For instance, if we write for two scalar functions f_1 and f_2, with $f_2(x) > 0$ $(x \in D)$,

$$f_1(x) \leq f_2(x)(1+O(1)) \ , \quad [h(x) \to 0] \ , \tag{5}$$

we mean

$$(f_1(x) - f_2(x)) / f_2(x) \leq O(1) \ , \quad [h(x) \to 0] \ , \tag{6}$$

in the sense of definition 1. We also write relations like

$$f_1(x)O(1) = O(1) \ , \quad [h(x) \to 0] \ , \tag{7}$$

which statement is to be interpreted as follows. For any function f_2 for which $f_2(x) = O(1)$, $[h(x) \to 0]$, one also has $f_1(x)f_2(x) = O(1)$, $[h(x) \to 0]$. In these cases the expressions involving O-symbols have to be considered as a class of functions (compare also the properties below).

Some rather trivial but often used properties are the following.

PROPERTIES 3.

(i) $f(x) = O(1) \ , \quad [f(x) \to 0] \ ,$

(ii) $O(1) + O(1) = O(1) \ , \quad [h(x) \to 0] \ ,$

(iii) $O(1)O(1) = O(1) \ , \quad [h(x) \to 0] \ ,$

(iv) $(1+O(1))^{-1} = 1+O(1) \ , \quad [h(x) \to 0] \ .$ □

The last three properties indicate that the O-symbol is easy to handle and that is our main reason for using it.
Another advantage of O-symbols above explicit constants is that the coefficients in the relations involving O- and $=$-symbols are more or less uniquely determined (compare $(1-\varepsilon)^{-1} = 1+\varepsilon+O(1)$, $[\varepsilon \to 0]$ and $1+\varepsilon \leq (1-\varepsilon)^{-1} \leq 1 + (1+1/3)\varepsilon$, $[0 < \varepsilon < \frac{1}{4}]$).

A disadvantage of the use of O-symbols is that we do not obtain explicit bounds. However, in all cases where we derive formulas with O-symbols it is possible to retrace the proof, replacing all O-formulas by estimates involving explicit numerical constants. That is, at every stage of the proof we are able to indicate definite numbers, where the asymptotic estimates only state the existence of such numbers (compare the proof of theorem 2.3.4 and the proof of proposition 2.3.12). But in most cases the final estimates are obtained by means of a considerable number of steps and in each step a factor 2 or so, in the estimates, is easily lost. Quite often it is possible to reduce such losses by a more careful examination.
We are primarily interested in studying how the matrix condition number κ and the constants C_1, C_2 affect the various error estimates. For this purpose the O-notation supplies sufficient information if it is used in an appropriate way, which means that one checks at every stage whether a formula holds uniformly with respect to the relevant parameters.

REMARK 4. Wilkinson [65] uses explicit bounds in his error analysis. The application of the basic relations mentioned in section 1.3 frequently leads in the first instance to bounds of the form

$$(8) \qquad (1-\varepsilon)^{\ell} \leq 1+\mu \leq (1+\varepsilon)^{\ell}, \quad (\ell \in \mathbb{N})$$

and these are somewhat inconvenient. In order to simplify such bounds Wilkinson makes the assumption that in all practical applications ℓ will be subject to the restriction $\ell\varepsilon < 1/10$. With this restriction one has

$$(9) \qquad (1+\varepsilon)^{\ell} < 1+(1.06)\ell\varepsilon, \quad (1-\varepsilon)^{\ell} > 1-(1.06)\ell\varepsilon.$$

Therefore he defines $\varepsilon_1 := (1.06)\varepsilon$, which is only marginally different from ε and enables him to replace relation (8) by $|\mu| < \ell\varepsilon_1$, whenever this is advantageous. However, this leads to many explicit constants like 12.36, 1.501, etc., which is why we refrained from following the same strategy.
Wózniakowski [80] uses the relation $\doteq$ which is defined as follows. Let f and g be two scalar functions defined on $[0,\varepsilon_0]$. Then $f(\varepsilon) \doteq g(\varepsilon)$ means that there exists a constant K and a scalar function h such that

$f(\varepsilon) = g(\varepsilon)(1+h(\varepsilon))$, where $|h(\varepsilon)| \le K\varepsilon$ for $0 \le \varepsilon \le \varepsilon_0$. The relation $f(\varepsilon) \dot{\le} g(\varepsilon)$ now means $f(\varepsilon) \le g(\varepsilon)$ or $f(\varepsilon) \doteq g(\varepsilon)$. These relations enabled him to ignore terms of order ε^2 in the presence of a term of order ε. Not very much attention is payed by him to uniformity as far as C_1, C_2 and κ are concerned. For example, we distinguish between

$$\varepsilon\kappa^{\frac{1}{2}} + \varepsilon^2\kappa^2 = \varepsilon\kappa^{\frac{1}{2}}(1+o(1)) \quad , \quad [\varepsilon\kappa^{3/2} \to 0] \ ,$$

and

$$\varepsilon\kappa^{\frac{1}{2}} + \varepsilon^2\kappa = \varepsilon\kappa^{\frac{1}{2}}(1+o(1)) \quad , \quad [\varepsilon\kappa^{\frac{1}{2}} \to 0] \ ,$$

whereas Wóźniakowski would write in both cases

$$\varepsilon\kappa^{\frac{1}{2}} + \varepsilon^2\kappa^2 \doteq \varepsilon\kappa^{\frac{1}{2}} \ ,$$

$$\varepsilon\kappa^{\frac{1}{2}} + \varepsilon^2\kappa \doteq \varepsilon\kappa^{\frac{1}{2}} \ .$$

1.6. Test problems and implementations

In carrying out computational experiments for testing mathematical software there are two main types of test problems (cf. Crowder, Dembo and Mulvey [79]): those which are representative real-world application problems and those which are "constructed" problems. The first type is used to give an indication of the behavior for practical problems, whereas the second type is used to investigate specific aspects of a method which might be exercised infrequently in application of the method on real-world problems. We only performed numerical tests with problems of the second type, generated pseudorandomly. Our test problems are designed to verify the validity of our analytical results, which deal with attainable accuracy of approximate solutions and give upper bounds for convergence ratios (cf. section 1.3). Moreover, we want to investigate whether and under which conditions these estimates are best-possible, or essentially best-possible in the sense that they contain the correct power of κ. This last goal justifies the use of constructed problems, where the characteristics of the population, from which a problem is drawn, are known and can be controlled. If an estimate turns out to be best-possible for some class of constructed test problems, then there remains the question to what extent

this class is representative for real-world problems. The answer to this question will be discussed only incidentally.

Numerical experiments have only been carried out for the GM and the CGM. The algebraic performance of these two methods depends on the data A, b and the initial vector x_0. The numerical performance of these methods not only depends on A, b and x_0, but in addition on the way of implementation of the various arithmetical operations. In the following we describe how A, b and x_0 are constructed and how the arithmetical operations are implemented.

The choice of the matrix A

Every $n \times n$ *positive definite* matrix A can be written in terms of its *spectral decomposition*

$$A = U\Lambda U^T , \tag{1}$$

where

- U is an orthogonal $n \times n$ matrix, whose columns are a complete set of n *orthonormal eigenvectors* $u_1,\ldots,u_n$ of the matrix A;
- Λ is an $n \times n$ diagonal matrix whose diagonal entries λ_i are the n positive *eigenvalues* of A.

Without loss of generality we always assume $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n = 1$ (hence $\|A\| = 1$ and $\kappa = \lambda_1^{-1}$).
Obviously A is determined completely by Λ and U and hence choosing A is equivalent with choosing Λ and U.

The diagonal matrices Λ can be controlled in a trivial way by choosing its diagonal entries λ_i.

Algebraically, the GM and the CGM are invariant relative to orthogonal basic transformations (see remark 1.3.3). Consequently, the algebraic performance of these two methods is completely determined by the eigenvalue distribution of A and the *eigenvector components* (components with respect to the basis of eigenvectors) of the vectors b and x_0. Therefore an obvious choice for U would be U = I. In the presence of round-off however, the whole structure of A (and consequently the choice of U) affects the numerical performance (cf. remark 1.3.2). Since the round-off occurring at the computation of $fl(\Lambda x)$ is

certainly not representative for the real-world computation of fl(Ax), we have to take special arrangements if we choose U = I.
Another obvious choice is to construct pseudorandomly orthogonal matrices U. The two alternatives are evaluated in the sequel.

PSEUDORANDOMLY GENERATED ORTHOGONAL MATRICES. The construction of a pseudorandomly generated orthogonal matrix can be accomplished in various ways. For instance, U can be constructed as a product of a number of random Householder transformations or Givens transformations or by Gram-Schmidt orthogonalization performed on a matrix with random entries (cf. Stewart [80]). We only experimented with matrices U constructed as a product of a number of Householder transformations. Intuitively one feels that the number of Householder transformations must be rather large to guarantee the random character of U.

THE CASE U = I. Choosing U = I implies that we have a test matrix A with eigenvectors $e_1,\ldots,e_n$ (unit vectors). From a numerical point of view this is a rather special choice.

Implementation of the matrix by vector product computations

Once Λ and U are selected we have to decide how to implement the computation of $Av = U\Lambda U^T v$ for an arbitrary machine vector v.

PSEUDORANDOMLY GENERATED ORTHOGONAL MATRICES. In cases where U is chosen pseudorandomly an obvious way of implementing the computation of Av is first to assemble (and store) the matrix A by computing $A = fl(U\Lambda U^T)$ (in some way), and next to compute any matrix by vector product straightforwardly by computing inner products of rows of A and the vector involved. This way of implementation will be referred to as *assembled implementation* (AI). If the assembled A has not a rather special structure, then the round-off occurring if computing Av in this way agrees with the real-world situation (cf. remark 1.3.2). During the assembly of A round-off occurs, but if ε_K is appreciably less than unity (and assuming symmetry is preserved), then certainly the computed matrix is positive definite although its exact eigenvalue distribution is slightly different from the chosen one. In general, the matrix A, constructed in this way, is not sparse and every matrix by vector computation takes n^2 multiplications.

If U is constructed as a product of (not too many) elementary orthogonal transformations, then computing time can be reduced significantly by keeping A in product form. For instance, if U is constructed as a product of m *Householder transformations* $U = H_m \cdots H_1$, where each H_i corresponds to a random vector h_i $(i = 1,\ldots,m)$ such that

$$H_i := I - 2\, h_i h_i^T / (h_i, h_i) \quad , \tag{2}$$

then instead of assembling A and computing Av straightforwardly, this matrix by vector product could be computed (from right to left) from the relation

$$Av = H_m \cdots H_1 \Lambda H_1 \cdots H_m v \quad . \tag{3}$$

This way of implementation will be referred to as *product form implementation* (PFI). For each Householder transformation the computation of $H_i w$ costs $2n + 1$ multiplications (apart from the computation of $2 / (h_i, h_i)$ which has to be carried out only once). Hence, if Av is computed using (3), this costs about $(4m+1)n$ multiplications. Thus, from this point of view, implementing Av based on (3) is cheaper if (roughly) $m < n/4$. If m is much smaller than n, computational time is reduced significantly. However, as we shall discuss in section 3.4, for small values of m the round-off occurring at the computation of Av, based on (3), is certainly not representative for the real-world situation.

THE CASE U = I. In cases where we take the identity matrix for U we have $A = \Lambda$. One might think of computing Λv by just multiplying each component of v with the corresponding eigenvalue. However, this way of implementation is certainly not a real-world implementation. One has $fl(\Lambda v) = (\Lambda + E)v$, where $|E| \leq \varepsilon\Lambda$. Hence our general condition on round-off errors due to matrix by vector products, $\|E\| \leq \varepsilon C_1 \|\Lambda\|$, is certainly satisfied (with $C_1 = 1$). However, the vector Ev is approximately parallel to the vector Λv whereas for the real-world implementation this vector is rather randomly directed (cf. remark 1.3.2). To remedy this drawback we use a kind of *artificial floating point implementation* (AFI) for the computation of Λv in the following way. We first compute $u := fl(\Lambda v)$ and next add a vector u' with components chosen randomly from the interval $[-\delta\|\Lambda\|\|v\| \, , +\delta\|\Lambda\|\|v\|]$, where $\delta > 0$ is

a fixed number (fixed throughout the whole performance of the algorithm) called the *artificial relative precision*. If fla(Λv) denotes the vector Λv computed in this way, then we have

(4) $\quad fla(\Lambda v) = fl(fl(\Lambda v) + fl(\gamma e))$,

where e is a machine vector with components randomly chosen from the interval $[-1,+1]$ and $\gamma := fl(\delta\|v\|)$ (since $\|\Lambda\| = 1$). We assume that the computation of γ is carried out such that the relative error does not exceed $(\frac{1}{2}n+2)\varepsilon(1+o(1))$, under the restriction $n\varepsilon \to 0$. The following considerations show that, if $\delta \gg \varepsilon$, the computation of Λv by using (4) simulates the real-world computation of a general matrix by vector product, using floating point arithmetic with relative precision δ.

LEMMA 1.

(5) $\quad fla(\Lambda v) = \Lambda v + w$,

where the components of w *satisfy*

(6) $\quad w_j = \mu_j(\Lambda v)_j + \gamma(1+\sigma_j)e_j$

with

(7) $\quad |\mu_j|, |\sigma_j| \leq 2\varepsilon(1+o(1))$, $\quad [n\varepsilon \to 0] \quad (j = 1,\ldots,n)$.

Furthermore,

(8) $\quad fla(\Lambda v) = (\Lambda + E)v$,

where

(9) $\quad \|E\| \leq (n^{\frac{1}{2}}\delta + 2\varepsilon)(1+o(1))$, $\quad [n\varepsilon \to 0]$.

PROOF. All o-symbols are assumed to hold under the restriction $n\varepsilon \to 0$. Using the preliminaries of section 1.3 we note that each component of fla(Λv) satisfies

(10) $\quad (fla(\Lambda v))_j = (\lambda_j v_j(1+\varepsilon_j) + \gamma e_j(1+\tau_j))(1+\rho_j)$,

where $|\varepsilon_j|, |\tau_j|, |\rho_j| \leq \varepsilon$.
Consequently,

(11) $\quad w_j = (fla(\Lambda v) - \Lambda v)_j = \mu_j \lambda_j v_j + \gamma e_j(1+\sigma_j)$,

where

(12) $\quad |\mu_j| = |(1+\varepsilon_j)(1+\rho_j) - 1| \leq 2\varepsilon(1+O(1)) ,$

and

(13) $\quad |\sigma_j| = |(1+\tau_j)(1+\rho_j) - 1| \leq 2\varepsilon(1+O(1)) ,$

which proves (6).
Defining $E := wv^T / (v,v)$ we conclude from (5) that (8) is satisfied.
From (6) it follows that

(14) $\quad \|E\| \leq \|w\| / \|v\| \leq (2\varepsilon\|\Lambda v\| + \delta\|v\|\|e\|) / \|v\|(1+O(1)) \leq$

$$\leq (2\varepsilon + n^{\frac{1}{2}}\delta)(1+O(1)) ,$$

which proves (9). □

Since $|\mu_j(\Lambda v)_j| \leq 2\varepsilon\|v\|(1+O(1))$ and $|\gamma(1+\sigma_j)e_j| = \delta\|v\||e_j|(1+O(1))$ the vector w is approximately parallel to the random vector e, if $\delta \gg \varepsilon$. Hence, the vector w is randomly directed. Therefore, if $\delta \gg \varepsilon$, then $fla(\Lambda v)$ agrees with a real-world implementation of Av on a floating point machine with relative machine precision δ (but with $C_1 = n^{\frac{1}{2}}(1+O(1))$ instead of $C_1 = n^{3/2}(1+O(1))$; cf. remark 1.3.2). Of course, for every matrix by vector product that must be computed during the performance of the algorithm one has to choose a different random vector e, since otherwise all corresponding round-off vectors w are parallel. Each matrix by vector product based on (4) costs (apart from choosing n random numbers) $3n+1$ multiplications and 1 square root operation, which is favourable if compared with the assembled implementation.

The choice of b

As far as the choice of b is concerned we want to be able to control directly the eigenvector components s_j of the solution vector $\hat{x} = A^{-1}b$. With respect to AI or PFI this can be achieved by choosing s and computing directly $b = fl(U\Lambda s)$, or by computing first $\bar{x} = fl(Us)$ and next $b = fl(A\bar{x})$.

REMARK 2. If b is computed from $b = fl(U\Lambda s)$, then there holds for some C_1'

(15) $\quad b = (U\Lambda + E)s , \quad \|E\| \leq \varepsilon C_1'\|A\| ,$

and consequently $U^T A^{-1} b - s = \Lambda^{-1} U^T E s$, which indicates that the eigenvector components s_j' of $A^{-1}b$ not necessarily have a small relative error. Furthermore, it follows that $\|U^T A^{-1} b - s\| \leq \varepsilon C_1' \kappa \|s\|$. Similar results hold for the situation where b is computed from $\bar{x} = fl(Us)$, $b = fl(A\bar{x})$. □

For AFI the eigenvector components of $\hat{x}$ can be controlled in a trivial way by computing $b = fl(\Lambda s)$ (but not $b = fla(\Lambda s)$).

REMARK 3. If b is computed from $b = fl(\Lambda s)$, then one has

$$b = \Lambda(I+E)s \ , \quad |E| \leq \varepsilon I \ . \tag{16}$$

Consequently, $|(A^{-1}b - s)_j| = |(Es)_j| \leq \varepsilon |s_j|$, which implies that componentwise $A^{-1}b$ equals s up to machine precision ε. □

The choice of x_0

As far as the choice of x_0 is concerned we want to be able to control directly the eigenvector components e_j of the initial error $\hat{x} - x_0$ ($\hat{x} - x_0 = Ue$). For AI or PFI this can be achieved by choosing e and computing $x_0 = fl(U(s-e))$ or $x_0 = fl(\bar{x} - Ue)$, where $\bar{x} = fl(Us)$. For AFI this can be accomplished by computing $x_0 = fl(s-e)$.

Implementation of the basic dyadic arithmetical operations and the inner product computations

If in a test problem matrix by vector product computations are based on AI or PFI, then the other basic dyadic arithmetical operations, i.e., vector addition, vector subtraction, scalar by vector product and scalar division, are implemented in the obvious way. The inner product computations are also implemented in the obvious way, described in remark 1.3.1.

If in a test problem matrix by vector product computations are based on AFI with artificial relative precision δ ($\delta \gg \varepsilon$), then the other basic operations are adapted in the following way (x, y are machine vectors, a, b are machine numbers).

$$fla(x \pm y) = fl(fl(x \pm y) + fl(\mu\tilde{e})) \ , \tag{17}$$

$$fla(ax) = fl(fl(ax) + fl(\tau\tilde{e})) \ , \tag{18}$$

(19) $\quad fla((x,y)) = fl(fl(x,y) + fl(\delta'\|x\|\|y\|))$,

(20) $\quad fla(a/b) = fl((a/b)(1+\delta''))$,

where $\mu := fl(\delta\|x \pm y\|)$, $\tau := fl(\delta\|ax\|)$, $\bar{e}$ and $\tilde{e}$ are machine vectors with components randomly chosen from the interval $[-1,+1]$ and δ' and δ'' are scalars randomly chosen from the interval $[-\delta,+\delta]$.

REMARK 4. We note that formulas (17), (18) and (19) do not agree with the basic relations and inequalities (3) to (10) of section 1.3. For instance, in general the diagonal matrix F_1 defined by the relation

(21) $\quad fla(x+y) = (I+F_1)(x+y)$,

does not satisfy $|F_1| \leq \delta I$. However, there exists a not necessarily diagonal matrix F_1 satisfying (21) for which, neglecting terms of the order ε, $\|F_1\| \leq \delta$ holds. This also applies to the other round-off matrices F_2, G_1, G_2, D corresponding to the specific fla-operations. In our round-off error analysis we never use the fact that for floating point arithmetic the round-off matrices are diagonal; we only use the upper bound for their norms. Consequently, our round-off error analysis also holds for fla-arithmetic.
An important property of the AFI is that this implementation is invariant relative to orthogonal basis transformations.
A more obvious adaptation of the operations would be

(22) $\quad fla(x \pm y) = fl((I+\Phi_1)(x+y))$,

(23) $\quad fla(ax) = fl((I+\Phi_2)ax)$,

(24) $\quad fla((x,y)) = fl(((I+\Phi_3)x,y))$,

where Φ_1, Φ_2, Φ_3 are diagonal matrices with diagonal entries randomly chosen from the interval $[-\delta,+\delta]$, since these formulas agree automatically with the basic relations and inequalities (3) to (10) of section 1.3. However, these fla-operations do not always correspond to the real-world implementation on a machine with relative machine precision δ. If $U = I$, then the elements of a vector x are its eigenvector components. Hence, if the vector $\hat{x}$ has a special structure of eigenvector components, then ultimately also the approximations x_i will have that special structure. This implies, for instance, that

ultimately the round-off vectors due to the fla-addition (22) in updates like $x_{i+1} = x_i + a_i p_i$ will be more or less parallel to the (special) direction of $\hat{x}$. This unrealistic situation cannot occur for fla-addition based on (17). □

REMARK 5. AFI enables us to simulate even one-round-off error analysis (cf. section 1.3). For instance, if a one-round-off error analysis is carried out, only taking into account round-off at the matrix by vector product computations, then AFI of Av with $\delta \gg \varepsilon$, and performing all other arithmetical operations in normal floating point arithmetic, simulates the situation corresponding to this one-round-off error analysis. □

If we want to perform different tests with the same eigenvalue distribution and the same eigenvector components for $\hat{x}$ and $\hat{x} - x_0$ but with different round-off patterns, then for AI and PFI this can be achieved by selecting different matrices U_m (which means selecting different stes of vectors $\{h_1,\ldots,h_m\}$ for the Householder transformations). In case we deal with AFI this can be achieved by selecting different random numbers from the interval $[-\delta,+\delta]$.

We had carried out quite a lot of tests problems based on AI and PFI before we got aware of the advantages of AFI. However, we did not repeat all of these former experiments using AFI.

The numerical experiments were performed on the Burroughs B7700 computer ($B = 8$, $t = 13$, $\varepsilon = \frac{1}{2}8^{-12} \sim 7.3_{10}\text{-}12$). For the generation of random numbers we used the arithmetic function RANDOM, intrinsic to Burroughs Extended Algol, described in: B7000/B6000 Series, System Software, Operational Guide, Vol. 1, p. 9.2.4 (1977).

CHAPTER 2

DESCENT METHODS (DM)

2.1. *Introduction*

In this chapter we consider the numerical process of solving a definite linear system

(1) $\quad Ax = b \ ,$

by a descent method (DM). Every descent method for solving (1) is coupled with a so called objective function $F: \mathbb{R}^n \to \mathbb{R}^n$, This objective function is chosen in such a way that the solution $\hat{x}$ of the linear system is a global minimum of F.

The fundamental underlying structure of descent methods (see Luenberger [73]) is as follows. Starting at an initial point one determines, according to a fixed rule, a direction of movement, and then moves in that direction to a minimum of the given objective function F on that line. At the new point a new direction is determined and the process is repeated. The main difference between various descent methods rests with the rule by which successive directions of movement are selected. Once the direction is chosen, the method determines the point on the corresponding line for which the objective function attains its minimal value. This indicates a second difference, namely the choice of the objective function.

In contrast with direct methods, like for instance Gaussian elimination, the descent methods do not alter the original matrix. In fact, it is possible to avoid storing the matrix explicitly. All that is required is a subroutine that produces Ax for a given vector x. This is one of the main reasons why descent methods became attractive for solving large sparse linear systems. Full advantage can be taken of the sparsity structure of A and no assumptions need to be made about the pattern of nonzeros. Also the storage requirements are quite modest and the implementation is easy.

Some well-known descent methods are the Gauss-Seidel method, the Gauss-Southwell method, the gradient method (or steepest descent method) and the conjugate gradient method.
In this thesis we restrict ourselves to the case where the $n \times n$ matrix A is positive definite. We consider only descent methods for which the objective function F is represented by the quadratic form

$$F(x) = ((\hat{x} - x), A(\hat{x} - x)) , \tag{2}$$

where $\hat{x} \neq 0$ is the solution of the linear system.
In the Euclidean norm (2) can be written as

$$F(x) = \| A^{\frac{1}{2}}(\hat{x} - x) \|^2 . \tag{3}$$

All descent methods mentioned before are based on this objective function. For descent methods based on objective functions of the form $\| A^{\alpha}(\hat{x} - x) \|^2$ $(2\alpha \in \mathbb{N})$ one can derive similar results as are derived in this thesis for the case $\alpha = \frac{1}{2}$.

We now summarize the contents of Chapter 2.
In section 2 we deduce some well-known elementary algebraic properties of the descent methods. The main reason of deducing these properties here is that they are basic for studying the behavior of the methods in the presence of round-off. We also formulate explicitly the well-known descent methods mentioned above.
In executing a descent method there are two possible ways of computing the residual vector $r_i := b - Ax_i$, belonging to each successive approximation x_i. One way is to compute the residual vector directly from this definition; residuals computed in this way are called *true residual vectors*. The second way is to compute the residual vector by updating, using the recurrence relation for two successive residual vectors; residual vectors computed in this way are called *recursive residual vectors*. It turns out that there is a great difference between algorithms using true residuals and recursive residuals in the presence of round-off.
Section 3 deals with the numerical behavior of descent methods if recursive residuals are used. It consists of two subsections. In the first subsection a round-off error analysis is presented of one step of the process and subsequently this result is used to prove the stepwise linear convergence of the objective function, expressed in terms

of the recursive residual vector. In the second subsection we derive an upperbound for the value of the objective function, expressed in terms of the computed approximations, for large values of i.
Section 4 deals with the numerical behavior of the descent methods if true residuals are used. First we give a round-off error analysis of one step of the process. Next we use this result to prove the stepwise linear convergence of the objective function, expressed in terms of the true residual vector and we give an upper bound for the attainable accuracy of the computed approximations. Finally we apply the theoretical results of this section to the Gauss-Southwell method in order to show how the general theory leads to assertions on good-behavior and numerical stability in a specific example.

2.2. *Algebraic properties of descent methods*

In this section we formulate the DM and deduce some important algebraic properties, i.e. properties that are valid if no round-off occurs. We shall interprete these results for some specific DM's.
Given a definite system

(1) $Ax = b$,

then the DM, corresponding to a given sequence of arbitrary nonzero vectors $\{p_i\}$, is defined by the following statements.

Descent Method (DM)
Choose an initial point x_0;
$r_0 := b - Ax_0$; $i := 0$;
while $r_i \neq 0$ do
begin

(2) $a_i := (r_i, p_i) / (p_i, Ap_i)$;

(3) $x_{i+1} := x_i + a_i p_i$;

(4) (5) $r_{i+1} := \begin{cases} \text{either } b - Ax_{i+1} ; & \text{(TRDM)} \\ \text{or } \quad r_i - a_i Ap_i ; & \text{(RRDM)} \end{cases}$

$i := i + 1$

end.

The residual vector r_{i+1} can be computed from either formula (4) or formula (5).

If the residual vector is computed from (4), then this residual vector is called a *true residual vector* (cf. section 2.1), and a DM where all residual vectors are computed from (4) is called a *true residual descent method* (TRDM).

From statement (3) it follows that

$$b - Ax_{i+1} = (b - Ax_i) - a_i Ap_i , \tag{6}$$

which gives the recurrence relation (5) if translated in terms of the residual vectors.

Therefore, if the residual vector is computed from (5), this residual vector is called a *recursive residual vector* and a DM where all residual vectors are computed from (5) is called a *recursive residual descent method* (RRDM).

Of course, if exact arithmetic is used, the approximations $\{x_i\}$ generated by RRDM and by TRDM are exactly the same. However, this certainly is not the case when both methods are performed using floating point arithmetic.

As far as the computational work is concerned, the most expensive operation is in general the matrix by vector product. For TRDM two matrix by vector products are needed, for RRDM only one. For both methods, apart from the vector p_i, one needs to store the vectors x_i, r_i and Ap_i only during the step from i to i + 1. There is no need to know all vectors p_i in advance, they could as well be computed (and stored for one step) as the process proceeds.

REMARK 1. Of course, computing all residual vectors either from (4) or (5) is not absolutely necessary. One might as well compute r_{i+1} from relation (4) every (say) 10 steps and use formula (5) in all other steps. This will be called a *mixed descent method* (MDM). □

We now prove some well-known, elementary algebraic properties of DM's. Since algebraically RRDM and TRDM are equivalent, there is no need to distinguish between them.

THEOREM 2. *At each step a DM minimizes the objective function*

$$F(x) := ((\hat{x} - x), A(\hat{x} - x)) = \| A^{\frac{1}{2}}(\hat{x} - x)\|^2 \tag{7}$$

along the line $x = x_i + ap_i$ *and*

$$\text{(i)} \qquad \frac{\|A^{\frac{1}{2}}(\hat{x} - x_{i+1})\|^2}{\|A^{\frac{1}{2}}(\hat{x} - x_i)\|^2} = \frac{\|A^{-\frac{1}{2}} r_{i+1}\|^2}{\|A^{-\frac{1}{2}} r_i\|^2} = 1 - \gamma_i^2$$

where

$$\text{(8)} \qquad \gamma_i := \frac{|(r_i, p_i)|}{\|A^{-\frac{1}{2}} r_i\| \|A^{\frac{1}{2}} p_i\|} .$$

Further

$$\text{(ii)} \qquad \|A^{-\frac{1}{2}} r_{i+1}\|^2 + \|a_i A^{\frac{1}{2}} p_i\|^2 = \|A^{-\frac{1}{2}} r_i\|^2 ,$$

$$\text{(iii)} \qquad (r_{i+1}, p_i) = 0 .$$

PROOF. We have

$$\text{(9)} \qquad F(x_i + ap_i) = \|A^{\frac{1}{2}}(\hat{x} - x_i - ap_i)\|^2 =$$

$$= F(x_i) - 2a(A(\hat{x} - x_i), p_i) + a^2 \|A^{\frac{1}{2}} p_i\|^2 =$$

$$= F(x_i) + \|A^{\frac{1}{2}} p_i\|^2 \left(a - \frac{(A(\hat{x} - x_i), p_i)}{\|A^{\frac{1}{2}} p_i\|^2} \right)^2 - \frac{(A(\hat{x} - x_i), p_i)^2}{\|A^{\frac{1}{2}} p_i\|^2} ,$$

which is minimal for

$$\text{(10)} \qquad a = \frac{(A(\hat{x} - x_i), p_i)}{\|A^{\frac{1}{2}} p_i\|^2} = \frac{(r_i, p_i)}{(p_i, Ap_i)} = a_i$$

and the minimal value $F(x_i + a_i p_i) = \|A^{\frac{1}{2}}(\hat{x} - x_{i+1})\|^2$ satisfies

$$\text{(11)} \qquad \|A^{\frac{1}{2}}(\hat{x} - x_{i+1})\|^2 = \|A^{\frac{1}{2}}(\hat{x} - x_i)\|^2 - (r_i, p_i)^2 / \|A^{\frac{1}{2}} p_i\|^2 .$$

Since $A^{\frac{1}{2}}(\hat{x} - x_i) = A^{-\frac{1}{2}}(A\hat{x} - Ax_i) = A^{-\frac{1}{2}} r_i$ and $(r_i, p_i)^2 / \|A^{\frac{1}{2}} p_i\|^2 = \|a_i A^{\frac{1}{2}} p_i\|^2$, formulas (i) and (ii) follow readily from (11). Using (5) we obtain

$$\text{(12)} \qquad (r_{i+1}, p_i) = (r_i, p_i) - a_i (p_i, Ap_i) = 0 ,$$

which proves (iii). □

The vector $A^{\frac{1}{2}}(\hat{x} - x_i)$ is something in between the error vector $\hat{x} - x_i$ and the residual vector $A(\hat{x} - x_i)$. Since a DM minimizes the error $\|A^{\frac{1}{2}}(\hat{x} - x_i\|$ at each separate step, it seems natural to measure the error this way, instead of measuring $\|\hat{x} - x_i\|$ or $\|A(\hat{x} - x_i\|$. Therefore we call this error the *natural error* and $A^{\frac{1}{2}}(\hat{x} - x_i)$ is called the *natural error vector* (cf. section 1.4).

REMARK 3. The gradient vector of the objective function $F(x)$ at point x_{i+1} equals

$$(13) \qquad -2A(\hat{x} - x_{i+1}) = -2(b - Ax_{i+1}) = -2r_{i+1} .$$

Consequently, relation (iii) states that the gradient vector of the objective function at the minimal point on the line $x = x_i + ap_i$ is orthogonal to the direction of that line. This is a well-known necessary condition for minimization (see Luenberger [73]). □

From (i) of theorem 2 it follows that

$$(14) \qquad \|A^{\frac{1}{2}}(\hat{x} - x_{i+1})\|^2 = \|A^{\frac{1}{2}}(\hat{x} - x_i)\|^2 \prod_{\ell=0}^{i} (1 - \gamma_\ell^2) ,$$

and this infinite product diverges to zero iff $\Sigma_{\ell=0}^{i} \gamma_\ell^2$ diverges. Hence we have the following corollary of theorem 2.

COROLLARY 4. *The sequence* $\{x_i\}$, *generated by a DM converges to the solution* $\hat{x}$ *iff* $\Sigma_{i=0}^{\infty} \gamma_i^2$ *diverges.* □

Note that $\cos^{-1} \gamma_i$ is the angle between the vectors $A^{-\frac{1}{2}} r_i$ and $A^{\frac{1}{2}} p_i$.

If there exists a $\gamma > 0$ and an infinite subset N of $\mathbb{N}_0$ such that $\gamma_i > \gamma$ for all $i \in N$, then the series $\Sigma_{i=0}^{\infty} \gamma_i^2$ diverges. Stated differently, if the angle between $A^{-\frac{1}{2}} r_i$ and $A^{\frac{1}{2}} p_i$ is bounded away from $\pi/2$ for an infinite subset of vectors p_i, then the natural error tends to zero.

For the Gauss-Southwell method, the gradient method and the conjugate gradient method we shall show that the choice $N = \mathbb{N}_0$ is possible. This leads to the second corollary of theorem 2.

COROLLARY 5. *If* $\{x_i\}$, $\{r_i\}$ *are generated by a DM and if there exists a* $\gamma > 0$ *such that* $\gamma_i > \gamma$ *for all* $i \geq 0$ *then the natural error converges step-wise linearly to zero and for all* $i \geq 0$

$$\frac{\|A^{\frac{1}{2}}(\hat{x} - x_{i+1})\|^2}{\|A^{\frac{1}{2}}(\hat{x} - x_i)\|^2} = \frac{\|A^{-\frac{1}{2}} r_{i+1}\|^2}{\|A^{-\frac{1}{2}} r_i\|^2} \leq 1 - \gamma^2 . \tag{15}$$

□

In the cases we are dealing with it is in general difficult to determine directly a lower bound for γ_i. Therefore, in our numerical analysis of DM's we often consider, instead of the parameter γ_i, the parameters α_i and β_i defined by

$$\alpha_i := \frac{\|r_i\| \, \|p_i\|}{|(r_i, p_i)|} \tag{16}$$

and

$$\beta_i := \frac{\|r_i\| \|A^{\frac{1}{2}} p_i\|}{\|A^{\frac{1}{2}}\| \, |(r_i, p_i)|} . \tag{17}$$

It turns out that γ_i can be bounded in terms of α_i and β_i; this is the contents of the following lemma.

LEMMA 6. *Suppose* r *and* p *are two arbitrary vectors for which* $(r,p) \neq 0$ *and let*

$$\alpha := \|r\|\|p\| \, / \, |(r,p)| ,$$

$$\beta := \|r\|\|A^{\frac{1}{2}} p\| \, / \, (\|A^{\frac{1}{2}}\| \, |(r,p)|) ,$$

$$\gamma := |(r,p)| \, / \, (\|A^{-\frac{1}{2}} r\|\|A^{\frac{1}{2}} p\|) ,$$

then

$$\kappa^{-\frac{1}{2}}\alpha \leq \beta \leq \gamma^{-1} \leq \kappa^{\frac{1}{2}}\beta \leq \kappa^{\frac{1}{2}}\alpha , \tag{18}$$

where κ *is the condition number of* A.

PROOF. The various inequalities follow easily from

$$\|A^{-\frac{1}{2}}\|^{-1}\|A^{-\frac{1}{2}} r\| \leq \|r\| \leq \|A^{\frac{1}{2}}\|\|A^{-\frac{1}{2}} r\|$$

and

$$\|A^{\frac{1}{2}}\|^{-1}\|A^{\frac{1}{2}} p\| \leq \|p\| \leq \|A^{-\frac{1}{2}}\|\|A^{\frac{1}{2}} p\| .$$

□

REMARK 7. Instead of minimizing the objective function at each separate step of the process, one might be concerned only with diminishing the objective function at each separate step. We have (see (9)):

$$f(a) := F(x_i + ap_i) = F(x_i) - 2a(r_i,p_i) + a^2\|A^{\frac{1}{2}} p_i\|^2 ,$$

which is a quadratic function in a and $f(0) = F(x_i)$. Since f is minimal at a_i, f is symmetric around a_i and consequently $f(a) < f(0)$ for all a satisfying $|a_i - a| < |a_i - 0|$. Stated differently, if $a = \omega a_i$, for some $0 < \omega < 2$, then $F(x_i + ap_i) < F(x_i)$. Hence, if instead of (2) one takes $a_i := \omega_i(r_i,p_i) / (p_i,Ap_i)$, where $0 < \omega_i < 2$, then the natural error decreases at the step from i to i + 1. The factor ω_i is called *relaxation factor*. If all relaxation factors satisfy the condition $\delta < \omega_i < 2 - \delta$ for some $\delta \in (0,1)$, then for this process a convergence result similar to (15) holds. Note that the sequence $\{\omega_i\}$ influences the sequences $\{\alpha_i\}$, $\{\beta_i\}$ and $\{\gamma_i\}$ and also the convergence ratio.
A well-known DM using relaxation factors is the method of systematic overrelaxation (see Gauss-Seidel method). □

The remaining part of this section is devoted to a review of some basic algebraic properties of the Gauss-Seidel method, the Gauss-Southwell method, the gradient method and the conjugate gradient method.

1. *The Gauss-Seidel method*

The Gauss-Seidel method (as well as the Gauss-Southwell method) belongs to the class of so-called *coordinate descent methods*. In these methods each direction vector p_i is a unit vector. Therefore at each separate step only one component of x_i is changed. Moreover, (cf. theorem 2(iii)), every residual vector has one zero component. A subclass of the coordinate descent methods is the class of *cyclic coordinate descent methods*, to which the Gauss-Seidel method belongs.
In these methods the direction vectors p_i are cyclically chosen out of the set $\{e_1,\ldots,e_n\}$ of unit vectors. The objective function is sequentially minimized with respect to different components of x. There are a number of ways in which this concept can be developed into a complete algorithm.
In the Gauss-Seidel method one takes successively $p_0 := e_1$, $p_1 := e_2$, ..., $p_{n-1} := e_n$ and then repeats by taking $p_n := e_1$, $p_{n+1} := e_2$ and so on.

Consequently, for all $i \geq 0$

$$p_i := e_{k(i)} , \tag{19}$$

where

$$k(i) := 1 + i \bmod n . \tag{20}$$

The Gauss-Seidel method, applied to a definite system, converges average linearly. This result was first proven by Reich [49], using spectral radii. We here give a different proof using a compactness-argument.

Observe that for the Gauss-Seidel method, for every $i \geq 0$, the n steps from x_{i*n} to $x_{(i+1)*n}$ are exactly the same as the n steps from x_0' to x_n' if the process is started with $x_0' := x_{i*n}$.

Consequently, if we can prove that there exists an $L \in (0,1)$ such that $\|A^{\frac{1}{2}}(\hat{x} - x_n)\| \leq L\|A^{\frac{1}{2}}(\hat{x} - x_0)\|$ for any initial start vector $x_0 \neq \hat{x}$, then either the Gauss-Seidel process terminates (i.e. $x_i = \hat{x}$ for some $i \geq 0$) or $\|A^{\frac{1}{2}}(\hat{x} - x_{(i+1)*n})\| \leq L\|A^{\frac{1}{2}}(\hat{x} - x_{i*n})\|$ for all $i \geq 0$. In the latter case we have linear convergence on the average with an average convergence ratio no greater than $L^{1/n}$.

We now prove the existence of such an $L \in (0,1)$. From theorem 2 it follows

$$\|A^{\frac{1}{2}}(\hat{x} - x_n)\|^2 = \|A^{\frac{1}{2}}(\hat{x} - x_0)\|^2 - \sum_{i=0}^{n-1} (r_i, p_i)^2 / \|A^{\frac{1}{2}} p_i\|^2 . \tag{21}$$

Consequently, $\|A^{\frac{1}{2}}(\hat{x} - x_n)\| = \|A^{\frac{1}{2}}(\hat{x} - x_0)\|$ iff $(r_i, p_i) = 0$ for all $0 \leq i \leq n-1$. From the recurrence relation for r_i and the definition of a_i we obtain

$$r_i = r_0 - \sum_{\ell=0}^{i-1} a_\ell A p_\ell = r_0 , \tag{22}$$

if $(r_i, p_i) = 0$ for all $0 \leq i \leq n-1$. Hence, if $\|A^{\frac{1}{2}}(\hat{x} - x_n)\| = \|A^{\frac{1}{2}}(\hat{x} - x_0)\|$ then $(r_0, p_i) = 0$ for all $0 \leq i \leq n-1$ and since $p_0, \ldots, p_n$ are linearly independent this implies $r_0 = 0$. Therefore, if $r_0 \neq 0$ (i.e. $x_0 \neq \hat{x}$) then certainly $\|A^{\frac{1}{2}}(\hat{x} - x_n)\| < \|A^{\frac{1}{2}}(\hat{x} - x_0)\|$. Obviously, if

$$L := \max_{\|\hat{x} - x_0\| = 1} \frac{\|A^{\frac{1}{2}}(\hat{x} - x_n)\|}{\|A^{\frac{1}{2}}(\hat{x} - x_0)\|} , \tag{23}$$

then $L < 1$. Since $\hat{x} - x_n$ depends homogeneously and linearly on $\hat{x} - x_0$, (23) implies that $\|A^{\frac{1}{2}}(\hat{x} - x_n)\| \leq L\|A^{\frac{1}{2}}(\hat{x} - x_0)\|$ for all $x_0 \neq \hat{x}$. Note that the foregoing proof also holds under weaker conditions with respect to the direction vectors $\{p_i\}$.

A modification of the Gauss-Seidel method is the method of *systematic overrelaxation* (SOR). This DM generates the direction vectors p_i exactly in the same way, but instead of actually minimizing the objective function along that direction, which means computing a_i from (2), one computes $a_i := \omega(r_i,p_i) / (p_i,Ap_i)$. Here the relaxation factor ω is greater than 1 and it is introduced to improve the convergence ratio (cf. remark 7).

2. *The Gauss-Southwell method (GSM)*

This method belongs to the class of so-called *directed coordinate descent methods*. Instead of assigning the sequence of unit vectors a priori in carrying out line minimization, the coordinate to be changed is chosen such that it corresponds to the largest (in absolute value) component of the gradient vector.
Consequently, for all $i \geq 0$

$$p_i := e_{k(i)} , \tag{24}$$

where $k(i)$ satisfies

$$|(r_i,e_{k(i)})| \geq |(r_i,e_j)| , \tag{25}$$

for all $1 \leq j \leq n$.
Since $\|r_i\|^2 \leq n(r_i,e_{k(i)})^2$ we find, taking into account the definitions (16) and (17), that

$$\beta_i \leq \alpha_i = \frac{\|r_i\|\|e_{k(i)}\|}{|(r_i,e_{k(i)})|} \leq n^{\frac{1}{2}} , \tag{26}$$

and consequently, from lemma 6, $\gamma_i \geq (\beta_i \kappa^{\frac{1}{2}})^{-1} \geq (n\kappa)^{-\frac{1}{2}}$. Hence, from corollary 5, $(1 - (n\kappa^{-1}))^{\frac{1}{2}}$ is an upper bound for the convergence ratio. (Note that theorem 2(iii) yields $(r_{i+1},e_{k(i)}) = 0$ $(i \geq 0)$ and therefore, except for the first step, one has $\|r_i\|^2 \leq (n-1)(r_i,e_{k(i)})^2$, which leads to the sharper bound $(1 - ((n-1)\kappa)^{-1})^{\frac{1}{2}}$ for the convergence ratio.)

3. *The gradient method (steepest descent method) (GM)*

We recall that a DM searches for the minimum of the objective function $F(x) := (\hat{x} - x, A(\hat{x} - x))$. Obviously a good choice for moving towards $\hat{x}$ is to move in the (opposite) direction of the gradient, since this is the direction of steepest descent of the objective function. The gradient vector of F at x_i (cf. (13)) equals

$$(27) \qquad -2A(\hat{x} - x_i) = -2(b - Ax_i) = -2r_i \ .$$

The gradient method is based on this idea; as successive search directions one chooses the successive directions of the residual vector. This means, for all $i \geq 0$

$$(28) \qquad p_i := r_i \ .$$

Consequently, according to definitions (16) and (17),

$$(29) \qquad \beta_i \leq \alpha_i = \frac{\|r_i\|^2}{(r_i, r_i)} = 1 \ ,$$

and the convergence ratio is no greater than $(1 - 1/\kappa)^{\frac{1}{2}}$. However, directly estimating γ_i gives a sharper upper bound, for according to the definition of γ_i we obtain from the Kantorovich inequality (cf. section 1.2):

$$(30) \qquad \gamma_i^2 = \frac{\|r_i\|^4}{\|A^{-\frac{1}{2}} r_i\|^2 \|A^{\frac{1}{2}} r_i\|^2} \geq \frac{4\kappa}{(\kappa + 1)^2} \ ,$$

and hence (see corollary 5) $(\kappa - 1) / (\kappa + 1)$ is another (sharper) upper bound for the convergence ratio.

4. *The conjugate gradient method (CGM)*

This method belongs to the class of so-called *conjugate direction methods*. We review some characteristics of these methods (see Hestenes and Stiefel [52] and also Hestenes [80]. First a definition is given.

DEFINITION 8. *Let* A *be a symmetric* $n \times n$ *matrix, then the vectors* $x, y \in \mathbb{R}^n$ *are said to be* A*-orthogonal, or conjugate with respect to* A, *if* $(x, Ay) = 0$.

Note that, if A is positive definite, mutually conjugate nonzero vectors are linearly independent. Consequently the maximum number of mutually conjugate nonzero vectors then equals n.

A *conjugate direction method* is a DM in which the vectors $p_0, p_1, \ldots, p_{n-1}$ are mutually conjugate nonzero vectors. It causes no problem in having no more than n such vectors available, since after at most n steps the solution $\hat{x}$ is obtained. This property follows from the following consideration.

Let m be the smallest integer such that $\hat{x} - x_0$ is in the subspace spanned by $p_0, \ldots, p_{m-1}$. Clearly $m \leq n$, since the conjugate vectors are linearly independent. Furthermore,

$$\hat{x} - x_0 = \sum_{i=0}^{m-1} a_i' p_i , \tag{31}$$

where

$$a_i' := \frac{(A(\hat{x} - x_0), p_i)}{(p_i, Ap_i)} = \frac{(r_0, p_i)}{(p_i, Ap_i)} . \tag{32}$$

From the recurrence relation (5) for residual vectors we obtain

$$(r_j, p_i) = (r_{j-1}, p_i) - a_{j-1}(p_{j-1}, Ap_i) = (r_{j-1}, p_i) \quad (j-1 \neq i) . \tag{33}$$

Hence, $(r_i, p_i) = (r_0, p_i)$ and $a_i = a_i'$.
Since

$$x_m = x_0 + \sum_{i=0}^{m-1} a_i p_i$$

it follows that $x_m = \hat{x}$ and the algorithm ends after m steps.

For every DM one has for $i \geq 0$

$$A^{-1} r_{i+1} = \hat{x} - x_{i+1} = \hat{x} - x_0 - \sum_{k=0}^{i} a_k p_k . \tag{34}$$

Consequently, for a conjugate direction method it follows from (31) that

$$A^{-1} r_{i+1} = \sum_{k=i+1}^{m} a_k p_k \tag{35}$$

and hence, for any $j < i + 1$

(36) $$(r_{i+1},p_j) = (A^{-1} r_{i+1}, Ap_j) = 0 .$$

This means (compare remark 3) that the gradient vector of the objective function $F(x)$ at point x_{i+1} is orthogonal to all previous direction vectors $p_0,\ldots,p_i$. This establishes the second important algebraic property of conjugate direction methods, namely that x_{i+1} not only minimizes $F(x)$ along the line $x = x_i + ap_i$ (see theorem 2) but on the whole affine set passing through x_0 and spanned by $p_0,p_1,\ldots,p_i$ (cf. section 4.1).

The conjugate gradient method is the conjugate direction method that is obtained by constructing the successive directions by A-orthogonalization of the successive gradients, acquired as the process proceeds. The first step is identical to a gradient method step ($p_0 = r_0$). In each of the next steps one determines the (opposite) gradient vector (i.e. the residual vector) and adds to it a linear combination of the previous direction vectors in such a way that this new direction vector is A-orthogonal to the previous one. Proceeding in this way it happens that r_i ($i \geq 2$) is automatically A-orthogonal to $p_0,\ldots,p_{i-2}$. Hence, p_i ($i \geq 1$) can be determined from

(37) $$p_i = r_i + b_{i-1} p_{i-1} ,$$

where

(38) $$b_{i-1} := - \frac{(r_i,Ap_{i-1})}{(p_{i-1},Ap_{i-1})} .$$

From the definition of b_{i-1} it follows immediately that $(p_i,Ap_{i-1}) = 0$. From the definition of p_i and theorem 2(iii) it follows that for $i \geq 1$

(39) $$(r_i,p_i) = (r_i,r_i) + b_{i-1}(r_i,p_{i-1}) = (r_i,r_i) .$$

It is obvious that $(r_0,p_0) = (r_0,r_0)$.
From the definition of p_i it also follows that for $i \geq 1$

$$A^{\frac{1}{2}} p_i - b_{i-1} A^{\frac{1}{2}} p_{i-1} = A^{\frac{1}{2}} r_i .$$

Taking squared norms at both sides and using A-orthogonality, we obtain Pythagoras' theorem

(40) $$\|A^{\frac{1}{2}} p_i\|^2 + \|b_{i-1} A^{\frac{1}{2}} p_{i-1}\|^2 = \|A^{\frac{1}{2}} r_i\|^2 \quad (i \geq 1) .$$

Consequently, for $i \geq 1$

$$\|A^{\frac{1}{2}} p_i\| \leq \|A^{\frac{1}{2}} r_i\| \quad . \tag{41}$$

This inequality trivially also holds for $i = 0$.
Thus for the conjugate gradient method we have according to definitions (16) and (17), for all $i \geq 0$,

$$\alpha_i = \frac{\|p_i\|}{\|r_i\|} \leq \frac{\|A^{-\frac{1}{2}}\| \|A^{\frac{1}{2}} p_i\|}{\|r_i\|} \leq \frac{\|A^{-\frac{1}{2}}\| \|A^{\frac{1}{2}} r_i\|}{\|r_i\|} \leq \kappa^{\frac{1}{2}} \tag{42}$$

and

$$\beta_i = \frac{\|A^{\frac{1}{2}} p_i\|}{\|A^{\frac{1}{2}}\| \|r_i\|} \leq \frac{\|A^{\frac{1}{2}} r_i\|}{\|A^{\frac{1}{2}}\| \|r_i\|} \leq 1 \quad . \tag{43}$$

Therefore, in view of corollary 5 and lemma 6, $(1 - 1/\kappa)^{\frac{1}{2}}$ is an upper bound for the convergence ratio of the step-wise linear convergence ratio of the natural error. Analogously to the gradient method case one finds a better bound by using the γ_i. One has

$$\gamma_i^2 \geq 4\kappa / (\kappa + 1)^2 \tag{44}$$

and hence the convergence ratio is no greater than $(\kappa - 1) / (\kappa + 1)$.

Since $p_i = r_i + b_{i-1} p_{i-1}$ and $(r_i, p_{i-1}) = 0$, we have

$$\|p_i\|^2 = \|r_i\|^2 + b_{i-1}^2 \|p_{i-1}\|^2 \quad . \tag{45}$$

From (39) it follows that

$$a_i = (r_i, r_i) / (p_i, Ap_i) \quad , \tag{46}$$

and hence a_i might as well be computed from this relation.
A simple alternative formula can also be derived for b_i. From (5), (37) and (39) we obtain for $i \geq 1$

$$(r_{i+1}, r_i) = (r_i, r_i) - a_i (r_i, Ap_i) =$$

$$= (r_i, r_i) - a_i ((p_i, Ap_i) - b_{i-1} (p_{i-1}, Ap_i)) = 0 \quad . \tag{47}$$

It is obvious from the first equality that $(r_{i+1}, r_i) = 0$ also holds for $i = 0$.

From (5) it follows that $Ap_i = a_i^{-1}(r_i - r_{i+1})$ $(i \geq 0)$ and consequently, together with (46) and (47) we obtain for $i \geq 0$

$$(48) \qquad (r_{i+1}, Ap_i) = a_i^{-1}((r_{i+1}, r_i) - (r_{i+1}, r_{i+1})) =$$

$$= -(r_{i+1}, r_{i+1})(p_i, Ap_i) / (r_i, r_i) ,$$

which implies

$$(49) \qquad b_i = (r_{i+1}, r_{i+1}) / (r_i, r_i) ;$$

hence b_i might as well be computed from this relation.
The conjugate gradient methods based on these alternative formulas for a_i and b_i are discussed in Chapter 5.

2.3. *The recursive residual descent method (RRDM)*

In the presence of rounding errors the algebraic properties of DM's mentioned in the previous section, are affected by these rounding errors. For instance, even if at a certain step the relation $r_i = b - Ax_i$ holds exactly, then, performing one more step using the recurrence relation $r_{i+1} = r_i - a_i Ap_i$ for floating point computation of r_{i+1}, this recursive residual will differ from the exact residual $b - Ax_{i+1}$. This is due to the fact that the rounding errors occurring during the computation of $x_{i+1} = x_i + a_i p_i$ and occurring during the computation of $r_{i+1} = r_i - a_i Ap_i$ are independent. From this it will follows that we have to distinguish between RRDM and TRDM. In this section we shall investigate the behavior of RRDM if computations are carried out using floating point arithmetic. For TRDM this investigation is carried out in section 4.

2.3.1. *The numerical convergence of* $\{r_i\}$

We recall that the RRDM, corresponding to a given sequence of arbitrary nonzero vectors $\{p_i\}$ consists of the following statements.

RRDM

Choose an initial point x_0;

$r_0 := b - Ax_0$; $i := 0$;

while $r_i \neq 0$ do

begin

(1) $a_i := (r_i, p_i) / (p_i, Ap_i)$;

(2) $x_{i+1} := x_i + a_i p_i$;

(3) $r_{i+1} := r_i - a_i Ap_i$;

$i := i + 1$

end.

We observe that the sequence $\{r_i\}$ can be computed without computing the sequence $\{x_i\}$. Therefore, in this subsection we first analyse one step of RRDM, disregarding the computation of x_{i+1}, and next we add the computation of x_{i+1} to our considerations in subsection 2.3.2.

The following round-off error analysis is performed under the assumptions of section 1.3. The vectors $\{p_i\}$ corresponding to RRDM are supposed to be arbitrary nonzero machine vectors. The constants C_1 and C_2 refer to the constants corresponding to matrix by vector product computations and inner product computations as described in section 1.3. The capital characters D, E, F and G, appearing in the error analysis, will always refer to round-off matrices describing particular computations as mentioned in section 1.3. By a_i, r_i, r_{i+1} and p_i we always indicate the numbers and vectors as they are computed and stored by RRDM. For clearness' sake, (r_i, p_i) is the exact euclidean inner product of the stored vectors r_i and p_i, whereas $fl((r_i, p_i))$ denotes the computed value of this inner product. In the formulation of the lemmas and theorems we shall not always mention the restriction that r_i, p_i and (r_i, p_i) are assumed to be nonzero during the computations. Throughout the error analysis we use the O-symbol defined in section 1.5 and we neglect the possibility of overflow and underflow.

We are now ready to prove an analogue of theorem 2.2.2 in the presence of round-off. It assesses the influence of round-off on relation (i) of theorem 2.2.2.

THEOREM 1. *Let* r_i, p_i *be two nonzero machine vectors and let* r_{i+1} *be computed from one step RRDM, based on these vectors. Let*

$$\gamma_i := |(r_i,p_i)| \,/\, (\|A^{-\frac{1}{2}} r_i\| \|A^{\frac{1}{2}} p_i\|) \ . \tag{4}$$

Then we have

$$\frac{\|A^{-\frac{1}{2}} r_{i+1}\|^2}{\|A^{-\frac{1}{2}} r_i\|^2} = 1 - \gamma_i^2 + \eta_{i+1} \ , \tag{5}$$

where

$$|\eta_{i+1}| \le 4\varepsilon\kappa^{\frac{1}{2}}(3 + C_1\kappa^{\frac{1}{2}})(1 + o(1)) + \varepsilon C_2 \kappa^{\frac{1}{2}} o(1) \tag{6}$$

under the restriction

$$\varepsilon\kappa^{\frac{1}{2}}(1 + C_2 + C_1\kappa^{\frac{1}{2}}) \to 0 \ . \tag{7}$$

PROOF. We first consider the computation of a_i from (1).

$$\mathrm{fl}((r_i,p_i)) = ((I + D_i')r_i,p_i) = (r_i,p_i) + \tau_i \ , \tag{8}$$

$$|\tau_i| = |(D_i' r_i,p_i)| \le \varepsilon C_2 \|r_i\| \|p_i\| \ . \tag{9}$$

Further we have

$$\mathrm{fl}((p_i,Ap_i)) = ((I + D_i'')p_i,(A + E_i)p_i) = (p_i,Ap_i) + \xi_i \ , \tag{10}$$

$$\begin{aligned} |\xi_i| &= |(D_i'' p_i,Ap_i) + ((I + D_i'')p_i,E_i p_i)| \le \\ &\le \varepsilon C_2 \|p_i\| \|Ap_i\| + \varepsilon C_1 (1 + \varepsilon C_2)\|A\| \|p_i\|^2 \le \\ &\le \varepsilon C_2 \|p_i\| \|Ap_i\| + \varepsilon C_1 \|A\| \|p_i\|^2 (1 + o(1)) \ , \quad [\varepsilon C_2 \to 0] \ . \end{aligned} \tag{11}$$

This yields

$$a_i = \mathrm{fl}\left(\frac{\mathrm{fl}((r_i,p_i))}{\mathrm{fl}((p_i,Ap_i))}\right) = \frac{(r_i,p_i) + \tau_i}{(p_i,Ap_i) + \xi_i}(1 + \varepsilon_i) \ , \tag{12}$$

$$|\varepsilon_i| \le \varepsilon \ . \tag{13}$$

Hence,

$$a_i = \tilde{a}_i + \delta a_i' \tag{14}$$

where

$$\bar{a}_i := (r_i,p_i) / (p_i,Ap_i) \ , \tag{15}$$

$$\delta a_i' := \frac{\{\tau_i - \xi_i(r_i,p_i) / \|A^{\frac{1}{2}} p_i\|^2 + \varepsilon_i(r_i,p_i) + \tau_i\varepsilon_i\}}{\{\|A^{\frac{1}{2}} p_i\|^2(1 + \xi_i / \|A^{\frac{1}{2}} p_i\|^2)\}} \ . \tag{16}$$

From (11) we obtain

$$|\xi_i| / \|A^{\frac{1}{2}} p_i\|^2 \le \varepsilon C_2\kappa^{\frac{1}{2}} + \varepsilon C_1\kappa(1 + o(1)) \qquad [\varepsilon C_2 \to 0] \ . \tag{17}$$

Consequently, from (9), (11), (13), (16) and (17),

$$|\delta a_i'| \le \frac{\{\varepsilon C_2\|r_i\|\|p_i\| + (\varepsilon C_2\kappa^{\frac{1}{2}} + \varepsilon C_1\kappa)|(r_i,p_i)| + \varepsilon|(r_i,p_i)|\}}{\{\|A^{\frac{1}{2}} p_i\|^2(1 + o(1))\}} \tag{18}$$

under the restriction

$$\varepsilon(1 + C_2\kappa^{\frac{1}{2}} + C_1\kappa) \to 0 \ . \tag{19}$$

Since

$$|(r_i,p_i)| = |(A^{-\frac{1}{2}} r_i \ , A^{\frac{1}{2}} p_i)| \le \|A^{-\frac{1}{2}} r_i\|\|A^{\frac{1}{2}} p_i\| \tag{20}$$

it follows from (18) that

$$|\delta a_i'| \le (2\varepsilon C_2\kappa^{\frac{1}{2}} + \varepsilon C_1\kappa + \varepsilon)\|A^{-\frac{1}{2}} r_i\| / \|A^{\frac{1}{2}} p_i\|(1 + o(1)) \ , \tag{21}$$

under restriction (19), and in particular that

$$|\delta a_i'| \le \|A^{-\frac{1}{2}} r_i\| / \|A^{\frac{1}{2}} p_i\| \, o(1) \tag{22}$$

under restriction (19).

For the computation of r_{i+1} we have

$$r_{i+1} = (I + F_i'')(r_i - (I + F_i')a_i(A + E_i)p_i) \ , \tag{23}$$

or

$$r_{i+1} = r_i - a_iAp_i + \delta r_{i+1}' \ , \tag{24}$$

with

$$\delta r_{i+1}' = F_i''r_i - a_i(V_i'Ap_i + V_i''p_i) \ , \tag{25}$$

(26) $V_i' := F_i' + F_i''(I+F_i') \ , \quad \|V_i'\| \le 2\varepsilon(1+o(1)) \ , \qquad [\varepsilon \to 0] \ ,$

(27) $V_i'' := (I+F_i'')(I+F_i')E_i \ , \ \|V_i''\| \le \varepsilon C_1 \|A\|(1+o(1)) \ , \quad [\varepsilon \to 0] \ .$

A substitution of (14) in (24) shows that also

(28) $r_{i+1} = r_i - \hat{a}_i A p_i + \delta r_{i+1}'' \ ,$

where

(29) $\delta r_{i+1}'' = F_i'' r_i - (\delta a_i' A p_i + J_i' A p_i + J_i'' p_i) \ ,$

(30) $J_i' := a_i V_i' \ , \quad J_i'' := a_i V_i'' \ .$

From (20) it follows that

(31) $|\hat{a}_i| \le \|A^{-\frac{1}{2}} r_i\| / \|A^{\frac{1}{2}} p_i\| \ .$

Together with (14) and (22) this yields

(32) $|a_i| \le |\hat{a}_i| + |\delta a_i'| \le \|A^{-\frac{1}{2}} r_i\| / \|A^{\frac{1}{2}} p_i\| (1+o(1)) \ ,$

under restriction (19).
Consequently,

(33) $\|J_i'\| \le |a_i| \|V_i'\| \le 2\varepsilon \|A^{-\frac{1}{2}} r_i\| / \|A^{\frac{1}{2}} p_i\| (1+o(1)) \ ,$

(34) $\|J_i''\| \le |a_i| \|V_i''\| \le \varepsilon C_1 \|A\| \|A^{-\frac{1}{2}} r_i\| / \|A^{\frac{1}{2}} p_i\| (1+o(1)) \ ,$

under restriction (19).
We proceed by expressing η_{i+1} in terms of $\delta r_{i+1}''$. From (28) it follows that

(35) $A^{-\frac{1}{2}} r_{i+1} = A^{-\frac{1}{2}}(r_i - \hat{a}_i A p_i) + A^{-\frac{1}{2}} \delta r_{i+1}'' \ ,$

and, by taking squared norms of both sides of the equality, we obtain

(36) $\|A^{-\frac{1}{2}} r_{i+1}\|^2 = \|A^{-\frac{1}{2}}(r_i - \hat{a}_i A p_i)\|^2 + 2(A^{-1} r_i - \hat{a}_i p_i, \delta r_{i+1}'') + \|A^{-\frac{1}{2}} \delta r_{i+1}''\|^2 \ .$

From the definition of $\hat{a}_i$ we get (compare (2.2.11))

$$(37)\qquad \|A^{-\frac{1}{2}} r_{i+1}\|^2 = \|A^{-\frac{1}{2}} r_i\|^2 / \|A^{\frac{1}{2}} p_i\|^2 +$$

$$+ 2(A^{-1} r_i - \hat{a}_i p_i, \delta r''_{i+1}) + \|A^{-\frac{1}{2}} \delta r''_{i+1}\|^2 ,$$

which leads to the basic formula

$$(38)\qquad \frac{\|A^{-\frac{1}{2}} r_{i+1}\|^2}{\|A^{-\frac{1}{2}} r_i\|^2} = 1 - \gamma_i^2 + \eta_{i+1} ,$$

where γ_i is defined by (4) and

$$(39)\qquad \eta_{i+1} := \{2(A^{-1} r_i - \hat{a}_i p_i, \delta r''_{i+1}) + \|A^{-\frac{1}{2}} \delta r''_{i+1}\|^2\} / \|A^{-\frac{1}{2}} r_i\|^2 .$$

It remains to be proved that η_{i+1} satisfies (6) under the restriction (7).

Note that $(A^{-1} r_i - \hat{a}_i p_i, Ap_i) = 0$ and therefore the term $\delta a_i' Ap_i$ in (29) cancels when evaluating the inner product in the numerator of formula (39).

Consequently, from (29), (31), (33) and (34) we obtain, evaluating term by term,

$$(40)\qquad |(A^{-1} r_i - \hat{a}_i p_i, \delta r''_{i+1})| / \|A^{-\frac{1}{2}} r_i\|^2 \le$$

$$\le \{\|A^{-1} r_i\| \|F''_i\| \|r_i\| + |\hat{a}_i| \|p_i\| \|F''_i\| \|r_i\| + \|A^{-1} r_i\| \|J'_i\| \|Ap_i\| +$$

$$+ |\hat{a}_i| \|p_i\| \|J'_i\| \|Ap_i\| + \|A^{-1} r_i\| \|J''_i\| \|p_i\| +$$

$$+ |\hat{a}_i| \|p_i\| \|J''_i\| \|p_i\| \} / \|A^{-\frac{1}{2}} r_i\|^2 \le$$

$$\le 2\|F''_i\| \kappa^{\frac{1}{2}} + 2(\|J'_i\| \kappa^{\frac{1}{2}} + \|J''_i\| \|A^{-1}\|) \|A^{\frac{1}{2}} p_i\| / \|A^{-\frac{1}{2}} r_i\| \le$$

$$\le (6\varepsilon\kappa^{\frac{1}{2}} + 2\varepsilon C_1)(1 + o(1)) ,$$

under restriction (19).

Remains to be estimated the second order term in (39). This estimate does not affect the numerical constants appearing in the first order terms and therefore we may estimate rather roughly as far as numerical constants are concerned.

Since $(a+b+c+d)^2 \le 4(a^2+b^2+c^2+d^2)$, from (21), (29), (31), (33) and (34) we find

$$(41)\qquad \|A^{-\frac{1}{2}}\,\delta r''_{i+1}\|^2 / \|A^{-\frac{1}{2}}\,r_i\|^2 \le$$

$$\le 4\{\|F''_i\|^2\|A^{-1}\|\|r_i\|^2 + (\delta a'_i)^2\|A^{\frac{1}{2}}\,p_i\|^2 + \|J'_i\|^2\|A^{-1}\|\|Ap_i\|^2 +$$

$$+ \|J''_i\|^2\|A^{-1}\|\|p_i\|^2\} / \|A^{-\frac{1}{2}}\,r_i\|^2 \le$$

$$\le 4\{\|F''_i\|^2\kappa + ((\delta a'_i)^2 + \|J'_i\|^2\kappa + \|J''_i\|^2\|A^{-1}\|^2)\|A^{\frac{1}{2}}\,p_i\|^2 / \|A^{-\frac{1}{2}}\,r_i\|^2\} \le$$

$$\le 4\{\varepsilon^2\kappa + 4(4\varepsilon^2C_2^2\kappa + \varepsilon^2C_1^2\kappa^2 + \varepsilon^2) + 4\varepsilon^2\kappa + \varepsilon^2C_1^2\kappa^2\}(1+o(1)) ,$$

under restriction (19).
So, finally we obtain from (39), (40) and (41)

$$|\eta_{i+1}| \le 4\varepsilon\kappa^{\frac{1}{2}}(3+C_1\kappa^{\frac{1}{2}})(1+o(1)) + 4\varepsilon^2(4+5\kappa+5C_1^2\kappa^2+16C_2^2\kappa)(1+o(1)) ,$$

under restriction (19). As this inequality can be written in the more compact form (6), under the restriction (7), we have proved theorem 1. □

We note that the constant C_2 does not show up in the first order part of estimate (6). In the error analysis it only appears in the absolute error $\delta a'_i$ occurring at the computation of $\hat{a}_i$. The objective function $F(x_i + ap_i)$ is quadratic in a and hence, if we are at a distance δ from the point at which this function attains its maximum, the function value differs by an amount of the order δ^2 from the function value in that minimal point. Consequently, $\delta a'_i$ does not appear in (40), which explains the absence of C_2. Formulas (6) and (7), however, show that $\varepsilon C_2\kappa^{\frac{1}{2}}$ has to be small in order to have η_{i+1} small. A first order round-off error analysis would not have given this information.

REMARK 2. Theorem 1 can also be written in a form more closely related to (ii) of theorem 2.2.2. We have

$$\gamma_i^2 = (r_i,p_i)^2 / (\|A^{-\frac{1}{2}}\,r_i\|\|A^{\frac{1}{2}}\,p_i\|)^2 = \|\hat{a}_i\,A^{\frac{1}{2}}p_i\|^2 / \|A^{-\frac{1}{2}}\,r_i\|^2 ,$$

where, according to (15), $\hat{a}_i := (r_i,p_i) / (p_i,Ap_i)$.
Consequently, (5) can be written as

$$\|A^{-\frac{1}{2}} r_{i+1}\|^2 + \|\tilde{a}_i A^{\frac{1}{2}} p_i\|^2 = (1+\eta_{i+1})\|A^{-\frac{1}{2}} r_i\|^2 .$$

It follows in particular that, under the restriction (7),

$$\|A^{-\frac{1}{2}} r_{i+1}\| \leq (1+o(1))\|A^{-\frac{1}{2}} r_i\| ,$$

$$\|\tilde{a}_i A^{\frac{1}{2}} p_i\| \leq (1+o(1))\|A^{-\frac{1}{2}} r_i\| .$$

Using (14) and (22) this yields, under the restriction (7),

$$\|a_i A^{\frac{1}{2}} p_i\| \leq (1+o(1))\|A^{-\frac{1}{2}} r_i\| .$$

Retracing the proof of theorem 1 and replacing all o-symbols by definite estimates involving explicit numerical constants, one can prove that $|\eta_{i+1}| \leq 7/40$ if

$$\varepsilon\kappa^{\frac{1}{2}}(3+C_2+C_1\kappa^{\frac{1}{2}}) \leq \frac{1}{40} .$$

Hence

$$\|A^{-\frac{1}{2}} r_{i+1}\| \leq (1+\tfrac{1}{10})\|A^{-\frac{1}{2}} r_i\| ,$$

$$\|\tilde{a}_i A^{\frac{1}{2}} p_i\| \leq (1+\tfrac{1}{10})\|A^{-\frac{1}{2}} r_i\| .$$

Furthermore, it follows that

$$\|a_i A^{\frac{1}{2}} p_i\| \leq (1+\tfrac{2}{10})\|A^{-\frac{1}{2}} r_i\| . \qquad \square$$

REMARK 3. It is obvious from (5) that $\|A^{-\frac{1}{2}} r_{i+1}\|^2 \leq (1+|\eta_{i+1}|)\|A^{-\frac{1}{2}} r_i\|^2$, which means that the natural error $\|A^{-\frac{1}{2}} r_{i+1}\|$ cannot increase by more than a factor $(1+|\eta_{i+1}|)^{\frac{1}{2}}$ at the step from i to i + 1. In view of (6) and (7), $|\eta_{i+1}|$ is small if $\varepsilon\kappa^{\frac{1}{2}}(1+C_2+C_1\kappa^{\frac{1}{2}})$ is small. Formula (5) also shows that the natural error certainly decreases if the condition $\gamma_i^2 > \eta_{i+1}$ is satisfied. From lemma 2.2.6 it follows that in terms of the parameter β_i defined by (2.2.17), this condition is certainly satisfied if $\beta_i^2|\eta_{i+1}|\kappa < 1$. Because of (6) and (7) this last condition is fulfilled if $\varepsilon\kappa^{3/2}\beta_i^2(1+C_2+C_1\kappa^{\frac{1}{2}})$ is small enough. $\square$

Our relation (5) is phrased in terms of an absolute error η_{i+1}. We may as well try to estimate the relative error ν_{i+1}, defined by the relation

$$(42)\qquad \frac{\|A^{-\frac{1}{2}} r_{i+1}\|^2}{\|A^{-\frac{1}{2}} r_i\|^2} = 1 - \gamma_i^2(1+\nu_{i+1}) \ .$$

Obviously, $\nu_{i+1} = -\eta_{i+1}/\gamma_i^2$. In the proof of the next theorem we do not estimate ν_{i+1} from this relation, but we estimate ν_{i+1} directly. In this way we obtain a weaker sufficient condition on β_i, C_1, C_2 and κ to guarantee the monotonicity of the natural error $\|A^{-\frac{1}{2}} r_i\|$.

THEOREM 4. *Let r_i, p_i be two nonzero machine vectors, for which $(r_i,p_i) \neq 0$, and let r_{i+1} be computed from one step RRDM, based on these vectors. Let γ_i be defined as in theorem 1 and let, according to (2.2.16) and (2.2.17)*

$$(43)\qquad \alpha_i := \|r_i\|\|p_i\| / |(r_i,p_i)|$$

and

$$(44)\qquad \beta_i := \|r_i\|\|A^{\frac{1}{2}} p_i\| / (\|A^{\frac{1}{2}}\| |(r_i,p_i)|) \ .$$

Then we have

$$(45)\qquad \frac{\|A^{-\frac{1}{2}} r_{i+1}\|^2}{\|A^{-\frac{1}{2}} r_i\|^2} = 1 - \gamma_i^2(1+\nu_{i+1}) \ ,$$

where

$$(46)\qquad |\nu_{i+1}| \leq 2\varepsilon\{\kappa^{\frac{1}{2}}(2+C_1\kappa^{\frac{1}{2}}) + \kappa\beta_i(2+\beta_i) + \alpha_i(1+C_1\kappa)\}(1+o(1)) + \\ + \varepsilon C_2(\kappa^{\frac{1}{2}} + \alpha_i)o(1) \ ,$$

under the restriction

$$(47)\qquad \varepsilon\{\kappa^{\frac{1}{2}}(1+C_2+C_1\kappa^{\frac{1}{2}}) + C_2\alpha_i\} \to 0 \ .$$

PROOF. In the proof of the previous theorem we estimated the absolute error $\delta a_i'$ occurring at the computation of $\hat{a}_i$. In view of (8) we see that the relative error in the computation of the inner product (r_i,p_i) is not necessarily bounded if $(r_i,p_i) \to 0$ and therefore we evaded expressions for relative errors. In the present proof we follow the lines of the proof of theorem 1, the only difference being the use of

an estimate for the relative error of $\hat{a}_i$, in terms of the parameters α_i and β_i.

From (8) and (9) one has

$$\text{(48)} \qquad \mathrm{fl}((r_i,p_i)) = (r_i,p_i)(1+\lambda_i) \ ,$$

$$\text{(49)} \qquad |\lambda_i| = |\tau_i / (r_i,p_i)| \le \varepsilon C_2 \alpha_i \ .$$

Formulas (10) and (17) yield

$$\text{(50)} \qquad \mathrm{fl}((p_i,Ap_i)) = (p_i,Ap_i)(1+\mu_i) \ ,$$

$$\text{(51)} \qquad |\mu_i| = |\xi_i| / \|A^{\frac{1}{2}} p_i\|^2 \le \varepsilon C_2 \kappa^{\frac{1}{2}} + \varepsilon C_1 \kappa (1+o(1)) \quad [\varepsilon C_2 \to 0] \ .$$

Consequently

$$\text{(52)} \qquad a_i = \frac{(r_i,p_i)\ (1+\lambda_i)}{(p_i,Ap_i)(1+\mu_i)} (1+\varepsilon_i) = \hat{a}_i (1+\delta a_i'') \ ,$$

where $\hat{a}_i$ is defined by (15) and

$$\text{(53)} \qquad |\delta a_i''| = (\lambda_i - \mu_i + \varepsilon_i + \lambda_i \varepsilon_i) / (1+\mu_i) \ .$$

Hence, from (49), (51) and (13)

$$\text{(54)} \qquad |\delta a_i''| \le (\varepsilon C_2 \alpha_i + \varepsilon C_2 \kappa^{\frac{1}{2}} + \varepsilon C_1 \kappa + \varepsilon)(1+o(1))$$

under the restriction (19), and consequently

$$\text{(55)} \qquad |\delta a_i''| = o(1) \ ,$$

under the restriction

$$\text{(56)} \qquad \varepsilon\{(1 + C_2 \kappa^{\frac{1}{2}} + C_1 \kappa) + C_2 \alpha_i\} \to 0 \ .$$

The vector $\delta r_{i+1}''$ in (29) can be written as

$$\text{(57)} \qquad \delta r_{i+1}'' = F_i'' r_i - \hat{a}_i (\delta a_i'' Ap_i + M_i' Ap_i + M_i'' p_i) \ ,$$

$$\text{(58)} \qquad M_i' := \hat{a}_i^{-1} J_i' = (1+\delta a_i'') \|V_i'\| \ , \quad \|M_i'\| \le 2\varepsilon(1+o(1)) \ ,$$

$$\text{(59)} \qquad M_i'' := \hat{a}_i^{-1} J_i'' = (1+\delta a_i'') \|V_i''\| \ , \quad \|M_i''\| \le \varepsilon C_1 \|A\| (1+o(1)) \ ,$$

under the restriction (56).

Instead of transforming (37) into (38), we may also write

$$\frac{\|A^{-\frac{1}{2}} r_{i+1}\|^2}{\|A^{-\frac{1}{2}} r_i\|^2} = 1 - \gamma_i^2(1+\nu_{i+1}) \quad , \tag{60}$$

where γ_i is defined as in theorem 1 and

$$\nu_{i+1} := \{2(A^{-1} r_i - \hat{a}_i p_i, \delta r''_{i+1}) + \|A^{-\frac{1}{2}} \delta r''_{i+1}\|^2\} / \{\hat{a}_i (r_i, p_i)\} \, . \tag{61}$$

Analogously to (40) we obtain, evaluating term by term,

$$\begin{aligned} |(A^{-1} r_i - \hat{a}_i p_i, \delta r''_{i+1})| / |\hat{a}_i (r_i, p_i)| \leq \\ \leq \|A^{-1} r_i\| \|F''_i\| \|r_i\| \|A^{\frac{1}{2}} p_i\|^2 / (r_i, p_i)^2 + \|p_i\| \|F''_i\| \|r_i\| / |(r_i, p_i)| + \\ + \|A^{-1} r_i\| \|M'_i\| \|Ap_i\| / |(r_i, p_i)| + \|p_i\| \|M'_i\| \|Ap_i\| / \|A^{\frac{1}{2}} p_i\|^2 + \\ + \|A^{-1} r_i\| \|M''_i\| \|p_i\| / |(r_i, p_i)| + \|p_i\| \|M''_i\| \|p_i\| / \|A^{\frac{1}{2}} p_i\|^2 \leq \\ \leq \|F''_i\| \kappa \beta_i^2 + \|F''_i\| \alpha_i + \|M'_i\| \kappa \beta_i + \|M'_i\| \kappa^{\frac{1}{2}} + \|M''_i\| \|A^{-1}\| \alpha_i + \|M''_i\| \|A^{-1}\| \leq \\ \leq \varepsilon \{\kappa^{\frac{1}{2}} (2 + C_1 \kappa^{\frac{1}{2}}) + \kappa \beta_i (2 + \beta_i) + \alpha_i (1 + C_1 \kappa)\} (1 + o(1)) \quad , \end{aligned} \tag{62}$$

under the restriction (56).
Analogously to (41) we find, evaluating term by term, that

$$\begin{aligned} \|A^{-\frac{1}{2}} \delta r''_{i+1}\|^2 / |\hat{a}_i (r_i, p_i)| \leq \\ \leq 4\{\|F''_i\|^2 \|A^{-1}\| \|r_i\|^2 \|A^{\frac{1}{2}} p_i\|^2 / (r_i, p_i)^2 + (\delta a''_i)^2 + \\ + \|M'_i\|^2 \|A^{-1}\| \|Ap_i\|^2 / \|A^{\frac{1}{2}} p_i\|^2 + \|M''_i\|^2 \|A^{-1}\| \|p_i\|^2\} / \|A^{\frac{1}{2}} p_i\|^2 \leq \\ \leq 4\{\|F''_i\|^2 \kappa \beta_i^2 + (\delta a''_i)^2 + \|M'_i\|^2 \kappa + \|M''_i\|^2 \|A^{-1}\|^2\} \leq \\ \leq 4\{\varepsilon^2 \kappa \beta_i^2 + 4(\varepsilon^2 C_2^2 \alpha_i^2 + \varepsilon^2 C_2^2 \kappa + \varepsilon^2 C_1^2 \kappa^2 + \varepsilon^2) + \\ + 4\varepsilon^2 \kappa + \varepsilon^2 C_1^2 \kappa^2\} (1 + o(1)) \quad , \end{aligned} \tag{63}$$

under the restriction (56).
So, finally, we obtain from (61), (62) and (63)

(64) $$|\nu_{i+1}| \leq 2\varepsilon\{\kappa^{\frac{1}{2}}(2+C_1\kappa^{\frac{1}{2}}) + \kappa\beta_i(2+\beta_i) + \alpha_i(1+C_1\kappa)\}(1+o(1)) +$$

$$+ 4\varepsilon^2\{(4+4\kappa+4C_2^2\kappa+5C_1^2\kappa^2) + \kappa\beta_i^2 + 4C_2^2\alpha_i^2\}(1+o(1)) ,$$

under the restriction (56). As this inequality can be written in the more compact form (46), under the restriction (47), we have proved theorem 4. □

REMARK 5. We have (cf. (11), (50) and (51))

$$fl((p_i,Ap_i)) = (p_i,Ap_i)(1+\mu_i) ,$$

$$|\mu_i| = |\xi_i| / \|A^{\frac{1}{2}} p_i\|^2 \leq \varepsilon C_2\kappa^{\frac{1}{2}} + (1+\varepsilon C_2)\varepsilon C_1\kappa .$$

Consequently, if $\varepsilon C_2\kappa^{\frac{1}{2}} + (1+\varepsilon C_2)\varepsilon C_1\kappa < 1$, then $fl((p_i,Ap_i)) \neq 0$ if $p_i \neq 0$. Hence in that case the DM performed using floating point arithmetic cannot break down because of a zero denominator, and the algorithm will only end if $r_i = 0$ (neglecting, of course, underflow). □

REMARK 6. We observe that just as in theorem 1 the constant C_2 does not show up in the first order part of (46). From formulas (46) and (47) we conclude that, as far as C_2 concerns, $\varepsilon C_2(\kappa^{\frac{1}{2}} + \alpha_i)$ has to be small in order to have ν_{i+1} small. □

REMARK 7. In the previous theorem we frequently used the parameters α_i and β_i for estimating the various rounding errors. Expressions involving (say) $\|r_i\|\|Ap_i\| / (\|A\|\,|(r_i,p_i)|)$ were estimated in terms of these parameters. For the gradient method and the conjugate gradient method the introduction of more parameters does not give stronger results, since no sharper direct bounds than in terms of α_i and β_i are available for these expressions.
The numerical behavior of these methods is of our main interest and therefore theorem 4 is formulated in terms of α_i and β_i only. □

From (46) and (47) and the fact that $\alpha_i \geq 1$ it follows that $|\nu_{i+1}| = o(1)$ under the restriction

(65) $$\varepsilon\{\kappa^{\frac{1}{2}}(1+C_2) + \kappa\beta_i(1+\beta_i) + \alpha_i(1+C_2+C_1\kappa)\} \to 0 .$$

Consequently, as a corollary of theorem 4 we obtain the analogue of corollary 2.2.5 if floating point arithmetic is used.

COROLLARY 8. *Let* $\{r_i\}$ *be computed by RRDM with an arbitrary initial machine vector* x_0 *and suppose there exist constants* $\alpha,\beta,\gamma > 0$ *such that for all* $i \geq 0$

$$\alpha_i := \|r_i\|\|p_i\| / |(r_i,p_i)| \leq \alpha \,, \tag{66}$$

$$\beta_i := \|r_i\|\|A^{\frac{1}{2}} p_i\| / (\|A^{\frac{1}{2}}\| |(r_i,p_i)|) \leq \beta \,, \tag{67}$$

$$\gamma_i := |(r_i,p_i)| / (\|A^{-\frac{1}{2}} r_i\|\|A^{\frac{1}{2}} p_i\|) \geq \gamma \,, \tag{68}$$

then we have for $i \geq 0$

$$\frac{\|A^{-\frac{1}{2}} r_{i+1}\|^2}{\|A^{-\frac{1}{2}} r_i\|^2} \leq 1 - \gamma^2(1+o(1)) \tag{69}$$

under the restriction

$$\varepsilon\{\kappa^{\frac{1}{2}}(1+C_2) + \kappa\beta(1+\beta) + \alpha(1+C_2+C_1\kappa)\} \to 0 \,. \tag{70}$$

□

REMARK 9. From (69) and (70) it is obvious that the natural error $\|A^{-\frac{1}{2}} r_i\|$, and consequently r_i, tends to zero if $\varepsilon\{\kappa^{\frac{1}{2}}(1+C_2) + \kappa\beta(1+\beta) + \alpha(1+C_2+C_1\kappa)\}$ is small enough. We realize that from a practical point of view this is not a very interesting conclusion since convergence of the recursively computed residual r_i has no direct practical implication. However, from an academical point of view it is a rather surprising result, since there are not many iterative processes, used in practice, generating sequences that tend to zero. □

REMARK 10. From lemma 2.2.6 it follows that if one of the three parameters α_i, β_i, γ_i is bounded, then the other two parameters are bounded (bounded in accordance with (66), (67) and (68)). However, as we saw in section 2.2 for the algebraic case, it sometimes is possible to obtain sharper bounds by estimating each parameter separately. □

REMARK 11. In corollary 8 it is assumed that all γ_i are bounded away from zero uniformly for all $i \geq 0$. For the Gauss-Southwell method, the gradient method and the conjugate gradient method this is algebraically the case (cf. section 2.2) and, as we shall see, this enables us to prove this boundedness even if round-off occurs. In most cyclic coordinate DM's, however, $\gamma_i = 0$ can occur with exact arithmetic and then we lack an algebraic base for proving the existence of a uniform positive lower bound in the presence of round-off. On the other hand, for most of these algorithms (like e.g. Gauss-Seidel), there exists a $k \geq n$ such that in every k subsequent nonoverlapping steps there is at least one step for which γ_i is bounded away from zero and this bound (γ say) is uniform in i. Thus for every k subsequent nonoverlapping steps one can apply theorem 4 for the steps where $\gamma_i \geq \gamma$ and apply theorem 1 for the remaining steps in this subsequence in order to obtain results on the decrement of the natural error after these k steps. □

Comparing (69) and (2.2.15) we see that the convergence ratio of the numerical process approaches the convergence ratio of the algebraic process as (70) tends to zero. However, we are not primarily interested in the fact that the numerical convergence ratio is close to the algebraic convergence ratio if (70) is small, but we want to know under which explicit conditions the natural error $\|A^{-\frac{1}{2}} r_i\|$ tends to zero. From theorem 4, as we said before, we conclude that the natural error decreases at the step from i to i + 1 iff $\nu_{n+1} > -1$ in (45). Apparently, from (46) and (47), $|\nu_{i+1}| < 1$ if (65) is small. Unfortunately, these formulas do not supply an explicit bound for (65) in order to guarantee $|\nu_{i+1}| < 1$ uniformly in i. Here we encounter a situation where the disadvantage of the O-notation (it does not yield explicit error bounds) emerges. On the other hand, as we said already in section 1.5, one can easily retrace the proof and replace all O-symbols by definite estimates involving explicit numerical constants. To strengthen this assertion we shall execute this procedure for the foregoing proof. The reason of doing it in particular in this case is that theorem 4 is one of our basic results.

We shall show that, under the assumption

$$\varepsilon\{\kappa^{\frac{1}{2}}(2 + C_2 + C_1\kappa^{\frac{1}{2}}) + \kappa\beta_i(2 + \beta_i) + \alpha_i(1 + C_2 + C_1\kappa)\} \leq \frac{1}{8}, \tag{71}$$

one certainly has $|\nu_{i+1}| \le 11/16$.
We shall follow the lines of the proof of theorem 4, replacing the O-symbols by numerical constants.

We first obtain

$$|\mu_i| \le \varepsilon C_2 \kappa^{\frac{1}{2}} + \varepsilon(1+\varepsilon C_2) C_1 \kappa \le \varepsilon C_2 \kappa^{\frac{1}{2}} + \frac{9}{8}\varepsilon C_1 \kappa < \tag{72}$$

$$< \frac{9}{8}(\varepsilon C_2 \kappa^{\frac{1}{2}} + \varepsilon C_1 \kappa < \frac{9}{64} .$$

Hence, instead of (54) we get

$$|\delta a_i''| \le \frac{64}{55}(|\lambda_i| + |\mu_i| + |\gamma_i| + |\lambda_i \gamma_i|) < \tag{73}$$

$$< \frac{64}{55}(\varepsilon C_2 \alpha_i + \frac{9}{8}(\varepsilon C_2 \kappa^{\frac{1}{2}} + \varepsilon C_1 \kappa) + \varepsilon + \varepsilon^2 C_2 \alpha_i) <$$

$$< \frac{72}{55}(\varepsilon C_2 \alpha_i + \varepsilon C_2 \kappa^{\frac{1}{2}} + \varepsilon C_1 \kappa + 2\varepsilon) < \frac{9}{55} .$$

Furthermore we find

$$\|V_i'\| \le 2\varepsilon + \varepsilon^2 < \frac{9}{4}\varepsilon \le \frac{9}{32} ,$$

$$\|V_i''\| \le (1 + \|V_i'\|)\|E_i\| < \frac{41}{32}\varepsilon C_1 \|A\| ,$$

$$\|M_i'\| \le (1 + |\delta a_i''|)\|V_i'\| < \frac{64}{55}\frac{9}{4}\varepsilon = \frac{144}{55}\varepsilon < 2(1+\frac{4}{10})\varepsilon ,$$

$$\|M_i''\| \le (1 + |\delta a_i''|)\|V_i'\| < \frac{64}{55}\frac{41}{32}\varepsilon C_1 \|A\| = \frac{82}{55}\varepsilon C_1 \|A\| <$$

$$< (1+\frac{5}{10})\varepsilon C_1 \|A\| .$$

Consequently,

$$|(A^{-1} r_i - \hat{a}_i p_i, \delta r_{i+1}'')| / |\hat{a}_i(r_i, p_i)| \le \tag{74}$$

$$\le \varepsilon\kappa\beta_i^2 + \varepsilon\alpha_i + 2\varepsilon\kappa\beta_i(1+\frac{4}{10}) + 2\varepsilon\kappa^{\frac{1}{2}}(1+\frac{4}{10}) +$$

$$+ \varepsilon C_1 \kappa \alpha_i (1+\frac{5}{10}) + \varepsilon C_1 \kappa (1+\frac{5}{10}) <$$

$$< \frac{15}{10}\varepsilon(\kappa^{\frac{1}{2}}(2 + C_1\kappa^{\frac{1}{2}}) + \kappa\beta_i(2+\beta_i) + \alpha_i(1 + C_1\kappa)) \le \frac{3}{16} ,$$

and also

(75) $$\|A^{-\frac{1}{2}}\,\delta r''_{i+1}\|^2 / |\hat{a}_i(r_i,p_i)| \le$$

$$\le 4\{\varepsilon^2\kappa\beta_i^2 + (\tfrac{72}{55})^2(\varepsilon C_2\alpha_i + \varepsilon C_2\kappa^{\frac{1}{2}} + \varepsilon C_1\kappa + 2\varepsilon)^2 +$$

$$+ (\tfrac{144}{55})^2\,\varepsilon^2\kappa + (\tfrac{82}{55})^2\,\varepsilon^2 C_1^2\kappa^2\} \le$$

$$\le 4 * (\tfrac{82}{55})^2\{2\varepsilon^2\kappa\beta_i^2 + (\varepsilon C_2\alpha_i + \varepsilon C_2\kappa^{\frac{1}{2}} + \varepsilon C_1\kappa + 2\varepsilon)^2 +$$

$$+ (2\varepsilon\kappa^{\frac{1}{2}} + \varepsilon C_1\kappa)^2\} \le$$

$$\le 8 * (\tfrac{82}{55})^2\,\varepsilon^2(\kappa^{\frac{1}{2}}(2 + C_2 + C_1) + \kappa^{\frac{1}{2}}\beta_i + \alpha_i C_2)^2 \le$$

$$\le 8 * (\tfrac{82}{55})^2 * (\tfrac{1}{64}) < \tfrac{5}{16}\,.$$

Hence

$$|\nu_{i+1}| \le 2 * (\tfrac{3}{16}) + \tfrac{5}{16} = \tfrac{11}{16}\,.$$

Thus, as a more explicit version of theorem 4 we obtain

PROPOSITION 12. *Let* $\{r_i\}$ *be computed by RRDM with an arbitrary initial machine vector* x_0 *and let* α_i, β_i, γ_i *denote the parameters of theorem 4. Furthermore, let*

(76) $$\varepsilon\{\kappa^{\frac{1}{2}}(1 + C_2 + C_1\kappa^{\frac{1}{2}}) + \kappa\beta_i(1 + \beta_i) + \alpha_i(1 + C_2 + C_1\kappa)\} \le \tfrac{1}{8}\,,$$

then we have

(77) $$\frac{\|A^{-\frac{1}{2}}\,r_{i+1}\|^2}{\|A^{-\frac{1}{2}}\,r_i\|^2} \le 1 - \tfrac{5}{16}\,\gamma_i^2\,.$$ □

The restriction (76) is quite arbitrary and the bound $|\nu_{i+1}| \le \frac{11}{16}$ is deduced by a rather rough estimate; it can easily be improved.

Of course, the foregoing calculations also yield a more explicit version of corollary 8, i.e., if $\alpha_i \le \alpha$, $\beta_i \le \beta$, $\gamma_i \ge \gamma$ and if (76) holds when replacing α_i, β_i by α, β, then $\|A^{-\frac{1}{2}}\,r_{i+1}\|^2 / \|A^{-\frac{1}{2}}\,r_i\|^2 \le 1 - (5/16)\gamma^2$ and consequently $r_i \to 0$ $(i \to \infty)$.

REMARK 13. From lemma 2.2.6 it follows that

$$(78)\qquad \varepsilon\{\kappa^{\frac{1}{2}}(1+C_2+C_1\kappa^{\frac{1}{2}}) + \kappa\beta_i(1+\beta_i) + \alpha_i(1+C_2+C_1\kappa)\} <$$

$$< 4\varepsilon\kappa^{3/2}\beta_i^2(1+C_2+C_1\kappa^{\frac{1}{2}}) ,$$

and hence (cf. remark 3 and remark 9) the sufficient condition for convergence of r_i to zero, following from theorem 4, is weaker (and as we shall see in most cases essentially weaker) than the sufficient condition for convergence of r_i to zero, following from theorem 1. □

We conclude this section with an examination of the infinite sums $\Sigma_{\ell=0}^{\infty} \|A^{-\frac{1}{2}} r_\ell\|$ and $\Sigma_{\ell=0}^{\infty} \|a_\ell A^{\frac{1}{2}} p_\ell\|$. The results will be used in section 2.3.2.

From corollary 2.2.5 it follows that algebraically, if all $\gamma_i \geq \gamma$,

$$(79)\qquad \sum_{\ell=0}^{\infty} \|A^{-\frac{1}{2}} r_\ell\|^2 \leq \|A^{-\frac{1}{2}} r_0\|^2 \sum_{\ell=0}^{\infty} (1-\gamma^2)^\ell = \gamma^{-2}\|A^{-\frac{1}{2}} r_0\|^2 .$$

The analogue in the presence of round-off is expressed in (80).

LEMMA 14. *Let $\{r_i\}$ be computed by RRDM with an arbitrary initial machine vector x_0 and let α, β, γ denote the bounds of corollary 8. Then we have*

$$(80)\qquad \sum_{\ell=0}^{\infty} \|A^{-\frac{1}{2}} r_\ell\|^2 \leq \gamma^{-2}(1+o(1))\|A^{-\frac{1}{2}} r_0\|^2 ,$$

under the restriction

$$(81)\qquad \varepsilon\{\kappa^{\frac{1}{2}}(1+C_2) + \kappa\beta(1+\beta) + \alpha(1+C_2+C_1\kappa)\} \to 0 .$$

PROOF. This is a direct consequence of corollary 8. □

Algebraically (cf. theorem 2.2.2) we have $\|a_i A^{\frac{1}{2}} p_i\|^2 = \|A^{-\frac{1}{2}} r_i\|^2 + - \|A^{-\frac{1}{2}} r_{i+1}\|^2$, and consequently

$$(82)\qquad \sum_{\ell=0}^{i-1} \|a_\ell A^{\frac{1}{2}} p_\ell\|^2 = \|A^{-\frac{1}{2}} r_0\|^2 - \|A^{-\frac{1}{2}} r_i\|^2 .$$

Hence, since $\|A^{-\frac{1}{2}} r_i\| \to 0$ $(i \to \infty)$ under the conditions of corollary 2.2.5,

we obtain

$$\sum_{\ell=0}^{\infty} \|a_\ell A^{\frac{1}{2}} p_\ell\|^2 = \|A^{-\frac{1}{2}} r_0\|^2 . \tag{83}$$

The analogue in the presence of round-off is expressed in (84).

LEMMA 15. *Let* $\{r_i\}$ *be computed by RRDM with an arbitrary initial machine vector* x_0 *and let* α, β *denote the bounds of corollary 8. Then we have*

$$\sum_{\ell=0}^{\infty} \|a_\ell A^{\frac{1}{2}} p_\ell\|^2 = \|A^{-\frac{1}{2}} r_0\|^2 + o(1) \sum_{\ell=0}^{\infty} \|A^{-\frac{1}{2}} r_\ell\|^2 , \tag{84}$$

under the restriction (81).

PROOF. From (14), (22) and remark 2 it follows that

$$\begin{aligned} \|a_i A^{\frac{1}{2}} p_i\|^2 &= \|\bar{a}_i A^{\frac{1}{2}} p_i\|^2 + 2|\delta a_i'| \|\bar{a}_i A^{\frac{1}{2}} p_i\| + (\delta a_i')^2 \|A^{\frac{1}{2}} p_i\|^2 = \\ &= \|\bar{a}_i A^{\frac{1}{2}} p_i\|^2 + o(1)\|A^{-\frac{1}{2}} r_i\|^2 = \\ &= (1+o(1))\|A^{-\frac{1}{2}} r_i\|^2 - \|A^{-\frac{1}{2}} r_{i+1}\|^2 \end{aligned} \tag{85}$$

under the restriction $\varepsilon\kappa^{\frac{1}{2}}(1+C_2+C_1\kappa^{\frac{1}{2}}) \to 0$, and consequently also under restriction (81).

We obtain by summation

$$\sum_{\ell=0}^{\infty} \|a_\ell A^{\frac{1}{2}} p_\ell\|^2 = \|A^{-\frac{1}{2}} r_0\|^2 + o(1) \sum_{\ell=0}^{\infty} \|A^{-\frac{1}{2}} r_\ell\|^2 , \tag{86}$$

which proves (84). □

REMARK 16. From (28) we obtain

$$(r_{i+1}, p_i) = (r_i, p_i) - \bar{a}_i (p_i, Ap_i) + (\delta r''_{i+1}, p_i) = (\delta r''_{i+1}, p_i) . \tag{87}$$

From (21), (29), (33) and (34) we get

$$|(\delta r''_{i+1}, p_i)| \le \|A^{-\frac{1}{2}} \delta r''_{i+1}\| \|A^{\frac{1}{2}} p_i\| \le \tag{88}$$

$$\leq \{\|A^{-\frac{1}{2}}\|\|F_i''\|\|r_i\| + |\delta a_i'|\|A^{\frac{1}{2}} p_i\| + \|A^{-\frac{1}{2}}\|\|J_i'\|\|Ap_i\| +$$

$$+ \|A^{-\frac{1}{2}}\|\|J_i''\|\|p_i\|\}\|A^{\frac{1}{2}} p_i\| \leq$$

$$\leq \{\varepsilon\kappa^{\frac{1}{2}} + (2\varepsilon C_2\kappa^{\frac{1}{2}} + \varepsilon C_1\kappa + \varepsilon) + 2\varepsilon\kappa^{\frac{1}{2}} + \varepsilon C_1\kappa)(1 + o(1))\}\|A^{-\frac{1}{2}} r_i\|\|A^{\frac{1}{2}} p_i\| =$$

$$= o(1)\|A^{-\frac{1}{2}} r_i\|\|A^{\frac{1}{2}} p_i\| ,$$

under the restriction $\varepsilon\kappa^{\frac{1}{2}}(1 + C_2 + C_1\kappa^{\frac{1}{2}}) \to 0$.
Hence, under the same restriction, we have

$$\psi_i := |(A^{-\frac{1}{2}} r_{i+1}, A^{\frac{1}{2}} p_i)| / (\|A^{-\frac{1}{2}} r_{i+1}\|\|A^{\frac{1}{2}} p_i\|) = \tag{89}$$

$$= o(1)\|A^{-\frac{1}{2}} r_i\| / \|A^{-\frac{1}{2}} r_{i+1}\| .$$

We note that $\cos^{-1} \psi_i$ is the angle between the vectors $A^{-\frac{1}{2}} r_{i+1}$ and $A^{\frac{1}{2}} p_i$. From theorem 2.2.2(iii) we know that algebraically these vectors are orthogonal. From (89) we see that in the floating point case this orthogonality can be seriously disturbed if $\|A^{-\frac{1}{2}} r_i\| / \|A^{-\frac{1}{2}} r_{i+1}\|$ is large. Stated differently, the vectors $A^{-\frac{1}{2}} r_{i+1}$ and $A^{\frac{1}{2}} p_i$ are approximately orthogonal, unless $\|A^{-\frac{1}{2}} r_{i+1}\| \ll \|A^{-\frac{1}{2}} r_i\|$. □

REMARK 17. Substitution of result (80) into (84) yields

$$\sum_{\ell=0}^{\infty} \|a_\ell A^{\frac{1}{2}} p_\ell\|^2 = (1 + \gamma^{-2} o(1))\|A^{-\frac{1}{2}} r_0\|^2 , \tag{90}$$

under the restriction (81).
Of course, this implies that $a_i A^{\frac{1}{2}} p_i \to 0$ $(i \to \infty)$. □

2.3.2. *The numerical convergence of* $\{x_i\}$

In the previous subsection we disregarded the computation of x_i. However, the step-wise linear convergence of the computed recursive residuals to zero does not guarantee the convergence of x_i to $\hat{x}$, since in the presence of round-off the r_i will wander away from the true residuals $\hat{r}_i := b - Ax_i$ as we mentioned already at the beginning of this section.

The vector x_{i+1} is computed from the relation $x_{i+1} = x_i + a_i p_i$. Consequently, the error occurring at the computation of x_{i+1} can be of order $\varepsilon\|x_i\|$ at each step and this ultimately equals $\varepsilon\|\hat{x}\|$. Therefore, even if no round-off would occur at the computation of $r_{i+1} := r_i - a_i Ap_i$ the difference $\|\hat{r}_{i+1} - r_{i+1}\|$ may at each step increase by something of the order $\varepsilon\|A\|\|\hat{x}\|$. From this one can see that the assumption that the machine has strong arithmetic in the sense of Dekker [79] (see (1.3.1) and (1.3.2)), is not sufficient to guarantee even the uniform boundedness of $\hat{r}_i - r_i$ for all $i \geq 0$.

From his experiments Reid [71] found that $\hat{r}_i$ and r_i depart from each other rather slowly if the problem is well-conditioned. He showed that any errors that occur in the evaluation of a_i do not make a direct contribution to the difference between r_{i+1} and $\hat{r}_{i+1}$.

In the first part of our analysis we shall study the growth of the difference $\hat{r}_i - r_i$ as i increases, and next this result will be used to estimate the natural error $\|A^{\frac{1}{2}}(\hat{x} - x_i)\|$.

From the assumption that we have a machine with proper rounding arithmetic (cf. section 1.3), we shall arrive at the conclusion that the approximations x_i are uniformly bounded. These approximations are computed from the relation $x_{i+1} = x_i + a_i p_i$. Algebraically

$$\|a_i p_i\| \leq \|A^{-\frac{1}{2}}\|\|a_i A^{\frac{1}{2}} p_i\| \leq \|A^{-\frac{1}{2}}\|\|A^{-\frac{1}{2}} r_i\|$$

(cf. theorem 2.2.2) and hence $\{a_i p_i\}$ converges average linearly in the case of complete accuracy. From remark 2.3.1.2 it follows that this also holds in the presence of round-off. Hence we are in the situation as discussed in remark 1.3.4, indicating the uniform boundedness of $\{x_i\}$.

We conclude the numerical consideration of RRDM with a one-round-off error analysis (see section 1.3) which indicates the maximal magnitude of the true residual at the moment where in the numerical process $\|A^{\frac{1}{2}}(\hat{x} - x_{i+1})\| > \|A^{\frac{1}{2}}(\hat{x} - x_i)\|$ occurs.

We now first examine how much the true residual $\hat{r}_i := b - Ax_i$ and the computed recursive residual r_i can differ. It turns out that the sharpest result is obtained when using the natural norm $\|A^{-\frac{1}{2}}(\hat{r}_i - r_i)\|$.

LEMMA 1. *Let* $\{x_i\}$, $\{r_i\}$ *be computed by RRDM with an arbitrary initial machine vector* x_0 *and let* $\hat{r}_i := b - Ax_i$.
Then we have for all $i \geq 1$, *under the restriction* $\varepsilon \to 0$,

$$(1) \qquad \|A^{-\frac{1}{2}}(\hat{r}_i - r_i)\| \leq (1+\varepsilon\kappa^{\frac{1}{2}})^i\{\|A^{-\frac{1}{2}}(\hat{r}_0 - r_0)\| + i\varepsilon\|A^{\frac{1}{2}}\|\|\hat{x}\|\} +$$

$$+ \varepsilon(1+\varepsilon\kappa^{\frac{1}{2}})^i\Big\{(4\kappa^{\frac{1}{2}} + C_1\kappa)(1+o(1)) \sum_{\ell=0}^{i-1} \|a_\ell A^{\frac{1}{2}} p_\ell\| +$$

$$+ 2\kappa^{\frac{1}{2}} \sum_{\ell=0}^{i-1} \|A^{-\frac{1}{2}} r_\ell\|\Big\} .$$

PROOF. For convenience we introduce the abbreviation

$$(2) \qquad y_i := A^{-\frac{1}{2}}(\hat{r}_i - r_i) .$$

All o-symbols are assumed to hold under the restriction $\varepsilon \to 0$. For the computation of r_{i+1} we have, according to (2.3.1.25),

$$(3) \qquad r_{i+1} = r_i - a_i Ap_i + \delta r'_{i+1} ,$$

$$(4) \qquad \|\delta r'_{i+1}\| \leq (2\varepsilon\|a_i Ap_i\| + \varepsilon C_1\|A\|\|a_i p_i\|)(1+o(1)) + \varepsilon\|r_i\| .$$

Consequently,

$$(5) \qquad \|A^{-\frac{1}{2}}\delta r'_{i+1}\| \leq (2\varepsilon\kappa^{\frac{1}{2}} + \varepsilon C_1\kappa)\|a_i A^{\frac{1}{2}} p_i\|(1+o(1)) + \varepsilon\kappa^{\frac{1}{2}}\|A^{-\frac{1}{2}} r_i\| .$$

For the computation of x_{i+1} we have

$$(6) \qquad x_{i+1} = \mathrm{fl}(x_i + a_i p_i) = (I + F_i'''')(x_i + (I + F_i''')a_i p_i) =$$

$$= x_i + a_i p_i + \delta x_{i+1} ,$$

$$(7) \qquad \delta x_{i+1} := F_i'''' x_i + (F_i''' + F_i''''(I + F_i'''))a_i p_i .$$

Consequently,

$$(8) \qquad \|A^{\frac{1}{2}}\delta x_{i+1}\| \leq \varepsilon\|A^{\frac{1}{2}}\|\|x_i\| + 2\varepsilon\kappa^{\frac{1}{2}}\|a_i A^{\frac{1}{2}} p_i\|(1+o(1)) .$$

Since

$$(9) \qquad \|x_i\| = \|\hat{x} - A^{-1}\hat{r}_i\| \leq \|\hat{x}\| + \|A^{-\frac{1}{2}}\|(\|A^{-\frac{1}{2}}(\hat{r}_i - r_i)\| + \|A^{-\frac{1}{2}} r_i\|)$$

we obtain

(10) $\|A^{\frac{1}{2}}\,\delta x_{i+1}\| \le \varepsilon\|A^{\frac{1}{2}}\|\|\hat{x}\| + \varepsilon\kappa^{\frac{1}{2}}(\|y_i\| + \|A^{-\frac{1}{2}}\,r_i\| + 2\|a_i\,A^{\frac{1}{2}}p_i\|(1+o(1)))$.

From (6) we arrive at the following recursion for $\hat{r}_{i+1}$,

(11) $\qquad \hat{r}_{i+1} = b - Ax_{i+1} = b - Ax_i - a_i\,Ap_i - A\,\delta x_{i+1} = \hat{r}_i - a_i\,Ap_i - A\,\delta x_{i+1}$.

Combining this with recursion (3) for r_{i+1} we obtain the following recursion for y_{i+1},

(12) $\qquad y_{i+1} = y_i - A^{\frac{1}{2}}\,\delta x_{i+1} - A^{-\frac{1}{2}}\,\delta r'_{i+1}$.

In view of this and inequalities (5) and (10) we find

(13) $\qquad \|y_{i+1}\| \le (1+\varepsilon\kappa^{\frac{1}{2}})\|y_i\| + \varepsilon\|A^{\frac{1}{2}}\|\|\hat{x}\| +$

$$+ \varepsilon(4\kappa^{\frac{1}{2}} + C_1\kappa)\|a_i\,A^{\frac{1}{2}}p_i\|(1+o(1)) + 2\varepsilon\kappa^{\frac{1}{2}}\|A^{-\frac{1}{2}}\,r_i\| .$$

Backward repetition of this inequality yields

(14) $\qquad \|y_i\| \le (1+\varepsilon\kappa^{\frac{1}{2}})^i\|y_0\| +$

$$+ \varepsilon\sum_{\ell=0}^{i-1}\{(1+\varepsilon\kappa^{\frac{1}{2}})^{i-\ell-1}(\|A^{\frac{1}{2}}\|\|\hat{x}\| + (4\kappa^{\frac{1}{2}} + C_1\kappa)\|a_\ell\,A^{\frac{1}{2}}p_\ell\|(1+o(1)) +$$

$$+ 2\kappa^{\frac{1}{2}}\|A^{-\frac{1}{2}}\,r_\ell\|)\} .$$

Since $(1+\varepsilon\kappa^{\frac{1}{2}})^{i-\ell-1} < (1+\varepsilon\kappa^{\frac{1}{2}})^i$ we finally obtain

(15) $\qquad \|y_i\| \le (1+\varepsilon\kappa^{\frac{1}{2}})^i\{\|y_0\| + i\varepsilon\|A^{\frac{1}{2}}\|\|\hat{x}\|\} +$

$$+ \varepsilon(1+\varepsilon\kappa^{\frac{1}{2}})^i\Big\{(4\kappa^{\frac{1}{2}} + C_1\kappa)(1+o(1))\sum_{\ell=0}^{i-1}\|a_\ell\,A^{\frac{1}{2}}p_\ell\| +$$

$$+ 2\kappa^{\frac{1}{2}}\sum_{\ell=0}^{i-1}\|A^{-\frac{1}{2}}\,r_\ell\|\Big\} ,$$

as had to be proved. □

We are now ready to derive a bound for the difference $\|A^{-\frac{1}{2}}(\hat{r}_i - r_i)\|$ in terms of the solution $\hat{x}$, the initial residual $\hat{r}_0 = b - Ax_0$ and the number of iteration steps carried out.

THEOREM 2. *Let* $\{x_i\}$, $\{r_i\}$ *be computed by RRDM with an arbitrary initial machine vector* x_0 *and let* α, β, γ *denote the bounds of corollary 2.3.1.8.*
Then we have for $i \geq 1$

$$\|A^{-\frac{1}{2}}(\hat{r}_i - r_i)\| \leq \tag{16}$$

$$\leq \varepsilon(1+\varepsilon\kappa^{\frac{1}{2}})^i\{\kappa^{\frac{1}{2}} + C_1\kappa + \sqrt{i}\,(4\kappa^{\frac{1}{2}} + 2\gamma^{-1}\kappa^{\frac{1}{2}} + C_1\kappa)\}\|A^{-\frac{1}{2}}\hat{r}_0\|(1+o(1)) +$$

$$+ \varepsilon(1+\varepsilon\kappa^{\frac{1}{2}})^i\{i + C_1\kappa^{\frac{1}{2}} + \varepsilon\sqrt{i}\,C_1(4\kappa + 2\gamma^{-1}\kappa + C_1\kappa^{3/2})\}\|A^{\frac{1}{2}}\|\|\hat{x}\|(1+o(1))$$

under the restriction

$$\varepsilon\gamma^{-2}\{\kappa^{\frac{1}{2}}(1+C_2) + \kappa\beta(1+\beta) + \alpha(1+C_2+C_1\kappa)\} \to 0\ . \tag{17}$$

PROOF. For the computation of r_0 we have

$$r_0 = \mathrm{fl}(b - Ax_0) = (I+F)(b - (A+E)x_0) = (I+F)(\hat{r}_0 - Ex_0)\ . \tag{18}$$

Hence,

$$A^{-\frac{1}{2}} r_0 = A^{-\frac{1}{2}}\hat{r}_0 + A^{-\frac{1}{2}}F\hat{r}_0 - A^{-\frac{1}{2}}(I+F)Ex_0 \tag{19}$$

and, under the restriction $\varepsilon \to 0$,

$$\|A^{-\frac{1}{2}}(\hat{r}_0 - r_0)\| \leq \varepsilon\kappa^{\frac{1}{2}}\|A^{-\frac{1}{2}}\hat{r}_0\| + \varepsilon C_1\kappa^{\frac{1}{2}}(1+o(1))\|A^{\frac{1}{2}}\|\|x_0\|\ . \tag{20}$$

Since $x_0 = \hat{x} - A^{-1}\hat{r}_0$, we have $\|x_0\| \leq \|\hat{x}\| + \|A^{-\frac{1}{2}}\|\|A^{-\frac{1}{2}}\hat{r}_0\|$.
Substitution in (20) yields, under the restriction $\varepsilon \to 0$,

$$\|A^{-\frac{1}{2}}(\hat{r}_0 - r_0)\| \leq (\varepsilon(\kappa^{\frac{1}{2}} + C_1\kappa)\|A^{-\frac{1}{2}}\hat{r}_0\| + \varepsilon C_1\kappa^{\frac{1}{2}}\|A^{\frac{1}{2}}\|\|\hat{x}\|)(1+o(1))\ , \tag{21}$$

and consequently, since $\|A^{-\frac{1}{2}} r_0\| \leq \|A^{-\frac{1}{2}}(\hat{r}_0 - r_0)\| + \|A^{-\frac{1}{2}}\hat{r}_0\|$,

$$\|A^{-\frac{1}{2}} r_0\| \leq (\|A^{-\frac{1}{2}}\hat{r}_0\| + \varepsilon C_1\kappa^{\frac{1}{2}}\|A^{\frac{1}{2}}\|\|\hat{x}\|)(1+o(1))\ , \tag{22}$$

under the restriction $\varepsilon(\kappa^{\frac{1}{2}} + C_1\kappa) \to 0$.
From remark 2.3.1.17 and the Cauchy-Schwarz inequality we obtain

$$\sum_{\ell=0}^{i-1}\|a_\ell A^{\frac{1}{2}} p_\ell\| \leq \sqrt{i}\left(\sum_{\ell=0}^{\infty}\|a_\ell A^{\frac{1}{2}} p_\ell\|^2\right)^{\frac{1}{2}} \leq \sqrt{i}\,(1+\gamma^{-2}o(1))\|A^{-\frac{1}{2}} r_0\|\ , \tag{23}$$

under the restriction

$$\varepsilon\{\kappa^{\frac{1}{2}}(1+C_2) + \kappa\beta(1+\beta) + \alpha(1+C_2+C_1\kappa)\} \to 0 \ . \tag{24}$$

Consequently, under the restriction (17), using (22) we get

$$\sum_{\ell=0}^{i-1} \|a_\ell A^{\frac{1}{2}} p_\ell\| \le \sqrt{i}\ (1+o(1))\|A^{-\frac{1}{2}} r_0\| \le \tag{25}$$

$$\le \sqrt{i}\ (\|A^{-\frac{1}{2}} \hat{r}_0\| + \varepsilon C_1 \kappa^{\frac{1}{2}} \|A^{\frac{1}{2}}\| \|\hat{x}\|)(1+o(1)) \ .$$

(Note that restriction (24) is included in restriction (17) since $\gamma^{-2} \ge 1$.)

Analogously, from lemma 2.3.1.14, it follows that

$$\sum_{\ell=0}^{i-1} \|A^{-\frac{1}{2}} r_\ell\| \le \sqrt{i}\left(\sum_{\ell=0}^{\infty} \|A^{-\frac{1}{2}} r_\ell\|^2\right)^{\frac{1}{2}} \le \sqrt{i}\,\gamma^{-1}\|A^{-\frac{1}{2}} r_0\|(1+o(1)) \le \tag{26}$$

$$\le \sqrt{i}\,\gamma^{-1}(\|A^{-\frac{1}{2}} \hat{r}_0\| + \varepsilon C_1 \kappa^{\frac{1}{2}} \|A^{\frac{1}{2}}\| \|\hat{x}\|)(1+o(1))$$

under the restriction (24).

Substitution of (21), (25) and (26) into (1) completes the proof. □

REMARK 3. Since $\gamma_i^{-1} \le \beta_i \kappa^{\frac{1}{2}}$, theorem 2 is also valid if we replace all γ by $\beta\kappa^{\frac{1}{2}}$. □

REMARK 4. Instead of measuring the difference between $\hat{r}_i - r_i$ in terms of the norm $\|A^{-\frac{1}{2}}(\hat{r}_i - r_i)\|$ we may as well try to measure $\|\hat{r}_i - r_i\|$. From (12) it follows that

$$\hat{r}_{i+1} - r_{i+1} = \hat{r}_i - r_i - A\delta x_{i+1} - \delta r'_{i+1} \ , \tag{27}$$

and consequently

$$\|\hat{r}_i - r_i\| \le \|\hat{r}_0 - r_0\| + \sum_{\ell=1}^{i} \|A\,\delta x_\ell\| + \sum_{\ell=1}^{i} \|\delta r'_\ell\| \ . \tag{28}$$

Estimating each separate part we find under the restriction $\varepsilon \to 0$

$$\|\hat{r}_0 - r_0\| \le \varepsilon\|r_0\| + \varepsilon C_1 \|A\| \|x_0\| (1+o(1)) \ ,$$

$$\sum_{\ell=1}^{i} \|A\,\delta x_\ell\| \le \varepsilon\|A\| \sum_{\ell=0}^{i-1} \|x_\ell\| + 2\varepsilon\kappa^{\frac{1}{2}}\|A^{\frac{1}{2}}\|(1+o(1)) \sum_{\ell=0}^{i-1} \|a_\ell A^{\frac{1}{2}} p_\ell\| \ ,$$

$$\sum_{\ell=1}^{i} \|\delta r'_\ell\| \le \varepsilon(2 + C_1\kappa^{\frac{1}{2}})\|A^{\frac{1}{2}}\|(1 + o(1)) \sum_{\ell=0}^{i-1} \|a_\ell A^{\frac{1}{2}} p_\ell\| + \varepsilon\|A^{\frac{1}{2}}\| \sum_{\ell=0}^{i-1} \|A^{-\frac{1}{2}} r_\ell\| ,$$

where we subsequently used formulae (18), (16) and (3).
Substitution into (28) yields, under the restriction $\varepsilon \to 0$,

$$\|\hat{r}_i - r_i\| \le \varepsilon\left\{\|r_0\| + C_1\|A\|\|x_0\| + \|A\| \sum_{\ell=0}^{i-1} \|x_\ell\|\right\} + \tag{29}$$

$$+ \varepsilon\left\{(2 + 2\kappa^{\frac{1}{2}} + C_1\kappa^{\frac{1}{2}})\|A^{\frac{1}{2}}\|(1 + o(1)) \sum_{\ell=0}^{i-1} \|a_\ell A^{\frac{1}{2}} p_\ell\| + \right.$$

$$\left. + \|A^{\frac{1}{2}}\| \sum_{\ell=0}^{i-1} \|A^{-\frac{1}{2}} r_\ell\|\right\} .$$

The last two sums can be replaced by (25) and (26) under the restriction (17). The first part of estimate (29) is an a posteriori estimate; a similar a posteriori estimate can be written down in (1). Comparing (1) and (29) we observe that the corresponding factor for the last two sums in these estimates differ (apart from the numerical constants) by a factor $\|A^{-\frac{1}{2}}\|$ and hence, as far as these two sums are concerned, estimate (29) is sharper, even though it measures $\hat{r}_i - r_i$ in a different norm. However, replacing the first part of (29) by an a priori bound (as we did in lemma 1) destroys this superiority. □

From the inequality

$$\|A^{\frac{1}{2}}(\hat{x} - x_i)\| \le \|A^{-\frac{1}{2}}(\hat{r}_i - r_i)\| + \|A^{-\frac{1}{2}} r_i\| \tag{30}$$

and the fact that $\|A^{-\frac{1}{2}} r_i\| \to 0$ $(i \to \infty)$, if (17) is small enough, we see that (16) yields an estimate for the ultimate behavior of $\|A^{\frac{1}{2}}(\hat{x} - x_i)\|$. However, estimate (16) contains the number i of iteration steps and this estimate certainly is not bounded as i tends to infinity.
In the next theorem we take a suitable value of i. For this value of i the right-hand side of (30) is small in the sense that i is not too large to give an unacceptable bound for $\|A^{-\frac{1}{2}}(\hat{r}_i - r_i)\|$, using (16), and i is not too small to let $\|A^{-\frac{1}{2}} r_i\|$ be unacceptably large.

THEOREM 5. *Let* $\{x_i\}$, $\{r_i\}$ *be computed by RRDM with an arbitrary initial machine vector* x_0 *and let* α, β, γ *denote the bounds of corollary 2.3.1.8. Furthermore, let*

$$N := \mathrm{ent}(2\gamma^{-2} \log \frac{1}{\varepsilon} + 1) \quad . \tag{31}$$

Then we have

$$\|A^{\frac{1}{2}}(\hat{x} - x_N)\| \leq \tag{32}$$

$$\leq \varepsilon\{1 + \kappa^{\frac{1}{2}} + C_1\kappa + 2\gamma^{-1}(4\kappa^{\frac{1}{2}} + 2\gamma^{-1}\kappa^{\frac{1}{2}} + C_1\kappa) \log \frac{1}{\varepsilon}\}\|A^{-\frac{1}{2}} \hat{r}_0\|(1 + o(1)) +$$

$$+ \varepsilon\{1 + C_1\kappa^{\frac{1}{2}} + 2\gamma^{-2} \log \frac{1}{\varepsilon}\}\|A^{\frac{1}{2}}\|\|\hat{x}\|(1 + o(1)) \quad ,$$

under the restriction

$$\varepsilon\gamma^{-2}\{\kappa^{\frac{1}{2}}(1 + C_2) + \kappa\beta(1 + \beta) + \alpha(1 + C_2 + C_1\kappa)\} + \tag{33}$$

$$+ \varepsilon\gamma^{-1}\{\kappa^{\frac{1}{2}}(1 + \gamma^{-1} + C_1\kappa^{\frac{1}{2}}\} \log \frac{1}{\varepsilon} \to 0 \quad .$$

PROOF. From corollary 2.3.1.8 and the definition of N it follows that

$$\|A^{-\frac{1}{2}} r_N\| \leq (1 - \gamma^2(1 + o(1)))^{\frac{1}{2}N}\|A^{-\frac{1}{2}} r_0\| \leq \tag{34}$$

$$\leq (1 - \gamma^2(1 + o(1)))^{\gamma^{-2} \log 1/\varepsilon}\|A^{-\frac{1}{2}} r_0\| =$$

$$= \exp(-(1 + o(1)) \log \frac{1}{\varepsilon})\|A^{-\frac{1}{2}} r_0\| \quad ,$$

under the restriction (2.3.1.70) and consequently also under the restriction (17).
In fact, the o-symbol in (34) stands for the maximum ν of all $|\nu_{i+1}|$ $(0 \leq i \leq N-1)$ of theorem 2.3.1.4. From (2.3.1.46) and (2.3.1.47) it follows that ν can be bounded in terms of (2.3.1.70) and thus in terms of (17). Consequently, the expression $\exp(-(1 + o(1))\log 1/\varepsilon)$ in (34) can be replaced by $\varepsilon(1 + o(1))$ under the restriction (17).
Hence it follows with (22) that

$$\|A^{-\frac{1}{2}} r_N\| \leq \varepsilon(1 + o(1))(\|A^{-\frac{1}{2}} \hat{r}_0\| + o(1)\|A^{\frac{1}{2}}\|\|\hat{x}\|) \quad , \tag{35}$$

under the restriction (17). Furthermore, since $N \leq 2\gamma^{-2} \log \frac{1}{\varepsilon} + 1$ (and

$\varepsilon < e^{-1}$) we have $N^{\frac{1}{2}} \leq 2\gamma^{-1} \log 1/\varepsilon$ and

$$(1+\varepsilon\kappa^{\frac{1}{2}})^N \leq \exp(N\varepsilon\kappa^{\frac{1}{2}}) \leq \exp(\varepsilon\kappa^{\frac{1}{2}}(2\gamma^{-2} \log \tfrac{1}{\varepsilon}+1)) = 1+o(1) ,$$

under the restriction $\varepsilon\gamma^{-2}\kappa^{\frac{1}{2}} \log\frac{1}{\varepsilon} + \varepsilon\kappa^{\frac{1}{2}} \to 0$.
Substitution of these inequalities in theorem 2 and taking $i = N$ yields

$$(36) \qquad \|A^{-\frac{1}{2}}(\xi_N - r_N)\| \leq$$

$$\leq \varepsilon\{\kappa^{\frac{1}{2}} + C_1\kappa + 2\gamma^{-1}(4\kappa^{\frac{1}{2}} + 2\gamma^{-1}\kappa^{\frac{1}{2}} + C_1\kappa) \log\tfrac{1}{\varepsilon}\} \|A^{-\frac{1}{2}} f_0\|(1+o(1)) +$$

$$+ \varepsilon\{2\gamma^{-2} \log\tfrac{1}{\varepsilon} + 1 + C_1\kappa^{\frac{1}{2}}\}\|A^{\frac{1}{2}}\|\|\xi\|(1+o(1)) ,$$

under the restriction (33).
Note that the terms in (33), containing $\log 1/\varepsilon$, are needed to assure that $(1+\varepsilon\kappa^{\frac{1}{2}})^N = 1+o(1)$ and to assure that

$$\varepsilon N^{\frac{1}{2}}(4\kappa^{\frac{1}{2}} + 2\gamma^{-1}\kappa^{\frac{1}{2}} + C_1\kappa) = o(1) .$$

Inequality (32) now follows from (30), (35) and (36). □

REMARK 6. As in the case of theorem 2, theorem 5 is also valid if we replace all γ by $\beta\kappa^{\frac{1}{2}}$ (see remark 3).
Note that in fact $N^{\frac{1}{2}} \leq 2\gamma^{-1}(\log 1/\varepsilon)^{\frac{1}{2}}$ holds and that consequently in (32) the first $\log 1/\varepsilon$ may be replaced by $(\log 1/\varepsilon)^{\frac{1}{2}}$ and that also in (33) this replacement is allowed, but this is a rather small improvement and therefore deleted. □

REMARK 7. Instead of the estimates of subsection 2.3.1 we could have used

$$(37) \qquad \sum_{\ell=0}^{i-1} \|A^{-\frac{1}{2}} r_\ell\| \leq \|A^{-\frac{1}{2}} r_0\| \sum_{\ell=0}^{\infty} (1-\gamma^2(1+o(1)))^{\frac{1}{2}\ell} \leq$$

$$\leq \|A^{-\frac{1}{2}} r_0\| \sum_{\ell=0}^{\infty} (1-\tfrac{1}{2}\gamma^2(1+o(1)))^{\ell} \leq$$

$$\leq 2\gamma^{-2}\|A^{-\frac{1}{2}} r_0\|(1+o(1)) ,$$

and (see remark 2.3.1.2)

$$(38)\qquad \sum_{\ell=0}^{i-1} \|a_\ell A^{\frac{1}{2}} p_\ell\| \le (1+o(1)) \sum_{\ell=0}^{\infty} \|A^{-\frac{1}{2}} r_\ell\| \le 2\gamma^{-2}\|A^{-\frac{1}{2}} r_0\|(1+o(1)) ,$$

under the restriction (24).
Both bounds do not depend on i. However, for $0 \le i \le N$, the bound of (26) is of the same order as the bound supplied by (37), whereas the bound of (38) is a factor γ^{-1} larger than that of (23) (apart from a factor $\log 1/\varepsilon$). □

We now investigate what happens to the approximations x_i as the computations are carried out beyond iteration step N. As we mentioned already in the introduction of this subsection, our result will be based on lemma 1.3.3, due to the fact that we have a machine with proper rounding arithmetic, and the fact that $\|a_i p_i\|$ tends to zero as i tends to infinity, if (33) is small.

THEOREM 8. *Let* $\{x_i\}$, $\{r_i\}$ *be computed by RRDM with an arbitrary initial machine vector* x_0, *let* α, β, γ *denote the bounds of corollary 2.3.1.8 and let* N *be defined as in theorem 5. Then we have for all* $i \ge N$ *the inequality*

$$(39)\qquad \|A^{\frac{1}{2}}(\hat{x} - x_i)\| \le \varepsilon\{1+\kappa^{\frac{1}{2}} + C_1\kappa + 2\gamma^{-2}(1+B\kappa^{\frac{1}{2}}) +$$

$$+ 2\gamma^{-1}(4\kappa^{\frac{1}{2}} + 2\gamma^{-1}\kappa^{\frac{1}{2}} + C_1\kappa)\log\tfrac{1}{\varepsilon}\}\|A^{-\frac{1}{2}}\hat{r}_0\|(1+o(1)) +$$

$$+ \varepsilon\{1 + C_1\kappa^{\frac{1}{2}} + 2\gamma^{-2}\log\tfrac{1}{\varepsilon}\}\|A^{\frac{1}{2}}\|\|\hat{x}\|(1+o(1)) ,$$

under the restriction (33).

PROOF. Since $x_{i+1} = x_i + a_i p_i + \delta x_{i+1}$ (see (6)), we know that for $i > N$

$$(40)\qquad \|A^{\frac{1}{2}}(x_i - x_N)\| \le \sum_{\ell=N}^{i-1} \|a_\ell A^{\frac{1}{2}} p_\ell\| + \sum_{\ell=N+1}^{i} \|A^{\frac{1}{2}}\delta x_\ell\| .$$

Analogously to (38) we find for the first sum the estimate

$$(41)\qquad \sum_{\ell=N}^{i-1} \|a_\ell A^{\frac{1}{2}} p_\ell\| \le (1+o(1)) \sum_{\ell=N}^{\infty} \|A^{-\frac{1}{2}} r_\ell\| \le$$

$$\le 2\gamma^{-2}\|A^{-\frac{1}{2}} r_N\|(1+o(1)) ,$$

under the restriction (33) (note that (33) implies (17), which implies (24)).

This proves the convergence of the first sum. Since generally x_i will not tend to zero as i tends to infinity, the convergence of the second sum in (40) does not follow from (7). However, from lemma 1.3.3, due to the proper rounding arithmetic, we may conclude that instead of (6) one also has

$$x_{i+1} = x_i + (I+H_i)(I+F_i''')a_i p_i , \tag{42}$$

and therefore, instead of (8), under the restriction $\varepsilon \to 0$,

$$\|A^{\frac{1}{2}} \delta x_{i+1}\| \le \|A^{\frac{1}{2}}(F_i''' + H_i(I+F_i'''))a_i p_i\| \le \tag{43}$$

$$\le (\varepsilon + (B+\varepsilon)(1+o(1)))\kappa^{\frac{1}{2}}\|a_i A^{\frac{1}{2}} p_i\| .$$

Together with (41) this yields

$$\sum_{\ell=N+1}^{i} \|A^{\frac{1}{2}} \delta x_\ell\| \le (\varepsilon + (B+\varepsilon)(1+o(1)))\kappa^{\frac{1}{2}} \sum_{\ell=N}^{i-1} \|a_\ell A^{\frac{1}{2}} p_\ell\| \le \tag{44}$$

$$\le 2B\gamma^{-2}\kappa^{\frac{1}{2}}\|A^{-\frac{1}{2}} r_N\|(1+o(1)) ,$$

under the restriction (33).

From (34) and (22) we know that under the restriction (17)

$$\|A^{-\frac{1}{2}} r_N\| \le \varepsilon(1+o(1))(\|A^{-\frac{1}{2}} f_0\| + \varepsilon C_1 \kappa^{\frac{1}{2}}\|A^{\frac{1}{2}}\|\|\hat{x}\|) . \tag{45}$$

Hence, from (40), (41), (44) and (45) it follows that

$$\|A^{\frac{1}{2}}(x_i - x_N)\| \le 2\gamma^{-2}(1+B\kappa^{\frac{1}{2}})\|A^{-\frac{1}{2}} r_N\|(1+o(1)) \le \tag{46}$$

$$\le 2\varepsilon\gamma^{-2}(1+B\kappa^{\frac{1}{2}})\|A^{-\frac{1}{2}} f_0\|(1+o(1)) + \varepsilon o(1)\|A^{\frac{1}{2}}\|\|\hat{x}\|$$

under the restriction (33). (We remark that the second o-symbol stands for $2\varepsilon\gamma^{-2}(1+B\kappa^{\frac{1}{2}})C_1\kappa^{\frac{1}{2}}$; if the base B of the floating point numbers is regarded as a fixed integer, then the term containing B can be omitted in (33), as is actually done.)

Since $\|A^{\frac{1}{2}}(\hat{x} - x_i)\| \le \|A^{\frac{1}{2}}(\hat{x} - x_N)\| + \|A^{\frac{1}{2}}(x_i - x_N)\|$, inequality (39) follows from (32) and (46). □

Again we mention that in theorem 8 all γ's may be replaced by $\beta\kappa^{\frac{1}{2}}$.

REMARK 9. Comparing (32) and (39) we see that the bounds for the natural error $\|A^{\frac{1}{2}}(\hat{x} - x_i)\|$ at step N and at all later steps do not differ essentially. Here we shall not present a detailed discussion on the influence of and the interaction between the separate terms occurring in estimate (39). This discussion will be given for the gradient method in Chapter 3 and for the conjugate gradient method in Chapter 4. Note that, if $r_i = 0$ for some $i \geq 0$, then certainly (32) holds for $\|A^{\frac{1}{2}}(\hat{x} - x_i)\|$ if $i \leq N$ and (39) holds if $i > N$. □

Algebraically for every DM, $\|A^{-\frac{1}{2}} r_i\|$ converges step-wise linearly to zero (cf. section 1.4) under the conditions of corollary 2.2.5 and consequently $\|A^{-\frac{1}{2}} r_{i+1}\| < \|A^{-\frac{1}{2}} r_i\|$ holds for $i \geq 0$. We have seen that for RRDM this also holds numerically if expression (2.3.1.70) is small enough. Algebraically, for every DM, $\|A^{\frac{1}{2}}(\hat{x} - x_i)\| = \|A^{-\frac{1}{2}} r_i\|$ and consequently also $\|A^{\frac{1}{2}}(\hat{x} - x_{i+1})\| < \|A^{\frac{1}{2}}(\hat{x} - x_i)\|$ holds for all $i \geq 0$. However, in the presence of round-off the relation $\|A^{\frac{1}{2}}(\hat{x} - x_i)\| = \|A^{-\frac{1}{2}} r_i\|$ does not hold and consequently one might ask whether $\|A^{\frac{1}{2}}(\hat{x} - x_i)\|$ converges step-wise linearly for all $i \geq 0$. Another question concerns the magnitude of the error $\hat{x} - x_i$, measured in some norm, at the moment where the step-wise convergence of $\|A^{\frac{1}{2}}(\hat{x} - x_i)\|$ is destroyed by round-off errors.

These questions are not only interesting for RRDM but also for TRDM. In the next section, where we consider TRDM, we do not give an upper bound for the natural error $\|A^{\frac{1}{2}}(\hat{x} - x_i)\|$ at a fixed step ($i = N$), neither do we give a limes superior result for the natural error, but we only deal with the problem just mentioned, viz., the monotonic decrease of $\|A^{\frac{1}{2}}(\hat{x} - x_i)\|$ is disturbed at a certain step, what can be said about the error $\|\hat{x} - x_i\|$? In the remaining part of the present subsection we consider this problem for RRDM which enables us to point out the difference with TRDM. For brevity, we treat the case that there is only one type of arithmetical operations during the process where round-off occurs, whereas all other arithmetical operations are assumed to be carried out with complete accuracy. This kind of incomplete error analysis is a so-called one-round-off error analysis (cf. section 1.3). The only operation involving round-off will be the matrix by vector product computations, i.e. $fl(Ap_i) = (A + E_i)p_i$, $\|E_i\| \leq \varepsilon C_i \|A\|$ (we assume that $A * x_0$ is carried out with complete

accuracy). The reason for this choice is that in the restrictions and estimates of the foregoing theorems the terms including C_1, corresponding to matrix by vector computations, contain the largest powers of κ and therefore this operation seems to have the largest influence on the numerical behavior.
Of course, all results deduced so far in this chapter for RRDM and using floating point arithmetic for all arithmetic operations, are also valid in the one-round-off case at hand. It is obvious that terms appearing in estimates and restrictions and not containing the factor C_1 can be omitted, since they entered the round-off error analysis because of other arithmetical operations than the matrix by vector product computations. For instance, restriction (17) is replaced by restriction (49).

THEOREM 10. *Let* $\{x_i\}$, $\{r_i\}$ *be computed by RRDM with an arbitrary initial machine vector* x_0. *Assume that only the matrix by vector products* Ap_i $(i \geq 0)$ *are carried out in floating point arithmetic, all other arithmetical operations being executed exactly. Let* α_i, α, β, γ_i *denote the parameters and bounds of corollary 2.3.1.8.*
Then we have for all $i \geq 0$

$$\frac{\|A^{\frac{1}{2}}(\hat{x} - x_{i+1})\|^2}{\|A^{\frac{1}{2}}(\hat{x} - x_i)\|^2} = 1 - \gamma_i^2 \frac{\|A^{-\frac{1}{2}} r_i\|^2}{\|A^{-\frac{1}{2}} \hat{r}_i\|^2} (1 + \omega_{i+1}) , \tag{47}$$

where $\hat{r}_i := b - Ax_i$ *and*

$$|\omega_{i+1}| \leq 2\varepsilon \sqrt{1}\, \alpha_i C_1 \kappa^{\frac{1}{2}} \|A^{\frac{1}{2}}\| \frac{\|A^{-\frac{1}{2}} \hat{r}_0\|}{\|r_i\|} (1 + o(1)) + o(1) , \tag{48}$$

under the restriction

$$\varepsilon \gamma^{-2} \alpha C_1 \kappa \to 0 . \tag{49}$$

PROOF. Since $x_{i+1} = x_i + a_i p_i$, or equivalently, $A^{-\frac{1}{2}} \hat{r}_{i+1} = A^{-\frac{1}{2}} \hat{r}_i + - a_i A^{\frac{1}{2}} p_i$, we have for $A^{-\frac{1}{2}} \hat{r}_{i+1}$ $(= A^{\frac{1}{2}}(\hat{x} - x_{i+1}))$

$$\|A^{-\frac{1}{2}} \hat{r}_{i+1}\|^2 = \|A^{-\frac{1}{2}} \hat{r}_i\|^2 - 2a_i(\hat{r}_i, p_i) + a_i^2 \|A^{\frac{1}{2}} p_i\|^2 . \tag{50}$$

In view of

(51) $$(p_i, fl(Ap_i)) = (p_i, (A+E_i)p_i) = (p_i, Ap_i)(1+\mu_i) ,$$

(52) $$|\mu_i| = |(p_i, E_i p_i)| / \|A^{\frac{1}{2}} p_i\|^2 \le \varepsilon C_1 \|A\| \|p_i\|^2 / \|A^{\frac{1}{2}} p_i\|^2 \le \varepsilon C_1 \kappa ,$$

we have the following relation for a_i

(53) $$a_i = \frac{(r_i, p_i)}{(p_i, fl(Ap_i))} = \frac{(r_i, p_i)}{(p_i, Ap_i)} \frac{1}{(1+\mu_i)} = \frac{(r_i, p_i)}{(p_i, Ap_i)} (1+\delta a_i'') ,$$

(54) $$|\delta a_i''| = |\mu_i / (1+\mu_i)| = \varepsilon C_1 \kappa (1 + o(1)) = o(1) ,$$

under the restriction $\varepsilon C_1 \kappa \to 0$.
For the computation of r_{i+1} we obtain

(55) $$r_{i+1} = r_i - a_i fl(Ap_i) = r_i - a_i (A+E_i) p_i =$$

$$= r_i - a_i Ap_i + \delta r'_{i+1} ,$$

(56) $$\delta r'_{i+1} = \|a_i E_i p_i\| \le \varepsilon C_1 \kappa^{\frac{1}{2}} \|A^{\frac{1}{2}}\| \|a_i A^{\frac{1}{2}} p_i\| .$$

From $x_{i+1} = x_i + a_i p_i$ it follows that the residual vector $\hat{r}_i$ satisfies $\hat{r}_{i+1} = \hat{r}_i - a_i Ap_i$ and hence, together with (55), this yields $\hat{r}_{i+1} - r_{i+1} = \hat{r}_i - r_i - \delta r'_{i+1}$. Since $A * x_0$ is assumed to be computed with complete accuracy, we have $r_0 = \hat{r}_0$ and hence

(57) $$\hat{r}_i = r_i + \delta \bar{r}_i , \quad \delta \bar{r}_i := - \sum_{\ell=1}^{i} \delta r'_\ell .$$

Consequently,

(58) $$(\hat{r}_i, p_i) = (r_i, p_i)(1+\lambda_i) ,$$

$$|\lambda_i| = |(\delta \bar{r}_i, p_i) / (r_i, p_i)| \le (\|\delta \bar{r}_i\| / \|r_i\|) \alpha_i .$$

Substitution of (53) and (58) into (50) yields

(59) $$\|A^{-\frac{1}{2}} \hat{r}_{i+1}\|^2 = \|A^{-\frac{1}{2}} \hat{r}_i\|^2 +$$

$$- (r_i, p_i)^2 / \|A^{\frac{1}{2}} p_i\|^2 (1 + 2\lambda_i (1 + \delta a_i'') - (\delta a_i'')^2) ,$$

or (using that $\gamma_i = |(r_i, p_i)| / (\|A^{-\frac{1}{2}} r_i\| \|A^{\frac{1}{2}} p_i\|)$),

$$(60)\qquad \frac{\|A^{-\frac{1}{2}}\,\hat{r}_{i+1}\|^2}{\|A^{-\frac{1}{2}}\,\hat{r}_i\|^2} = 1 - \gamma_i^2\,\frac{\|A^{-\frac{1}{2}}\,r_i\|^2}{\|A^{-\frac{1}{2}}\,\hat{r}_i\|^2}\,(1+\omega_{i+1})\ ,$$

$$(61)\qquad \omega_{i+1} = 2\lambda_i(1+\delta a_i'') - (\delta a_i'')^2\ .$$

From (2.3.1.84) and (56) we obtain (cf. (29))

$$(62)\qquad \|\delta\bar{r}_i\| \le \sum_{\ell=1}^{i}\|\delta r_\ell'\| \le \varepsilon C_1\kappa^{\frac{1}{2}}\|A^{\frac{1}{2}}\|\sum_{\ell=0}^{i-1}\|a_\ell A^{\frac{1}{2}}p_\ell\| \le$$

$$\le \varepsilon\sqrt{i}\,C_1\kappa^{\frac{1}{2}}\|A^{\frac{1}{2}}\|\left(\sum_{\ell=0}^{\infty}\|a_\ell A^{\frac{1}{2}}p_\ell\|^2\right) \le$$

$$\le \varepsilon\sqrt{i}\,C_1\kappa^{\frac{1}{2}}\|A^{\frac{1}{2}}\|\|A^{-\frac{1}{2}}\hat{r}_0\|\,(1+o(1))\ ,$$

under the restriction (49).
Estimate (48) now follows from (54), (58) and (62). □

REMARK 11. From (57) it follows that

$$(63)\qquad \|A^{-\frac{1}{2}}\,r_i\|^2 / \|A^{-\frac{1}{2}}\,\hat{r}_i\|^2 = 1+\theta_i\ ,$$

$$(64)\qquad \theta_i \le \frac{\|A^{-\frac{1}{2}}\,\delta\bar{r}_i\|}{\|A^{-\frac{1}{2}}\,\hat{r}_i\|}\left(2 + \frac{\|A^{-\frac{1}{2}}\,\delta\bar{r}_i\|}{\|A^{-\frac{1}{2}}\,\hat{r}_i\|}\right)\ ,$$

and hence (47) can be written as

$$(65)\qquad \frac{\|A^{\frac{1}{2}}(\hat{x}-x_{i+1})\|^2}{\|A^{\frac{1}{2}}(\hat{x}-x_i)\|^2} = 1 - \gamma_i^2(1+\omega_{i+1}')\ ,$$

where $\omega_{i+1}' = \omega_{i+1}(1+\theta_i) + \theta_i$. For our purposes, however, formula (47) is more appropriate. □

The type of estimate (48) for ω_{i+1} is different from the type of estimate for η_{i+1} as given in theorem 2.3.1.1 and for ν_{i+1} as given in theorem 2.3.1.4. The estimate for η_{i+1} is uniformly bounded for $i \ge 0$ and this also holds for ν_{i+1} if α_i is uniformly bounded for $i \ge 0$. The estimate of ω_{i+1} certainly does not have this property since it contains a factor i in the numerator and a factor $\|r_i\|$ in the denominator for which algebraically $\|r_i\| \to 0$ $(i \to \infty)$.

Another minor distinction between η_{i+1} and ν_{i+1} on the one hand and ω_{i+1} on the other hand, is the presence of the initial natural error $\|A^{-\frac{1}{2}} \hat{r}_0\|$.

We recall that we are interested in what can be said about the magnitude of the error $\hat{x} - x_i$, measured in some norm, at the moment when for the first time $\|A^{\frac{1}{2}}(\hat{x} - x_{i+1})\| > \|A^{\frac{1}{2}}(\hat{x} - x_i)\|$. It follows from (47) that at this specific moment certainly $\omega_{i+1} < -1$ and hence $|\omega_{i+1}| > 1$. In order to demonstrate how this last inequality can be used to estimate $\hat{x} - x_i$ in terms of the norm $\|A(\hat{x} - x_i)\|$, we first consider an explicit version of theorem 10.

By retracing the proof of theorem 10, like we did in subsection 2.3.1 for theorem 2.3.1.4, one can prove that if

$$\varepsilon\gamma^{-2} C_1 \kappa(1+\alpha) \leq \frac{1}{8}, \tag{66}$$

then estimates (62) and (48) can be replaced by explicit bounds, i.e.,

$$\|\delta\bar{r}_i\| \leq \frac{5}{4} \varepsilon \sqrt{i}\, C_1 \kappa^{\frac{1}{2}} \|A^{\frac{1}{2}}\| \|A^{-\frac{1}{2}} \hat{r}_0\| \tag{67}$$

and

$$|\omega_{i+1}| \leq \frac{6}{5} \varepsilon \sqrt{i}\, \alpha_i C_1 \kappa^{\frac{1}{2}} \|A^{\frac{1}{2}}\| \frac{\|A^{-\frac{1}{2}} \hat{r}_0\|}{\|r_i\|} + \left(\frac{1}{15}\right)^2 . \tag{68}$$

We now assume that (66) is satisfied and that $\|A^{\frac{1}{2}}(\hat{x} - x_{i+1})\| > \|A^{\frac{1}{2}}(\hat{x} - x_i)\|$ holds for some $i \geq 0$. Then $|\omega_{i+1}| > 1$ certainly implies

$$\|r_i\| < 2 \frac{1}{4} \varepsilon \sqrt{i}\, \alpha_i C_1 \kappa^{\frac{1}{2}} \|A^{\frac{1}{2}}\| \|A^{-\frac{1}{2}} \hat{r}_0\| . \tag{69}$$

Since $\|\hat{r}_i\| \leq \|\delta\bar{r}_i\| + \|r_i\|$ and $\alpha_i \geq 1$, we obtain, together with (67), for the residual vector $A(\hat{x} - x_i)$ $(= \hat{r}_i)$,

$$\|A(\hat{x} - x_i)\| \leq 3 \frac{9}{20} \varepsilon \sqrt{i}\, \alpha_i C_1 \kappa^{\frac{1}{2}} \|A^{\frac{1}{2}}\| \|A^{-\frac{1}{2}} \hat{r}_0\| . \tag{70}$$

Summarizing, if (66) is satisfied and $\|A^{\frac{1}{2}}(\hat{x} - x_{i+1})\| > \|A^{\frac{1}{2}}(\hat{x} - x_i)\|$, then the residual vector $A(\hat{x} - x_i)$ satisfies (70).

We now return to the general case. Using o-symbols, the following corollary of theorem 10 follows from the same arguments as used in the explicit example.

COROLLARY 12. *Consider RRDM performed under the conditions of theorem 10. If* $\|A^{\frac{1}{2}}(\hat{x} - x_{i+1})\| > \|A^{\frac{1}{2}}(\hat{x} - x_i)\|$ *for some* $i \geq 0$, *then the residual vector* $A(\hat{x} - x_i)$ *satisfies*

$$\|A(\hat{x} - x_i)\| \leq 3\varepsilon \sqrt{i}\, \alpha_i\, C_1 \kappa^{\frac{1}{2}} \|A^{\frac{1}{2}}\| \|A^{-\frac{1}{2}} \hat{r}_0\| (1 + o(1)) , \tag{71}$$

under the restriction (49). □

The meaning of (71) diminishes as i increases. However, if i is no greater than N, defined in theorem 5, then certainly $\sqrt{i} \leq 2\gamma^{-1} \log 1/\varepsilon$ and consequently

$$\|A(\hat{x} - x_i)\| \leq 6\varepsilon\gamma^{-1} \alpha_i\, C_1 \kappa^{\frac{1}{2}} \|A^{\frac{1}{2}}\| \|A^{-\frac{1}{2}} \hat{r}_0\| \log \frac{1}{\varepsilon} (1 + o(1)) \tag{72}$$

under the restriction (49). Note that (72) contains the uniform bound γ for γ_i and the actual value of α_i. For values of i greater than N we have already in view of estimate (39) of theorem 8,

$$\|A(\hat{x} - x_i)\| \leq \varepsilon(1 + 2\gamma^{-1} \log \frac{1}{\varepsilon})\, C_1 \kappa \|A^{-\frac{1}{2}} \hat{r}_0\| (1 + o(1)) , \tag{73}$$

under the restriction $\varepsilon\gamma^{-2} \alpha C_1 \kappa + \varepsilon\gamma^{-1} C_1 \kappa \log 1/\varepsilon \to 0$, and hence, for the residual vector

$$\|A(\hat{x} - x_i)\| \leq \varepsilon(1 + 2\gamma^{-1} \log \frac{1}{\varepsilon}) C_1 \kappa \|A^{\frac{1}{2}}\| \|A^{-\frac{1}{2}} \hat{r}_0\| (1 + o(1)) . \tag{74}$$

(The absence in (73) of those terms in (39) not containing C_1 has been explained already; all terms in (39) containing $\|\hat{x}\|$ are absent because of the supposed exact computation of $r_0 = b - Ax_0$ and $x_{i+1} = x_i + a_i p_i$.)

So, we can use (71) if monotonicity breaks down before iteration step N and we can use (73) or (74) if this happens after iteration step N. In fact, if monotonicity breaks down at a certain step, one might as well stop the iterative process since then apparently algebraic properties are drastically disturbed by round-off errors. The problem of observing in practice, when monotonicity of $\|A^{\frac{1}{2}}(\hat{x} - x_i)\|$ breaks down is discussed in remark 2.4.9.

2.4. *The true residual descent methods (TRDM)*

In this section we consider the numerical behavior of TRDM. The only difference between RRDM and TRDM is the way in which the residual vector r_i is determined. In TRDM one determines r_i from the relation $r_i = b - Ax_i$, in fact this relation explains the name residual vector. The vectors x_i and r_i are directly coupled at each iteration step. Round-off occurring at the computation of x_i immediately affects the computed vector r_i and therefore TRDM seems to be more self-restoring than RRDM, where the sequence $\{r_i\}$ could be computed without even computing the sequence $\{x_i\}$. Moreover, the difference between the computed residual r_i and the exact residual $b - Ax_i$ is only caused by computational round-off at one step, whereas for RRDM we saw that this difference is influenced by all previous round-off errors, except for the round-off during the computation of a_i. In computing the residual from the relation $r_i = b - Ax_i$, a round-off error is introduced which ultimately is at least of the order $\varepsilon C_1 \|A\| \|\hat{x}\|$. Consequently, in the case of finite accuracy, it is rather unlikely that r_i tends to zero as i tends to infinity. Here we have a first, although rather obvious, difference between the numerical behavior of TRDM and RRDM.
As far as computational work is concerned, for RRDM only one matrix by vector product $A * p_i$ is computed at each iteration step, whereas for TRDM there are two, viz., $A * p_i$ and $A * x_i$. Therefore, from this point of view RRDM is to be preferred.

Recall that TRDM, corresponding to a given sequence of arbitrary non-zero vectors $\{p_i\}$, consists of the following statements.

TRDM
Choose an initial point x_0;
$r_0 := b - Ax_0$; $i := 0$;
while $r_i \neq 0$ do
begin

(1) $a_i := (r_i, p_i) / (p_i, Ap_i)$;

(2) $x_{i+1} := x_i + a_i p_i$;

(3) $r_{i+1} := b - Ax_i$;

$i := i + 1$

end.

It is obvious that for TRDM we cannot study the numerical convergence of $\{r_i\}$ separately from the numerical convergence of $\{x_i\}$. The following round-off error analysis is performed under the same conditions and conventions as in the previous section and it is also rather parallel and equal in character to the analysis done there. The presence of the same symbols $\hat{a}_i$, $\delta a_i'$, $\delta r_{i+1}'$, etc., does not indicate that they stand for exactly the same quantities, but it only expresses a certain correspondence.

Before starting off the round-off error analysis of TRDM, we first deduce some auxiliary results concerning the computation of a residual vector b - Ax. This result will be used in many considerations where the computation of a true residual vector is carried out.

LEMMA 1. *Let* b, x *be two machine vectors and let*

$$f := b - Ax , \quad r := fl(b - Ax) . \tag{4}$$

Then we have

$$(f,r) = \|f\|^2(1 + o(1)) = \|r\|^2(1 + o(1)) , \tag{5}$$

under the restriction $\varepsilon(1 + C_1\varphi) \to 0$

$$(f, A^{-1}r) = \|A^{-\frac{1}{2}}f\|^2 (1 + o(1)) = \|A^{-\frac{1}{2}}r\|^2 (1 + o(1)) , \tag{6}$$

under the restriction $\varepsilon\kappa^{\frac{1}{2}}(1 + C_1\psi) \to 0$, *and*

$$(f,Ar) = \|A^{\frac{1}{2}}f\|^2 (1 + o(1)) = \|A^{\frac{1}{2}}r\|^2 (1 + o(1)) , \tag{7}$$

under the restriction $\varepsilon(\kappa^{\frac{1}{2}} + C_1\chi) \to 0$, *where*

$$\varphi := \|A\|\|x\| / \|f\| , \quad \psi := \|A^{\frac{1}{2}}\|\|x\| / \|A^{-\frac{1}{2}}f\| , \tag{8}$$

$$\chi := \|A^{3/2}\|\|x\| / \|A^{1/2}f\| .$$

PROOF. We have

$$r = fl(b - Ax) = (I + F)(b - (A + E)x) = f + \delta r , \tag{9}$$

$$\delta r := F(b - Ax) - (I + F)Ex . \tag{10}$$

Consequently,

$$\|\delta r\| \le \varepsilon\|f\| + \varepsilon(1 + \varepsilon)C_1\|A\|\|x\| \tag{11}$$

and hence, under the restriction $\varepsilon \to 0$,

$$\|\delta r\| / \|\hat{r}\| \leq \varepsilon(1 + C_1\varphi(1 + o(1))) , \tag{12}$$

$$\|A^{-\frac{1}{2}} \delta r\| / \|A^{-\frac{1}{2}} \hat{r}\| \leq \varepsilon\kappa^{\frac{1}{2}}(1 + C_1\psi(1 + o(1))) , \tag{13}$$

$$\|A^{\frac{1}{2}} \delta r\| / \|A^{\frac{1}{2}} \hat{r}\| \leq \varepsilon(\kappa^{\frac{1}{2}} + C_1\chi(1 + o(1))) , \tag{14}$$

or

$$\|\delta r\| / \|\hat{r}\| = o(1) , \text{ under the restriction } \varepsilon(1 + C_1\varphi) \to 0 , \tag{16}$$

$$\|A^{-\frac{1}{2}} \delta r\| / \|A^{-\frac{1}{2}} \hat{r}\| = o(1) , \text{ under the restriction } \varepsilon\kappa^{\frac{1}{2}}(1 + C_1\psi) \to 0 , \tag{17}$$

$$\|A^{\frac{1}{2}} \delta r\| / \|A^{\frac{1}{2}} \hat{r}\| = o(1) , \text{ under the restriction } \varepsilon(\kappa^{\frac{1}{2}} + C_1\chi) \to 0 . \tag{18}$$

The first equalities in (5), (6) and (7) follow immediately from (9) and the appropriate inequality (16), (17) or (18). The second equalities follow from the fact that for $\ell = 0, -\frac{1}{2}, \frac{1}{2}$ one has

$$\left| \|A^{\ell} r\| - \|A^{\ell} \hat{r}\| \right| \leq \|A^{\ell} \delta r\| \leq \|A^{\ell} \hat{r}\| o(1) ,$$

under the appropriate restriction. □

REMARK 2. Note that

$$\psi \leq \varphi \leq \chi \leq \kappa^{\frac{1}{2}}\varphi \leq \kappa\psi . \tag{18}$$

Consequently, in the restriction of lemma 1 all ψ may be replaced by φ or χ, all φ may be replaced by χ or $\kappa^{\frac{1}{2}}\psi$, and all χ may be replaced by $\kappa^{\frac{1}{2}}\varphi$ or $\kappa\psi$. □

We now deduce a theorem where the influence of round-off on relation (i) of theorem 2.2.2 is expressed in terms of an absolute error. The proof is very similar to the proof of theorem 2.3.1.1.

THEOREM 3. *Let* x_i, p_i *be two arbitrary machine vectors* ($p_i \neq 0$) *and let* x_{i+1} *be computed from one step TRDM, based on these vectors. Let*

$$\hat{r}_i := b - Ax_i , \tag{19}$$

$$\gamma_i := |(\hat{r}_i, p_i)| / (\|A^{-\frac{1}{2}} \hat{r}_i\| \|A^{\frac{1}{2}} p_i\|) , \tag{20}$$

(21) $$\psi_i := \|A^{\frac{1}{2}}\|\|x_i\| \,/\, \|A^{-\frac{1}{2}} f_i\| .$$

Then we have

(22) $$\frac{\|A^{\frac{1}{2}}(\hat{x} - x_{i+1})\|^2}{\|A^{\frac{1}{2}}(\hat{x} - x_i)\|^2} = 1 - \gamma_i^2 + \eta_{i+1} ,$$

where

(23) $$|\eta_{i+1}| \leq 4\varepsilon(2\kappa^{\frac{1}{2}} + \psi_i)(1 + o(1)) + \varepsilon\kappa^{\frac{1}{2}}(C_2 + C_1\kappa^{\frac{1}{2}} + C_1\psi_i)o(1) ,$$

under the restriction

(24) $$\varepsilon\{\kappa^{\frac{1}{2}}(1 + C_2 + C_1\kappa^{\frac{1}{2}}) + (1 + C_1\kappa^{\frac{1}{2}})\psi_i\} \to 0 .$$

PROOF. From (9) and (13) we know that the computed vector r_i, under the restriction $\varepsilon \to 0$, satisfies

(25) $$r_i = f_i + \delta r_i ,$$

(26) $$\|A^{-\frac{1}{2}} \delta r_i\| \,/\, \|A^{-\frac{1}{2}} f_i\| \leq \varepsilon\kappa^{\frac{1}{2}}(1 + C_1\psi_i(1 + o(1))) .$$

From the proof of theorem 2.3.1.1 it follows that for the computed a_i there holds

(27) $$a_i = \mathrm{fl}\left(\frac{\mathrm{fl}((r_i, p_i))}{\mathrm{fl}((p_i, Ap_i))}\right) = \frac{(r_i, p_i)}{(p_i, Ap_i)} + \delta a_i' ,$$

(28) $$|\delta a_i'| \leq (2\varepsilon C_2\kappa^{\frac{1}{2}} + \varepsilon C_1\kappa + \varepsilon)\|A^{-\frac{1}{2}} r_i\| \,/\, \|A^{\frac{1}{2}} p_i\|(1 + o(1)) ,$$

under the restriction

(29) $$\varepsilon(1 + C_2\kappa^{\frac{1}{2}} + C_1\kappa) \to 0 .$$

Together with lemma 1 this yields

(30) $$|\delta a_i'| \leq (2\varepsilon C_2\kappa^{\frac{1}{2}} + \varepsilon C_1\kappa + \varepsilon)\|A^{-\frac{1}{2}} f_i\| \,/\, \|A^{\frac{1}{2}} p_i\|(1 + o(1)) ,$$

under the restriction

(31) $$\varepsilon\kappa^{\frac{1}{2}}(1 + C_2 + C_1\kappa^{\frac{1}{2}} + C_1\psi_i) \to 0 .$$

Substitution of $(r_i, p_i) = (f_i, p_i) + (\delta r_i, p_i)$ into (27) yields

(32) $a_i = \hat{a}_i + \delta\bar{a}_i$,

where

(33) $\hat{a}_i := (f_i, p_i) / (p_i, Ap_i)$,

(34) $\delta\bar{a}_i := (\delta r_i, p_i) / \|A^{\frac{1}{2}} p_i\|^2 + \delta a_i'$.

From (26) we obtain, under the restriction $\varepsilon \to 0$,

(35) $|(\delta r_i, p_i)| / \|A^{\frac{1}{2}} p_i\|^2 \le \|A^{-\frac{1}{2}} \delta r_i\| / \|A^{\frac{1}{2}} p_i\| \le$

$\le \varepsilon\kappa^{\frac{1}{2}}(1 + C_1\psi_i(1 + o(1)))\|A^{-\frac{1}{2}} f_i\| / \|A^{\frac{1}{2}} p_i\|$.

Together with (30) this yields, under the restriction (31),

(36) $|\delta\bar{a}_i| \le \varepsilon(1 + \kappa^{\frac{1}{2}} + 2C_2\kappa^{\frac{1}{2}} + C_1\kappa + C_1\kappa^{\frac{1}{2}}\psi_i)\|A^{-\frac{1}{2}} f_i\| / \|A^{\frac{1}{2}} p_i\|(1 + o(1))$.

This obviously implies, under the restriction (31),

(37) $|\delta\bar{a}_i| \le \|A^{-\frac{1}{2}} f_i\| / \|A^{\frac{1}{2}} p_i\| o(1)$.

For the computation of x_{i+1} we have

(38) $x_{i+1} = fl(x_i + a_i p_i) = (I + F_i''')(x_i + (I + F_i'')a_i p_i) =$

$= x_i + \hat{a}_i p_i + \delta x_{i+1}$,

(39) $\delta x_{i+1} := F_i''' x_i + \delta\bar{a}_i p_i + V_i p_i$,

(40) $V_i := a_i(F_i'' + F_i'''(I + F_i''))$.

Consequently, from (32), (37) and the fact that $|\hat{a}_i| \le \|A^{-\frac{1}{2}} f_i\| / \|A^{\frac{1}{2}} p_i\|$ we obtain

(41) $\|V_i\| \le 2\varepsilon(|\hat{a}_i| + |\delta\bar{a}_i|)(1 + o(1)) \le$

$\le 2\varepsilon\|A^{-\frac{1}{2}} f_i\| / \|A^{\frac{1}{2}} p_i\|(1 + o(1))$,

under the restriction (31).

We now express η_{i+1} in terms of δx_{i+1}. From (38) it follows that

(42) $$A^{\frac{1}{2}}(\hat{x} - x_{i+1}) = A^{\frac{1}{2}}(\hat{x} - x_i) - \hat{a}_i A^{\frac{1}{2}} p_i - A^{\frac{1}{2}} \delta x_{i+1} ,$$

and hence, by taking squared norm of both sides,

(43) $$\|A^{\frac{1}{2}}(\hat{x} - x_{i+1})\|^2 = \|A^{\frac{1}{2}}(\hat{x} - x_i) - \hat{a}_i A^{\frac{1}{2}} p_i\|^2 +$$

$$- 2(f_i - \hat{a}_i A p_i, \delta x_{i+1}) + \|A^{\frac{1}{2}} \delta x_{i+1}\|^2 .$$

From the definition of $\hat{a}_i$ we obtain

(44) $$\|A^{\frac{1}{2}}(\hat{x} - x_{i+1})\|^2 = \|A^{\frac{1}{2}}(\hat{x} - x_i)\|^2 - (f_i, p_i)^2 / \|A^{\frac{1}{2}} p_i\|^2 +$$

$$- 2(f_i - \hat{a}_i A p_i, \delta x_{i+1}) + \|A^{\frac{1}{2}} \delta x_{i+1}\|^2 ,$$

which leads to the basic formula

(45) $$\frac{\|A^{\frac{1}{2}}(\hat{x} - x_{i+1})\|^2}{\|A^{\frac{1}{2}}(\hat{x} - x_i)\|^2} = 1 - \gamma_i^2 + \eta_{i+1} ,$$

where γ_i is defined by (20) and

(46) $$\eta_{i+1} := \{2(\hat{a}_i A p_i - f_i, \delta x_{i+1}) + \|A^{\frac{1}{2}} \delta x_{i+1}\|^2\} / \|A^{-\frac{1}{2}} f_i\|^2 .$$

It remains to be proved that η_{i+1} satisfies (23) under the restriction (24). Note that $(f_i - \hat{a}_i A p_i, p_i) = 0$ and therefore the term $\delta\bar{a}_i p_i$ in (39) cancels when evaluating the inner product in the numerator of (46). Consequently, from (33), (39) and (41) we obtain, evaluating term by term

(47) $$(f_i - \hat{a}_i A p_i, \delta x_{i+1}) / \|A^{-\frac{1}{2}} f_i\|^2 \leq$$

$$\leq \{\|f_i\| \|F_i'''\| \|x_i\| + |\hat{a}_i| \|A p_i\| \|F_i'''\| \|x_i\| + \|f_i\| \|V_i\| \|p_i\| +$$

$$+ |\hat{a}_i| \|A p_i\| \|V_i\| \|p_i\|\} / \|A^{-\frac{1}{2}} f_i\|^2 \leq$$

$$\leq 2\|F_i'''\| \psi_i + 2\|V_i\| \kappa^{\frac{1}{2}} \|A^{\frac{1}{2}} p_i\| / \|A^{-\frac{1}{2}} f_i\| \leq 2\varepsilon(\psi_i + 2\kappa^{\frac{1}{2}}(1 + o(1)))$$

under the restriction (31).

For the second order term in (46) we obtain from (33), (36), (39) and (41)

(48) $$\|A^{\frac{1}{2}}\delta x_{i+1}\|^2 / \|A^{-\frac{1}{2}} f_i\|^2 \leq$$

$$\leq 4\{\|F_i'''\|^2\|A\|\|x_i\|^2 + (\delta\bar{a}_i)^2\|A^{\frac{1}{2}}p_i\|^2 + \|V_i\|^2\|A\|\|p_i\|^2\} / \|A^{-\frac{1}{2}} f_i\|^2 \leq$$

$$\leq 4\{\varepsilon^2\psi_i^2 + ((\delta\bar{a}_i)^2 + \|V_i\|^2\kappa)\|A^{\frac{1}{2}}p_i\|^2 / \|A^{-\frac{1}{2}} f_i\|^2\} \leq$$

$$\leq 4\{\varepsilon^2\psi_i^2 + 8(\varepsilon^2 + \varepsilon^2\kappa + 4\varepsilon^2C_2^2\kappa + \varepsilon^2C_1^2\kappa^2 + \varepsilon^2C_1^2\kappa\psi_i^2) + 4\varepsilon^2\}(1 + o(1))$$

under the restriction (31).
So, finally, we obtain from (46), (47) and (48)

(49) $$|\eta_{i+1}| \leq 4\varepsilon(2\kappa^{\frac{1}{2}} + \psi_i)(1 + o(1)) +$$

$$+ 4\varepsilon^2(12 + 8\kappa + 32C_2^2\kappa + 8C_1^2\kappa^2 + \psi_i^2 + 8C_1^2\kappa\psi_i^2)(1 + o(1))$$

under the restriction (31).
As this inequality can be written in the more compact form of (23), under the restriction (24), we have proven theorem 3. □

Note that both constants C_1 and C_2 do not occur in the first order part of (23). In the error analysis these constants only appear in the absolute error $\delta\bar{a}_i$ and, as is explained for the recursive residual cases, this absolute error does not appear in the first order part of (47).

REMARK 4. Theorem 3 can also be written in a form closer related to expression (ii) of theorem 2.2.2. We have

$$\gamma_i^2 = (f_i,p_i)^2 / (\|A^{-\frac{1}{2}} f_i\|\|A^{\frac{1}{2}}p_i\|)^2 = \|\hat{a}_i A^{\frac{1}{2}}p_i\|^2 / \|A^{-\frac{1}{2}} f_i\|^2 ,$$

where, according to (33), $\hat{a}_i = (f_i,p_i) / (p_i,Ap_i)$.
Consequently, (22) can be written as

$$\|A^{\frac{1}{2}}(\hat{x} - x_{i+1})\|^2 + \|\hat{a}_i A^{\frac{1}{2}}p_i\|^2 = (1 + \eta_{i+1})\|A^{\frac{1}{2}}(\hat{x} - x_i)\|^2 .$$

It follows in particular that, under the restriction (24),

$$\|A^{\frac{1}{2}}(\hat{x} - x_{i+1})\| \leq (1 + o(1))\|A^{\frac{1}{2}}(\hat{x} - x_i)\| ,$$

$$\|\hat{a}_i A^{\frac{1}{2}}p_i\| \leq (1 + o(1))\|A^{\frac{1}{2}}(\hat{x} - x_i)\| .$$

Using (32) and (37) this yields, under the restriction (24),

$$\|a_i A^{\frac{1}{2}} p_i\| \leq (1 + o(1))\|A^{\frac{1}{2}}(\hat{x} - x_i)\| .$$

Retracing the proof of theorem 3 and replacing all o-symbols by definite estimates involving explicit numerical constants one can prove that $|\eta_{i+1}| \leq 22/40$ if

$$\varepsilon\kappa^{\frac{1}{2}}(2 + C_2 + C_1\kappa^{\frac{1}{2}}) \leq \frac{1}{40} ,$$

and

$$\varepsilon(1 + C_1\kappa^{\frac{1}{2}})\psi_i \leq \frac{1}{16} .$$

Hence

$$\|A^{\frac{1}{2}}(\hat{x} - x_{i+1})\| \leq (1 + \frac{3}{10})\|A^{\frac{1}{2}}(\hat{x} - x_i)\| ,$$

$$\|\hat{a}_i A^{\frac{1}{2}} p_i\| \leq (1 + \frac{3}{10})\|A^{\frac{1}{2}}(\hat{x} - x_i)\| .$$

Furthermore, it follows that

$$\|a_i A^{\frac{1}{2}} p_i\| \leq (1 + \frac{4}{10})\|A^{\frac{1}{2}}(\hat{x} - x_i)\| . \qquad \square$$

REMARK 5. It is obvious from (22) that the natural error $\|A^{\frac{1}{2}}(\hat{x} - x_i)\|$ cannot increase more than a factor $(1 + |\eta_{i+1}|)^{\frac{1}{2}}$ at the step from i to i + 1 (cf. remark 2.3.1.3). In view of (23) and (24), η_{i+1} is small if

$$(50) \qquad \varepsilon\{\kappa^{\frac{1}{2}}(1 + C_2 + C_1\kappa^{\frac{1}{2}}) + (1 + C_1\kappa^{\frac{1}{2}})\psi_i\}$$

is small. The second part of (50) depends on ψ_i and consequently on x_i, whereas the first part only depends on the machine, the implementation and the matrix involved. The restriction $\varepsilon(1 + C_1\kappa)\psi_i \leq 1/16$ (say) is satisfied iff $\|A^{\frac{1}{2}}(\hat{x} - x_i)\| \geq 16\varepsilon(1 + C_1\kappa^{\frac{1}{2}})\|A^{\frac{1}{2}}\|\|x_i\|$ and consequently, at the step where $\varepsilon(1 + C_1\kappa^{\frac{1}{2}})\psi_i \leq 1/16$ is not satisfied (assuming that C_1 does not depend on κ), the error $\|A^{\frac{1}{2}}(\hat{x} - x_i)\|$ is of the order of magnitude of the inherent natural error (cf. section 1.3). Therefore, at that moment we might as well stop the iterative process. $\square$

In order to obtain results concerning the monotonicity of the natural error $\|A^{\frac{1}{2}}(\hat{x} - x_i)\|$ the next theorem (where we estimate relative errors)

is more appropriate than the previous theorem. The proof is very similar to the proof of theorem 2.3.1.4.

THEOREM 6. *Let* x_i, p_i *be two arbitrary machine vectors* ($p_i \neq 0$) *and let* x_{i+1} *be computed from one step TRDM, based on these vectors. Let* γ_i, $\hat{r}_i$ *be defined as in theorem 3 and let, according to (2.2.16), (2.2.17) and (8) the numbers* α_i, β_i, φ_i *be defined by*

$$\alpha_i := \|\hat{r}_i\|\|p_i\| \,/\, |(\hat{r}_i, p_i)| \,, \tag{51}$$

$$\beta_i := \|\hat{r}_i\|\|A^{\frac{1}{2}} p_i\| \,/\, (\|A^{\frac{1}{2}}\| \, |(\hat{r}_i, p_i)|) \,, \tag{52}$$

$$\varphi_i := \|A\|\|x_i\| \,/\, \|\hat{r}_i\| \,. \tag{53}$$

Then we have

$$\frac{\|A^{\frac{1}{2}}(\hat{x} - x_{i+1})\|^2}{\|A^{\frac{1}{2}}(\hat{x} - x_i)\|^2} = 1 - \gamma_i^2(1 + \nu_{i+1}) \,, \tag{54}$$

where

$$|\nu_{i+1}| \leq 2\varepsilon\{2\kappa^{\frac{1}{2}} + 2\alpha_i + \beta_i(1+\beta_i)\varphi_i\}(1 + o(1)) + \\ + \varepsilon\{(C_2\kappa^{\frac{1}{2}} + C_1\kappa) + \alpha_i C_2 + C_1\alpha_i\varphi_i\}o(1) \,, \tag{55}$$

under the restriction

$$\varepsilon\{\kappa^{\frac{1}{2}}(1 + C_2 + C_1\kappa^{\frac{1}{2}}) + \alpha_i(1 + C_2) + (\beta_i + C_1\alpha_i)\varphi_i\} \to 0 \,. \tag{56}$$

PROOF. From (9) and (10) we know that the computed vector r_i satisfies

$$r_i = \hat{r}_i + \delta r_i \,, \tag{57}$$

$$\|\delta r_i\| \,/\, \|\hat{r}_i\| \leq \varepsilon(1 + C_1\varphi_i(1 + o(1))) \,, \tag{58}$$

under the restriction $\varepsilon \to 0$.
If we define $\alpha_i' := \|r_i\|\|p_i\| \,/\, |(r_i, p_i)|$, then we may conclude (cf. theorem 2.3.1.4)

$$a_i = \mathrm{fl}\left(\frac{\mathrm{fl}((r_i, p_i))}{\mathrm{fl}((p_i, Ap_i))}\right) = \frac{(r_i, p_i)}{(p_i, Ap_i)}(1 + \delta a_i'') \,, \tag{59}$$

(60) $$|\delta a_i''| \le (\varepsilon C_2 \alpha_i' + \varepsilon C_2 \kappa^{\frac{1}{2}} + \varepsilon C_1 \kappa + \varepsilon)(1 + o(1)) ,$$

under the restriction

(61) $$\varepsilon(1 + C_2 \kappa^{\frac{1}{2}} + C_1 \kappa) \to 0 .$$

However, we want to have expressions in terms of α_i defined by (51). From (57) and (58) we obtain under the restriction $\varepsilon \to 0$,

(62) $$(r_i, p_i) = (\hat{f}_i, p_i)(1 + \tau_i) ,$$

(63) $$|\tau_i| = |(\delta r_i, p_i)| / |(\hat{f}_i, p_i)| \le (\|\delta r_i\| / \|\hat{f}_i\|)\alpha_i \le$$

$$\le \varepsilon\alpha_i + \varepsilon C_1 \alpha_i \varphi_i (1 + o(1)) ,$$

and hence $|\tau_i| = o(1)$ under the restriction

(64) $$\varepsilon\alpha_i + \varepsilon C_1 \alpha_i \varphi_i \to 0 .$$

(Note that $\alpha_i \ge 1$ and hence restriction (64) implies the restriction $\varepsilon \to 0$.)
In view of this and taking into account lemma 1 we find

(65) $$\alpha_i' = \alpha_i (1 + o(1))$$

under the restriction (64). Substitution of (62) into (59) yields

(66) $$a_i = \hat{a}_i (1 + \delta\tilde{a}_i) ,$$

(67) $$\hat{a}_i := (\hat{f}_i, p_i) / \|A^{\frac{1}{2}} p_i\|^2 ,$$

(68) $$\delta\tilde{a}_i := \tau_i (1 + \delta a_i'') + \delta a_i'' .$$

From (60), (61), (64) and (65) we obtain

(70) $$|\delta\tilde{a}_i| \le \varepsilon\{(1 + C_2 \kappa^{\frac{1}{2}} + C_1 \kappa) + \alpha_i (1 + C_1) + C_1 \alpha_i \varphi_i\}(1 + o(1)) ,$$

under the restriction

(71) $$\varepsilon\{\kappa^{\frac{1}{2}}(C_2 + C_1 \kappa^{\frac{1}{2}}) + \alpha_i + C_1 \alpha_i \varphi_i\} \to 0 .$$

The vector δx_{i+1} in (39) can be written as

(72) $$\delta x_{i+1} = F_i''' x_i + \hat{a}_i(\delta\tilde{a}_i p_i + M_i p_i) ,$$

(73) $$M_i := \hat{a}_i^{-1} V_i = (1 + \delta\tilde{a}_i)(F_i'' + F_i'''(I + F_i'')) ,$$

(74) $$\|M_i\| \leq 2\varepsilon(1 + o(1)) ,$$

under the restriction

(75) $$\varepsilon\{\kappa^{\frac{1}{2}}(C_2 + C_1\kappa^{\frac{1}{2}}) + \alpha_i(1 + C_2) + C_1\alpha_i\varphi_i\} \to 0 .$$

Instead of transforming (44) into (45) we may also write

(76) $$\frac{\|A^{\frac{1}{2}}(\hat{x} - x_{i+1})\|^2}{\|A^{\frac{1}{2}}(\hat{x} - x_i)\|^2} = 1 - \gamma_i^2(1 + \nu_{i+1}) ,$$

where γ_i is defined as in theorem 3 and

(77) $$\nu_{i+1} := \{2(\hat{f}_i - \hat{a}_i Ap_i, \delta x_{i+1}) - \|A^{\frac{1}{2}} \delta x_{i+1}\|^2\} / \{\hat{a}_i(\hat{f}_i, p_i)\} .$$

Analogously to (47) we find, evaluating term by term, that

(78) $$|(\hat{f}_i - \hat{a}_i Ap_i, \delta x_{i+1})| / |\hat{a}_i(\hat{f}_i, p_i)| \leq$$

$$\leq \|\hat{f}_i\|\|F_i'''\|\|x_i\|\|A^{\frac{1}{2}} p_i\|^2 / (\hat{f}_i, p_i)^2 + \|Ap_i\|\|F_i'''\|\|x_i\| / |(\hat{f}_i, p_i)| +$$

$$+ \|\hat{f}_i\|\|M_i\|\|p_i\| / |(\hat{f}_i, p_i)| + \|Ap_i\|\|M_i\|\|p_i\| / \|A^{\frac{1}{2}} p_i\|^2 \leq$$

$$\leq \|F_i'''\|\beta_i^2\varphi_i + \|F_i'''\|\beta_i\varphi_i + \|M_i\|\alpha_i + \|M_i\|\kappa^{\frac{1}{2}} \leq$$

$$\leq \varepsilon\beta_i^2\varphi_i + \varepsilon\beta_i\varphi_i + 2\varepsilon\alpha_i(1 + o(1)) + 2\varepsilon\kappa^{\frac{1}{2}}(1 + o(1)) ,$$

under the restriction (75).
Similarly to (48) we obtain, evaluating term by term, that there holds

(79) $$\|A^{\frac{1}{2}} \delta x_{i+1}\|^2 / |\hat{a}_i(\hat{f}_i, p_i)| \leq$$

$$\leq 4\{\|F_i'''\|^2\|A\|\|x_i\|^2 \|A^{\frac{1}{2}} p_i\|^2 / (\hat{f}_i, p_i)^2 + (\delta\tilde{a}_i)^2 +$$

$$+ \|M_i\|^2\|A\|\|p_i\|^2 / \|A^{\frac{1}{2}} p_i\|^2\} \leq$$

$$\leq 4\{\varepsilon^2\varphi_i^2\beta_i^2 + 4\varepsilon^2(4(1 + C_2^2\kappa + C_1^2\kappa^2) +$$

$$+ 2\alpha_i^2(1+C_2^2) + C_1^2\alpha_i^2\varphi_i^2) + 4\varepsilon^2\kappa\}(1+o(1)) ,$$

under the restriction (75).

So, finally, we obtain from (77), (78) and (79)

$$(80) \qquad |\nu_{i+1}| \leq 2\varepsilon\{2\kappa^{\frac{1}{2}} + 2\alpha_i + \beta_i(1+\beta_i)\varphi_i\}(1+o(1)) +$$

$$+ 4\varepsilon^2\{16(1+C_2^2\kappa+C_1^2\kappa^2) + 4\kappa + 8\alpha_i^2(1+C_2^2) +$$

$$+ (\beta_i^2 + 4C_1^2\alpha_i^2)\varphi_i^2\}(1+o(1)) ,$$

under the restriction (75).

As this inequality can be written in the more compact form of (55), under the restriction (56), we have proved theorem 6. □

Note that remark 2.3.1.5 and remark 2.3.1.7, concerning the main theorems of section 2.3.1, also apply to theorem 3 and theorem 6.

REMARK 7. Just like in theorem 3 the constants C_1 and C_2 do not occur in the first order part of (55). Formulae (55) and (56) indicate that $\varepsilon(C_2\kappa^{\frac{1}{2}} + C_1\kappa) + \alpha_i C_2 + (\beta_i + C_1\alpha_i)\varphi_i$ has to be small in order to have ν_{i+1} small. The following simple straightforward one-round-off error analysis stresses this result for the term $C_1\alpha_i\varphi_i$. Suppose that during the step from x_i to x_{i+1} only round-off occurs at the computation of $A * x_i$. Then, retracing the proof of theorem 1.4.6, we obtain successively

$$(81) \qquad \delta r_i = -E_i x_i , \quad \|\delta r_i\| / \|f_i\| < \varepsilon C_1\varphi_i , \quad \delta a_i'' = 0 ,$$

$$|\delta\tilde{a}_i| = |\tau_i| \leq \varepsilon C_1\alpha_i\varphi_i , \quad \delta x_{i+1} = \tilde{a}_i\,\delta\tilde{a}_i\,p_i ,$$

$$(82) \qquad \nu_{i+1} = -\|A^{\frac{1}{2}}\delta x_{i+1}\|^2 / \{\tilde{a}_i(f_i,p_i)\} = -(\delta\tilde{a}_i)^2 .$$

Consequently, $|\nu_{i+1}| \leq (\varepsilon C_1\alpha_i\varphi_i)^2$, which implies $|\nu_{i+1}| \leq \varepsilon C_1\alpha_i\varphi_i\, o(1)$ under the restriction $\varepsilon C_1\alpha_i\varphi_i \to 0$. The other terms in (55) and (56) containing C_1 are present due to the computation of $A * p_i$. The above derivation also indicates that the bound $|\nu_{i+1}| \leq (\varepsilon C_1\alpha_i\varphi_i)^2$ is sharp. □

From (55) and (56) it follows that $|\nu_{i+1}| = o(1)$ under the restriction

$$\varepsilon\{\kappa^{\frac{1}{2}}(1 + C_2 + C_1\kappa^{\frac{1}{2}}) + \alpha_i(1 + C_2) + (\beta_i + \beta_i^2 + C_1\alpha_i)\varphi_i\} \to 0 \ . \tag{83}$$

As we mentioned already in remarks 2.3.1.10 and 2.3.1.11, algebraically α_i, β_i and γ_i are uniformly bounded for all $i \geq 0$ for the Gauss-Southwell method, the gradient method and the conjugate gradient method. We shall see that this fact enables us to prove the boundedness of these parameters in the presence of round-off as long as $\varepsilon C_1\varphi_i$ is of order 1, i.e., the residual $\|\hat{r}_i\|$ is not less than something of the order $\varepsilon C_1\|A\|\|x_i\|$. (Note that algebraically $\varphi_i \to \infty$ $(i \to 0)$.)
For instance for the gradient method (cf. (2.2)) we have, algebraically, $\alpha_i = 1$. In the floating point case one takes $p_i = r_i$ (the computed residual) and hence in that case it follows from lemma 1 that

$$\alpha_i := \|\hat{r}_i\|\|r_i\| \,/\, |(\hat{r}_i, r_i)| = 1\cdot(1 + o(1)) \ , \tag{84}$$

under the restriction $\varepsilon(1 + C_1\varphi_i) \to 0$.
Stated differently, as long as $\varepsilon C_1\varphi_i$ is small the parameter α_i, corresponding to the process performed in floating point arithmetic, approximately equals the α_i of the algebraic process.
This is one reason why (55) has only significance as long as $\varepsilon C_1\varphi_i$ is of order 1. Another reason, of course, is the appearance of φ_i in (55) and (56) itself which even requires $\varepsilon(\beta_i + \beta_i^2 + C_1\alpha_i)\varphi_i$ to be small.
In corollary 2.3.1.8, which is a direct consequence of theorem 2.3.1.4, we assumed uniform bounds for α_i, β_i and γ_i. From the previous considerations it will be clear that it is unrealistic to assume this for TRDM in the presence of round-off and therefore an analogue of corollary 2.3.1.8 is omitted.
However, we are mainly interested in the monotonicity of the natural error $\|A^{\frac{1}{2}}(\hat{x} - x_i)\|$. From theorem 6 it follows that the natural error decreases at the step from i to i + 1 iff $\nu_{i+1} > -1$ in (54). Apparently, from (55) and (56), $|\nu_{i+1}| < 1$ if (83) is small. As we noted already in the case of RRDM, these formulae do not supply an explicit bound for (83) in order to guarantee $|\nu_{i+1}| < 1$, owing to the use of the o-notation. Therefore we now first turn to an explicit version of theorem 6. Since theorem 6 is one of our basic theorems we give a full proof of this explicit version, although there is nothing new in it.
We shall show that under the assumptions (97) and (98) of proposition

8 (to be stated presently) there certainly holds $|\nu_{i+1}| < 4/5$ in theorem 6.

We shall follow the lines of the proof of theorem 6, replacing the o-symbols by numerical constants. Proceeding in this way we first obtain from lemma 1, instead of (58),

$$(85) \qquad \|\delta r_i\| / \|\hat{r}_i\| \leq \varepsilon + \varepsilon(1+\varepsilon)C_1\varphi_i \leq (\tfrac{41}{40})(\varepsilon + \varepsilon C_1\varphi_i) \leq$$

$$\leq (\tfrac{41}{40})(\tfrac{1}{40} + \tfrac{1}{4}) < \frac{3}{10} .$$

From the proof of theorem 2.3.1.4 we find for $\delta a_i''$

$$(86) \qquad |\delta a_i''| \leq (|\lambda_i| + |\mu_i| + |\varepsilon_i| + |\lambda_i\varepsilon_i|) / (1 - |\mu_i|) ,$$

$$(87) \qquad |\lambda_i| \leq \varepsilon C_2 \|r_i\| \|p_i\| / |(r_i, p_i)| ,$$

$$(88) \qquad |\mu_i| \leq \varepsilon C_2 \kappa^{\frac{1}{2}} + \varepsilon C_1 \kappa (1 + \varepsilon C_2) \leq (\tfrac{41}{40})(\varepsilon C_2 \kappa^{\frac{1}{2}} + \varepsilon C_1 \kappa) \leq$$

$$\leq (\tfrac{41}{40})(\tfrac{1}{40}) < \frac{3}{100} ,$$

$$(89) \qquad |\varepsilon_i| \leq \varepsilon .$$

For τ_i we obtain, instead of (63),

$$(90) \qquad |\tau_i| \leq (\|\delta r_i\| / \|\hat{r}_i\|)\alpha_i \leq (\tfrac{41}{40})(\varepsilon\alpha_i + \varepsilon C_1 \alpha_i \varphi_i) < \frac{3}{10} ,$$

and consequently, using $\|r_i\| \leq \|\hat{r}_i\|(1 + \|\delta r_i\| / \|\hat{r}_i\|) \leq (13/10)\|\hat{r}_i\|$, one has

$$(91) \qquad |\lambda_i| \leq \varepsilon C_2 \alpha_i (\|r_i\| / \|\hat{r}_i\|)(|(\hat{r}_i, p_i)| / |(r_i, p_i)|) \leq$$

$$\leq \varepsilon C_2 \alpha_i (\tfrac{13}{10})(1 - \tfrac{3}{10})^{-1} < 2\varepsilon C_2 \alpha_i .$$

Hence, instead of (60), we obtain from (86), (88), (89) and (91)

$$(92) \qquad |\delta a_i''| \leq (\tfrac{100}{97})(2\varepsilon C_2 \alpha_i + (\tfrac{41}{40})(\varepsilon C_2 \kappa^{\frac{1}{2}} + \varepsilon C_1 \kappa) + \varepsilon + 2\varepsilon^2 C_2 \alpha_i) \leq$$

$$\leq 2(\tfrac{100}{97})(\tfrac{41}{40})(\varepsilon C_2 \alpha_i + \varepsilon C_2 \kappa^{\frac{1}{2}} + \varepsilon C_1 \kappa + \varepsilon) \leq$$

$$\leq 2(\tfrac{100}{97})(\tfrac{41}{40})(\tfrac{1}{40}) < \frac{6}{100} ,$$

and using (90) this yields

(93) $$|\delta\tilde{a}_i| \leq |\tau_i|(1+|\delta a_i''|) + |\delta a_i''| < (\frac{3}{10})(1+\frac{6}{100}) + \frac{6}{100} =$$

$$= (\frac{378}{1000}) < \frac{4}{10} .$$

Instead of (74) we find from (93)

(94) $$\|M_i\| \leq (2\varepsilon+\varepsilon^2)(1+|\delta\tilde{a}_i|) < (\frac{14}{10})(2+\frac{1}{40})\varepsilon < 3\varepsilon .$$

So, finally, instead of (78), one has

(95) $$|(\bar{r}_i - \hat{a}_i A p_i, \delta x_{i+1})| / |\hat{a}_i(\bar{r}_i, p_i)| \leq$$

$$\leq \varepsilon\beta_i^2\varphi_i + \varepsilon\beta_i\varphi_i + 3\varepsilon\alpha_i + 3\varepsilon\kappa^{\frac{1}{2}} \leq (\frac{1}{8}) + (\frac{3}{40}) < \frac{2}{10} ,$$

and, by a slightly different estimate, we obtain instead of (79)

(96) $$\|A^{\frac{1}{2}}\delta x_{i+1}\|^2 / |\hat{a}_i(\bar{r}_i, p_i)| \leq 2(\delta\tilde{a}_i)^2 + 4(\varepsilon^2\beta_i^2\varphi_i^2 + 9\varepsilon^2\kappa) \leq$$

$$\leq 2(\frac{378}{1000})^2 + 4(\frac{1}{8})^2 + 36(\frac{1}{40})^2 < \frac{4}{10} .$$

Hence

$$|\nu_{i+1}| < 2(\frac{2}{10}) + \frac{4}{10} = \frac{4}{5} .$$

Thus, as a more explicit version of theorem 6 we obtain

PROPOSITION 8. *Let* x_i, p_i *be two machine vectors* ($p_i \neq 0$) *and let* x_{i+1} *be computed from one step TRDM, based on these vectors. Let* γ_i, α_i, β_i, $\bar{r}_i$ *and* φ_i *be defined as in theorem 6. Furthermore, let*

(97) $$\varepsilon\{\kappa^{\frac{1}{2}}(1+C_2+C_1\kappa^{\frac{1}{2}}) + \alpha_i(1+C_2)\} \leq \frac{1}{40} ,$$

and

(98) $$\varepsilon\{2\beta_i(1+\beta_i)\varphi_i + C_1\alpha_i\varphi_i\} \leq \frac{1}{4} ,$$

Then we have

(99) $$\frac{\|A^{\frac{1}{2}}(\bar{x}-x_{i+1})\|^2}{\|A^{\frac{1}{2}}(\bar{x}-x_i)\|^2} \leq 1 - \frac{1}{5}\gamma_i^2 .$$ □

REMARK 9. One may ask whether it is possible to verify the monotonicity of $\|A^{\frac{1}{2}}(\hat{x} - x_i)\|$ by some computation. It is obvious that the objective function $F(x_i) = (\hat{x} - x_i, A(\hat{x} - x_i)) = \|A^{\frac{1}{2}}(\hat{x} - x_i)\|^2$ cannot be computed, since the solution vector $\hat{x}$ is not known. On the other hand, however, we have, algebraically,

$$(100) \qquad F(x_i) - F(x_{i+1}) = (\hat{x} - x_i, A(\hat{x} - x_i)) - (\hat{x} - x_{i+1}, A(\hat{x} - x_{i+1})) =$$

$$= (x_i, Ax_i) - 2(x_i, A\hat{x}) - (x_{i+1}, Ax_{i+1}) + 2(x_{i+1}, A\hat{x}) =$$

$$= (x_i - x_{i+1}, A(x_i - x_{i+1})) + 2(x_{i+1}, Ax_i) - 2(x_{i+1}, Ax_{i+1}) +$$

$$- 2(b, x_i - x_{i+1}) .$$

Hence,

$$(101) \qquad G(x_i, x_{i+1}) := (x_i - x_{i+1}, A(x_i - x_{i+1})) - 2(b - Ax_{i+1}, x_i - x_{i+1}) =$$

$$= F(x_i) - F(x_{i+1}) .$$

This function can be computed and algebraically $F(x_i) > F(x_{i+1})$ is equivalent with $G(x_i, x_{i+1}) > 0$. Now the important question rises what can be said if $fl(G(x_i, x_{i+1})) < 0$. To illustrate the problem involved, we perform a one-round-off error analysis, where we assume that during the computation of $G(x_i, x_{i+1})$, defined by the expression in the middle of (101), only round-off occurs at the computation of $A * x_{i+1}$ (and not at the computation of $A(x_i - x_{i+1})$). We then have

$$(102) \qquad fl(G(x_i, x_{i+1})) = G(x_i, x_{i+1}) + \delta G(x_i, x_{i+1}) ,$$

$$(103) \qquad \delta G(x_i, x_{i+1}) = 2(E_i x_{i+1}, x_i - x_{i+1}) .$$

Consequently, from theorem 6, it follows that

$$(104) \qquad fl(G(x_i, x_{i+1})) = F(x_i) - F(x_{i+1}) + \delta G(x_i, x_{i+1}) =$$

$$= \frac{(f_i, p_i)^2}{\|A^{\frac{1}{2}} p_i\|^2} (1 + \nu_{i+1} + \theta_{i+1}) ,$$

where ν_{i+1} is estimated in (55) and

$$(105) \qquad |\theta_{i+1}| \leq 2|(E_i x_{i+1}, x_i - x_{i+1})| \|A^{\frac{1}{2}} p_i\|^2 / (f_i, p_i)^2 \leq$$

$$\leq 2\varepsilon C_1 \|A\|^2 \|x_{i+1}\| \|x_i - x_{i+1}\| \beta_i^2 / \|f_i\|^2 .$$

Hence, if $fl(G(x_i,x_{i+1})) < 0$ at a certain step, then at least one of the two inequalities $|\nu_{i+1}| > \frac{1}{2}$, $|\theta_{i+1}| > \frac{1}{2}$ holds. If for that particular step (97) is satisfied, then $|\nu_{i+1}| \geq \frac{1}{2}$ leads to the conclusion that

$$\|f_i\| \leq 8\varepsilon(2\beta_i(1+\beta_i) + C_1\alpha_i)\|A\|\|x_i\| , \tag{106}$$

which qualitatively is the same conclusion as one would obtain when break down of the monotonicity of $\|A^{\frac{1}{2}}(\hat{x} - x_i)\|$ could be verified (which would imply $|\nu_{i+1}| > 1$). On the other hand, $|\theta_{i+1}| \geq \frac{1}{2}$ implies

$$\|f_i\|^2 \leq 2\varepsilon\beta_i^2 C_1 \|A\|^2 \|x_{i+1}\| \|x_i - x_{i+1}\| . \tag{107}$$

Algebraically we have (cf. theorem 2.2.2)

$$\|x_i - x_{i+1}\| = \|a_i p_i\| \leq \|A^{-\frac{1}{2}}\| \|a_i A^{\frac{1}{2}} p_i\| \leq \|A^{-\frac{1}{2}}\| \|A^{-\frac{1}{2}} r_i\| \leq$$

$$\leq \|A^{-1}\| \|r_i\| .$$

Hence, from (107) we obtain no better estimate than

$$\|f_i\| \leq 2\varepsilon\beta_i^2 C_1 \kappa \|A\| \|x_{i+1}\| ,$$

which is unsatisfactorily.
Note that algebraically also $x_{i+1} = x_i + a_i p_i$ and hence

$$G(x_i,x_{i+1}) = (a_i p_i, a_i Ap_i) + 2(r_{i+1}, a_i p_i) = \tag{108}$$

$$= a_i(a_i(p_i,Ap_i) + 2(r_{i+1},p_i)) ,$$

and therefore $G(x_i,x_{i+1})$ could easily be computed from the (already computed) number a_i and vectors p_i, Ap_i, r_{i+1}. However, in the presence of round-off, even if the right-hand side of (108) is computed exactly, the same kind of a problem, as reflected in (107) arises, due to round-off at the computation of r_{i+1}. □

We conclude this section by showing how the theory developed in this section leads to assertions on the numerical behavior of a specific DM, the Gauss-Southwell method.

EXAMPLE. *The true residual Gauss-Southwell method (TRGSM)*

Recall that in TRGSM one takes $p_i := e_{k(i)}$, where k(i) corresponds to the largest (in absolute value) component of the computed residual $r_i = fl(b - Ax_i)$, i.e., $|(r_i,e_{k(i)})| \geq |(r_i,e_j)|$ for all $1 \leq j \leq n$. For application of our theory we need bounds for α_i, β_i and γ_i, as defined by (51), (52) and (20). Algebraically, $\alpha_i \leq n^{\frac{1}{2}}$ (cf. section 2.2). Using floating point arithmetic this bound is affected by round-off. We have $\alpha_i := \|\hat{r}_i\|\|e_{k(i)}\| / |(\hat{r}_i,e_{k(i)})|$. From the proof of lemma 1 we know that under the restriction $\varepsilon \to 0$

$$\hat{r}_i = r_i + \delta r_i \ , \quad \|\delta r_i\| / \|r_i\| \leq \varepsilon(1 + C_1\varphi_i(1 + o(1))) \ . \tag{109}$$

Hence,

$$(\hat{r}_i,e_{k(i)}) = (r_i,e_{k(i)})(1 + \xi_i) \ ,$$

where

$$|\xi_i| = \left|\frac{(\delta r_i,e_{k(i)})}{(r_i,e_{k(i)})}\right| \leq \frac{\|\delta r_i\|}{\|r_i\|}\frac{\|\hat{r}_i\|}{\|r_i\|}\frac{\|r_i\|}{|(r_i,e_{k(i)})|} \leq \tag{110}$$

$$\leq \varepsilon(1 + C_1\varphi_i)\frac{\|r_i\|}{|(r_i,e_{k(i)})|}(1 + o(1)) \leq$$

$$\leq \varepsilon n^{\frac{1}{2}}(1 + C_1\varphi_i)(1 + o(1)) \ ,$$

under the restriction

$$\varepsilon(1 + C_1\varphi_i) \to 0 \ . \tag{111}$$

Together with (110) we find $|\xi_i| = o(1)$ under the restriction $\varepsilon n^{\frac{1}{2}}(1 + C_1\varphi_i) \to 0$. So, finally, we obtain

$$\alpha_i = \frac{\|\hat{r}_i\|}{|(\hat{r}_i,e_{k(i)})|} = \frac{\|r_i\|}{|(r_i,e_{k(i)})|}\frac{\|\hat{r}_i\|}{\|r_i\|}\frac{|(r_i,e_{k(i)})|}{|(\hat{r}_i,e_{k(i)})|} \leq \tag{112}$$

$$\leq n^{\frac{1}{2}}(1 + o(1)) \ ,$$

under the restriction $\varepsilon n^{\frac{1}{2}}(1 + C_1\varphi_i) \to 0$.

Together with lemma 2.2.6 it follows that $\beta_i \leq n^{\frac{1}{2}}(1 + o(1))$ and $\gamma_i^{-1} \leq n^{\frac{1}{2}}(1 + o(1))$ under the same restriction.

Substitution in (55) and (56) yields for the parameter ν_{i+1} of (54) the estimate

(113) $$|\nu_{i+1}| \le 2\varepsilon\{2\kappa^{\frac12} + 2n^{\frac12} + (n^{\frac12}+n)\varphi_i\}(1+o(1)) +$$

$$+ \varepsilon\{(C_2\kappa^{\frac12} + C_1\kappa) + n^{\frac12}C_2 + C_1 n^{\frac12}\varphi_i\}o(1) ,$$

under the restriction

(114) $$\varepsilon\{\kappa^{\frac12}(1 + C_2 + C_1\kappa^{\frac12}) + n^{\frac12}(1+C_2) + n^{\frac12}(1+C_1)\varphi_i\} \to 0 .$$

Retracing the proof of (112) one finds that, if $\varepsilon n^{\frac12} \le 1/40$ and $\varepsilon n^{\frac12} C_1 \varphi_i \le 1/4$, then $\alpha_i \le 3n^{\frac12}$ and consequently $\beta_i \le 3n^{\frac12}$, $\gamma_i^{-1} \le 3(n\kappa)^{\frac12}$. Combining this and proposition 8 we obtain the following explicit statement for TRGSM.

PROPOSITION 10. *If x_{i+1} is computed for one step TRGSM, based on the machine vector x_i and if*

(115) $$\varepsilon\{\kappa^{\frac12}(1 + C_2 + C_1\kappa^{\frac12}) + 3n^{\frac12}(1+C_2)\} \le \frac{1}{40} ,$$

(116) $$3\varepsilon\{n^{\frac12}(2 + C_1) + 6n\}\varphi_i \le \frac14 ,$$

where $\varphi_i := \|A\|\|x_i\| / \|b - Ax_i\|$, then we have

(117) $$\frac{\|A^{\frac12}(\hat{x} - x_{i+1})\|^2}{\|A^{\frac12}(\hat{x} - x_i)\|^2} \le 1 - \frac{1}{45n\kappa} .$$ □

This leads to the following three important conclusions on TRGSM (if (115) is satisfied):

(i) If $\|A^{\frac12}(\hat{x} - x_{i+1})\| \ge \|A^{\frac12}(\hat{x} - x_i)\|$ holds for some $i \ge 0$, then $\|b - Ax_i\| \le 12\varepsilon(n^{\frac12}(2 + C_1) + 6n)\|A\|\|x_i\|$.

(ii) As long as $\|b - Ax_i\| \ge 12\varepsilon(n^{\frac12}(2 + C_1) + 6n)\|A\|\|x_i\|$, the natural error $\|A^{\frac12}(\hat{x} - x_i)\|$ converges step-wise linearly with a convergence ratio no greater than $(1 - (45n\kappa)^{-1})^{\frac12}$.

(iii) There exists an $i \ge 0$ such that $\|b - Ax_i\| < 12\varepsilon(n^{\frac12}(2 + C_1) + 6n)\|A\|\|x_i\|$. (Since otherwise (117) holds for all $i \ge 0$ which leads to the contradiction $\|b - Ax_i\| \to 0$ $(i \to \infty)$.)

Combining these three conclusions into one statement we obtain (if (115) is satisfied) the following result.

PROPOSITION 11. *If* $\{x_i\}$ *is generated by TRGSM with an arbitrary initial machine vector* x_0, *then the natural error* $\|A^{\frac{1}{2}}(\hat{x} - x_i)\|$ *converges stepwise linearly with a convergence ratio no greater than* $(1 - (15n\kappa)^{-1})^{\frac{1}{2}}$, *at least until the iteration step where the residual satisfies* $\|b - Ax_i\| \leq 12\varepsilon(n^{\frac{1}{2}}(2 + C_1) + 6n)\|A\|\|x_i\|$. □

This implies that TRGSM is well-behaved and, consequently, numerically stable and A-numerically stable (cf. section 1.4).

CHAPTER 3

THE GRADIENT METHOD (GM)

3.1. *Introduction*

One of the oldest and most widely known descent methods is the gradient method (often referred to as steepest descent method). A description of the method was first given by Cauchy in 1847. The application of the method is not restricted to the case where the objective function is quadratic. The objective function may be any differentiable function of several variables whose gradient is known explicitly. Therefore the method is also of great interest as a technique for nonlinear optimization problems. From a theoretical point of view the method is very important, since it is one of the simplest iterative methods for which a satisfactory analysis of the algebraic behavior exists. Many more advanced methods, like the conjugate gradient method, are often motivated by an attempt to modify the basic GM in such a way that the new method will have superior convergence properties.

As far as we know, Woźniakowski [80] is until now the only one who gave a complete round-off error analysis for the TRGM in order to obtain assertions on the numerical behavior of the TRGM. Our results on the numerical behavior of TRGM, derived in this chapter are superior to those of Woźniakowski in two aspects. Firstly, we prove step-wise linear convergence whereas Woźniakowski gives a result in terms of the limit superior. Secondly, we prove good-behavior, whereas Woźniakowski's result does not even imply numerical stability.

No published round-off error analysis of RRGM is known to us.

The GM is defined by the following statements.

Gradient method (GM)

Choose an initial point x_0;

$r_0 := b - Ax_0$; $i := 0$;

while $r_i \neq 0$ do

begin

(1) $\quad a_i := (r_i, r_i) / (r_i, Ar_i)$;

(2) $\quad x_{i+1} := x_i + a_i r_i$;

(3) $\quad r_{i+1} := \begin{cases} \text{either } b - Ax_{i+1} ; & \text{(TRGM)} \\ \text{or } \quad r_i - a_i Ar_i ; & \text{(RRGM)} \end{cases}$ (4)

$\quad i := i + 1$

end.

We use either (3) at all steps or (4) at all steps, and thus disregard the mixed gradient method (MGM) (cf. remark 2.2.1).
According to the definitions (2.2.16), (2.2.17) and (2.2.8) we have algebraically

$$(5) \qquad \begin{cases} \alpha_i = \|r_i\|^2 / |(r_i, r_i)| = 1 , \\ \beta_i = \|r_i\| \, \|A^{\frac{1}{2}} r_i\| / (\|A^{\frac{1}{2}}\| \, |(r_i, r_i)|) \leq 1 , \end{cases}$$

$$(6) \qquad \gamma_i = \|r_i\|^2 / (\|A^{-\frac{1}{2}} r_i\| \|A^{\frac{1}{2}} r_i\|) \geq 2\kappa^{\frac{1}{2}} / (\kappa + 1) .$$

From lemma 2.2.6 we also have $\beta_i \geq \kappa^{-\frac{1}{2}}$ and $\kappa^{-\frac{1}{2}} \leq \gamma_i \leq 1$. Consequently, corollary 2.2.5 yields

$$(7) \qquad \frac{\|A^{\frac{1}{2}}(\hat{x} - x_{i+1})\|^2}{\|A^{\frac{1}{2}}(\hat{x} - x_i)\|^2} = 1 - \gamma_i^2 \leq 1 - \frac{4\kappa}{(\kappa + 1)^2} = \left(\frac{\kappa - 1}{\kappa + 1}\right)^2 ,$$

which reflects the step-wise linear convergence to zero of the natural error with a convergence ratio no greater than $(\kappa - 1) / (\kappa + 1)$.

Another well-known algebraic property of the GM reads as follows.

THEOREM 1. *If* x_{i+1} *is computed from one step GM, based on* x_i *for which* $r_i = b - Ax_i \neq 0$, *then we have*

$$(8) \qquad \frac{\|\hat{x} - x_{i+1}\|^2}{\|\hat{x} - x_i\|^2} = \frac{\|A^{-1} r_{i+1}\|^2}{\|A^{-1} r_i\|^2} =$$

$$= 1 - \frac{\|r_i\|^2 \|A^{-\frac{1}{2}} r_i\|^2}{\|A^{\frac{1}{2}} r_i\|^2 \|A^{-1} r_i\|^2} \left(2 - \frac{\|r_i\|^4}{\|A^{-\frac{1}{2}} r_i\|^2 \|A^{\frac{1}{2}} r_i\|^2}\right) .$$

PROOF. If, in the equality $\hat{x} - x_{i+1} = \hat{x} - x_i - a_i r_i$, we take squared norms at both sides we obtain

$$(9) \qquad \|\hat{x} - x_{i+1}\|^2 = \|\hat{x} - x_i\|^2 - 2a_i(\hat{x} - x_i, r_i) + a_i^2 (r_i, r_i) .$$

Since $a_i = \|r_i\|^2 / \|A^{\frac{1}{2}} r_i\|^2$ and $\hat{x} - x_i = A^{-1} r_i$ we obtain (8) after some rearrangements. □

Since

$$(10) \qquad \frac{\|A^{\frac{1}{2}} r_i\|^2 \|A^{-1} r_i\|^2}{\|r_i\|^2 \|A^{-\frac{1}{2}} r_i\|^2} \leq \kappa , \qquad \frac{\|r_i\|^4}{\|A^{-\frac{1}{2}} r_i\|^2 \|A^{\frac{1}{2}} r_i\|^2} \leq 1 ,$$

the following corollary of theorem 1 is valid.

COROLLARY 2. *If* $\{x_i\}$ *is generated by the GM, then the error converges step-wise linearly to zero and*

$$(11) \qquad \frac{\|\hat{x} - x_{i+1}\|^2}{\|\hat{x} - x_i\|^2} = \frac{\|A^{-1} r_{i+1}\|^2}{\|A^{-1} r_i\|^2} \leq 1 - \frac{1}{\kappa} .$$ □

REMARK 3. The GM is invariant relative to orthogonal basis transformations. Let $\{x_i\}$, respectively $\{x_i'\}$ be generated by the GM corresponding to the systems $Ax = b$, $A'x' = b'$, respectively, with initial vectors x_0, x_0', respectively. If $A' = V^T A V$, $b' = V^T b$, $x_0' = V^T x_0$, where V is an orthogonal matrix, then $x_i' = V^T x_i$ for all $i \geq 0$. Since V is orthogonal this implies that also $\|A^\alpha(\hat{x} - x_i)\| = \|(A')^\alpha(\hat{x}' - x_i')\|$ $(\alpha = 0, \frac{1}{2}, 1)$. □

In the next two sections we investigate the validity of these properties if the GM is performed using floating point arithmetic. RRGM is studied in section 2; TRGM is studied in section 3. In section 4 we report on numerical experiments carried out with the GM.

3.2. *The recursive residual gradient method* (RRGM)

The results, deduced in this section, will be based on the results of section 2.3 for general RRDM's. The results obtained there are expressed in terms of the parameters α_i, β_i and γ_i defined by (2.3.1.43), (2.3.1.44) and (2.3.1.4). Therefore we have to estimate these parameters for RRGM. We arrive at exactly the same estimates (3.1.5) and (3.1.6) as in the algebraic case. The only difference is that now r_i stands for the recursively computed residual vector, whereas in the algebraic case $r_i = b - Ax_i$. Hence

$$(1) \qquad \alpha_i = 1 \ , \quad \kappa^{-\frac{1}{2}} \le \beta_i \le 1 \ , \quad \kappa^{-\frac{1}{2}} \le 2\kappa^{-\frac{1}{2}} / (\kappa+1) \le \gamma_i \le 1 \ .$$

As an immediate consequence of corollary 2.3.1.8 we obtain

PROPOSITION 1. *Let* $\{r_i\}$ *be computed by RRGM with an arbitrary initial machine vector* x_0, *then we have for* $i \ge 0$

$$(2) \qquad \frac{\|A^{-\frac{1}{2}} r_{i+1}\|^2}{\|A^{-\frac{1}{2}} r_i\|^2} \le 1 - \frac{4\kappa}{(\kappa+1)^2}(1+o(1)) \ ,$$

under the restriction

$$(3) \qquad \varepsilon\{\kappa^{\frac{1}{2}} C_2 + \kappa(1+C_1)\} \to 0 \ . \qquad \square$$

As a more explicit version we obtain from proposition 2.3.1.12

PROPOSITION 2. *Let* $\{r_i\}$ *be computed by RRDM with an arbitrary initial machine vector* x_0 *and let*

$$(4) \qquad \varepsilon\{(1+C_2) + \kappa^{\frac{1}{2}}(1+C_2) + \kappa(3+C_1)\} \le \frac{1}{8} \ ,$$

then $\{A^{-\frac{1}{2}} r_i\}$ *converges step-wise linearly to zero and for all* $i \ge 0$

one has

$$(5) \qquad \frac{\|A^{-\frac{1}{2}} r_{i+1}\|^2}{\|A^{-\frac{1}{2}} r_i\|^2} \leq 1 - \frac{5\kappa}{4(\kappa+1)^2} . \qquad \square$$

As we mentioned already in section 1.3, for many straightforward implementations there holds $C_1 \sim n^{3/2}$ and $C_1 \sim n$. Hence in the left-hand side of (4) the largest term is of order $n^{3/2}\kappa$.

As far as the computed sequence $\{x_i\}$ is concerned we only reformulate theorem 2.3.2.5 for RRGM.

PROPOSITION 3. *Let* $\{x_i\}$, $\{r_i\}$ *be computed by RRGM with an arbitrary initial machine vector* x_0 *and let* $N := \text{ent}(2\kappa \log 1/\varepsilon + 1)$, *then we have*

$$(6) \qquad \|A^{\frac{1}{2}}(\hat{x} - x_N)\| \leq$$

$$\leq \varepsilon\{1 + \kappa^{\frac{1}{2}} + C_1\kappa + 2(4\kappa + (2+C_1)\kappa^{3/2}) \log \tfrac{1}{\varepsilon}\} \|A^{\frac{1}{2}}(\hat{x} - x_0)\| (1 + o(1)) +$$

$$+ \varepsilon\{1 + C_1\kappa^{\frac{1}{2}} + 2\kappa \log \tfrac{1}{\varepsilon}\}\|A^{\frac{1}{2}}\|\|\hat{x}\|(1 + o(1)) ,$$

under the restriction

$$(7) \qquad \varepsilon\kappa^{3/2}\{C_2 + \kappa^{\frac{1}{2}}(1 + C_1) + (1 + C_1)\log \tfrac{1}{\varepsilon}\} \to 0 . \qquad \square$$

If $\|A^{\frac{1}{2}}(\hat{x} - x_0)\| \sim \|A^{\frac{1}{2}}\|\|\hat{x}\|$ (which is the case for instance if $x_0 = 0$ and $\|A^{\frac{1}{2}}\hat{x}\| \sim \|A^{\frac{1}{2}}\|\|\hat{x}\|$), then, apart from the (rather unimportant) factor $\log 1/\varepsilon$ ($\log 1/\varepsilon = 27.6$ if $\varepsilon = 10^{-12}$), we conclude from proposition 3 that essentially $\|A^{\frac{1}{2}}(\hat{x} - x_N)\| \sim \varepsilon(1 + C_1)\kappa^{3/2}\|A^{\frac{1}{2}}\|\|\hat{x}\|$. This is a very unsatisfactory result since A-numerical stability (cf. section 1.4) requires $\|A^{\frac{1}{2}}(\hat{x} - x_N)\| \sim \varepsilon\kappa^{\frac{1}{2}}\|A^{\frac{1}{2}}\|\|\hat{x}\|$, which is a factor κ smaller. Even if $\|A^{\frac{1}{2}}(\hat{x} - x_0)\| \leq \kappa^{-1}\|A^{\frac{1}{2}}\|\|\hat{x}\|$, still $\|A^{\frac{1}{2}}(\hat{x} - x_N)\| \sim \varepsilon\kappa\|A^{\frac{1}{2}}\|\|\hat{x}\|$, which does not guarantee A-numerical stability. However, if κ is not too large and if the required accuracy of the computed solution is not too high, one might decide to use RRGM instead of TRGM (which is well-behaved as we shall prove in the next section) because of time-saving. Besides of that, our numerical experiments (cf. section 3.4) indicate that estimate (6) is unsharp by a factor $\kappa^{\frac{1}{2}}$.

3.3. *The true residual gradient method (TRGM)*

Most of the results deduced in this section are based on the results of section 2.4 for a general TRDM. We also state a result concerning the monotonicity of the error $\|\hat{x} - x_i\|$ in the presence of round-off (for the algebraic case see theorem 3.1.1).
In order to translate the result of section 2.4 for TRGM we have to estimate the parameters α_i, β_i, γ_i defined by (2.4.20), (2.4.51) and (2.4.52). If $r_i = fl(b - Ax_i)$, $f_i = b - Ax_i$ and $\varphi_i := \|A\|\|x_i\| / \|f_i\|$, then we conclude from lemma 2.4.1 and lemma 2.2.6, under the restriction $\varepsilon(1 + C_1\varphi_i) \to 0$,

$$\alpha_i := \|f_i\|\|r_i\| / |(f_i, r_i)| = 1 + o(1) \ , \tag{1}$$

$$\beta_i := \|f_i\| \, \|A^{\frac{1}{2}} r_i\| / (\|A^{\frac{1}{2}}\| \, |(f_i, r_i)|) \leq \alpha_i = 1 + o(1) \ ,$$

$$\gamma_i^{-1} := \|A^{-\frac{1}{2}} f_i\|\|A^{\frac{1}{2}} r_i\| / |(f_i, r_i)| \leq \kappa^{\frac{1}{2}}\beta_i \leq \kappa^{\frac{1}{2}}(1 + o(1)) \ . \tag{2}$$

Substitution in (2.4.55) and (2.4.56) yields for the parameter ν_{i+1} of theorem 2.4.6 the inequality

$$|\nu_{i+1}| \leq 4\varepsilon\{1 + \kappa^{\frac{1}{2}} + \varphi_i\}(1 + o(1)) + \varepsilon\{\kappa^{\frac{1}{2}}(C_2 + C_1\kappa^{\frac{1}{2}}) + C_1\varphi_i\}o(1) \ , \tag{3}$$

under the restriction

$$\varepsilon\{\kappa^{\frac{1}{2}}(1 + C_2 + C_1\kappa^{\frac{1}{2}}) + (1 + C_1)\varphi_i\} \to 0 \ . \tag{4}$$

Retracing the proof of (1) one finds that, if $\varepsilon \leq 1/40$ and $\varepsilon C_1\varphi_i \leq 1/4$, then $\alpha_i \leq 2$ and consequently $\beta_i \leq 2$, $\gamma_i \geq (2\kappa^{\frac{1}{2}})^{-1}$.
Combining this and proposition 2.4.8 we obtain the following explicit version of this proposition for TRGM.

PROPOSITION 1. *If x_{i+1} is computed from one step TRGM based on an arbitrary machine vector x_i and if furthermore*

$$\varepsilon\{\kappa^{\frac{1}{2}}(1 + C_2 + C_1\kappa^{\frac{1}{2}}) + 2(1 + C_2)\} \leq \frac{1}{40} \ , \tag{5}$$

$$\varepsilon\{6 + C_1\}\varphi_i \leq \frac{1}{8} \ , \tag{6}$$

then we have

$$\frac{\|A^{\frac{1}{2}}(\hat{x} - x_{i+1})\|^2}{\|A^{\frac{1}{2}}(\hat{x} - x_i)\|^2} \leq 1 - \frac{1}{20\kappa} . \qquad (7)$$ □

From this proposition we can draw conclusions similar to those we derived for TRGSM from proposition 2.4.10, which conclusions can be combined into the following statement.

PROPOSITION 2. *If* $\{x_i\}$ *is generated by the TRGM with arbitrary initial machine vector* x_0 *and if* $\varepsilon\{\kappa^{\frac{1}{2}}(1 + C_2 + C_1\kappa^{\frac{1}{2}}) + 2(1 + C_2)\} \leq 1/40$, *then the natural error* $\|A^{\frac{1}{2}}(\hat{x} - x_i)\|$ *converges step-wise linearly with a convergence ratio no greater than* $(1 - (20\kappa)^{-1})^{\frac{1}{2}}$, *at least until the iteration step where the residual satisfies*

$$\|b - Ax_i\| \leq 8\varepsilon(6 + C_1)\|A\|\|x_i\| . \qquad (8)$$ □

This implies that TRGM is well-behaved and consequently numerically stable and A-numerically stable (cf. section 1.4).
Note that (8) can be used as a stopping criterion, provided an estimate for $\|A\|$ is available.

In theorem 3.1.1 and corollary 3.1.2 we stated the monotonicity of the error $\|\hat{x} - x_i\|$ for the algebraic (TR)GM. One may ask what can be said about this monotonicity if TRGM is performed in floating point arithmetic. Without giving a proof (which is very similar to the proof of theorem 2.4.6) we state the following analogue of theorem 3.1.1 in the presence of round-off.

THEOREM 3. *Let* x_{i+1} *be computed from one step TRGM based on an arbitrary machine vector* x_i. *Let* $f_i := b - Ax_i$ *and define*

$$\psi_i := \|A^{\frac{1}{2}}\|\|x_i\| \,/\, \|A^{-\frac{1}{2}} f_i\| , \qquad (9)$$

$$\sigma_i := \frac{(f_i, r_i)(r_i, A^{-1} r_i)}{\|A^{\frac{1}{2}} r_i\|^2 \|A^{-1} f_i\|^2} , \qquad (10)$$

$$\rho_i := \frac{\|r_i\|^2 (f_i, r_i)}{\|A^{\frac{1}{2}} r_i\|^2 (f_i, A^{-1} r_i)} . \qquad (11)$$

Then we have

$$\frac{\|\hat{x} - x_{i+1}\|^2}{\|\hat{x} - x_i\|^2} = 1 - \sigma_i(2 - \rho_i + \nu_{i+1}) \ , \tag{12}$$

where

$$|\nu_{i+1}| \le 2\varepsilon\{2\kappa^{\frac{1}{2}}(1 + C_2 + C_1\kappa^{\frac{1}{2}}) + 2(3 + C_2) + (1 + \kappa^{\frac{1}{2}}(1 + 2C_1))\psi_i\} \ , \tag{13}$$

under the restriction

$$\varepsilon\{\kappa^{\frac{1}{2}}(1 + C_2 + C_1\kappa^{\frac{1}{2}}) + (1 + C_1\kappa^{\frac{1}{2}})\psi_i\} \to 0 \ . \tag{14}$$

□

From lemma 2.4.1 we obtain under the restriction $\varepsilon\kappa^{\frac{1}{2}}(1 + C_1\psi_i) \to 0$

$$\sigma_i^{-1} = \frac{\|A^{\frac{1}{2}} r_i\|^2 \|A^{-1} r_i\|^2}{\|r_i\|^2 \|A^{-\frac{1}{2}} r_i\|^2}(1 + o(1)) \le \kappa(1 + o(1)) \ , \tag{15}$$

and

$$\rho_i \le \frac{\|A^{-\frac{1}{2}} r_i\| \|A^{-\frac{1}{2}} \hat{r}_i\|}{(\hat{r}_i, A^{-1} r_i)} = 1 + o(1) \ . \tag{16}$$

Hence we have the following corollary of theorem 2 (cf. (3.1.11)).

COROLLARY 4. *Let x_{i+1} be computed from one step TRGM based on an arbitrary machine vector x_i and let ψ_i be defined by (9). Then we have*

$$\frac{\|\hat{x} - x_{i+1}\|^2}{\|\hat{x} - x_i\|^2} \le 1 - \frac{1}{\kappa}(1 + o(1)) \ , \tag{17}$$

under the restriction

$$\varepsilon\kappa^{\frac{1}{2}}\{(1 + C_2 + C_1\kappa^{\frac{1}{2}}) + (1 + C_1)\psi_i\} \to 0 \ . \tag{18}$$

□

One can prove that, if (19) and (20) (to be stated presently) are satisfied, then $0 < \sigma_i^{-1} < (14/10)\kappa$, $|\rho_i| < (12/10)$ and $|\nu_{i+1}| < (7/10)$. Hence we have the following explicit version of corollary 4.

PROPOSITION 5. *Let* x_{i+1} *be computed from one step TRGM based on an arbitrary machine vector* x_i *and let* ψ_i *be defined by (9). Furthermore, let*

$$\varepsilon\{\kappa^{\frac{1}{2}}(1 + C_2 + C_1\kappa^{\frac{1}{2}}) + (3 + C_2)\} \leq \frac{1}{40}\,, \tag{19}$$

and

$$\varepsilon\{1 + \kappa^{\frac{1}{2}}(1 + 2C_1)\}\psi_i \leq \frac{1}{8}\,, \tag{20}$$

then we have

$$\frac{\|\hat{x} - x_{i+1}\|^2}{\|\hat{x} - x_i\|^2} \leq 1 - \frac{1}{14\kappa}\,. \tag{21}$$

□

Consequently, if (19) is satisfied, then for TRGM the error $\|\hat{x} - x_i\|$ converges step-wise linearly with a convergence ratio no greater than $(1 - (14\kappa)^{-1})^{\frac{1}{2}}$, at least until the iteration step where the natural error satisfies

$$\|A^{\frac{1}{2}}(\hat{x} - x_i)\| \leq 8\varepsilon(1 + \kappa^{\frac{1}{2}}(1 + 2C_1))\|A^{\frac{1}{2}}\|\|x_i\|\,, \tag{22}$$

(which implies that the monotonicity of the error cannot break down before the natural error reaches the level of the inherent natural error, cf. section 1.4).

Now assume for a moment that $\varepsilon\{\kappa^{\frac{1}{2}}(1 + C_2 + C_1\kappa^{\frac{1}{2}}) + (3 + 2C_2)\} \leq 1/40$, then both (5) and (19) are satisfied. If at a certain step (22) is not satisfied, then it follows from proposition 5 that the error decreases at the step from i to i + 1. Given $C_1 \geq 5$, the nonvalidity of (22) also implies that $\|A(\hat{x} - x_i)\| > 8\varepsilon(1 + 2C_1)\|A\|\|x_i\| \geq 8\varepsilon(6 + C_1)\|A\|\|x_i\|$. Hence (6) holds and consequently, from proposition 1, then also the natural error decreases at the step from i to i + 1. We observe that this reasoning does not imply that the natural error decreases at least as long as the error.

REMARK 6. Note that both (8) and (22) do not contain a term involving C_2. Thus the round-off errors occurring at the inner product computations do not influence the values of the (natural) error and the residual at a step where the monotonicity of the error or the natural

error breaks down. On the other hand, restrictions (5) and (19) indicate the allowable level if these round-off errors in order to have linear convergence. The round-off occurring at the inner product computations only influence directly the value of the parameter a_i: $fl(a_i) = \hat{a}_i(1 + \delta\tilde{a}_i)$ (cf. (2.4.66)). If $\delta\tilde{a}_i = 0$, then we have exact minimization of the objective function $\|A^{\frac{1}{2}}(\bar{x} - x)\|^2$ along the line $x = x_i + ar_i$. For $\delta\tilde{a}_i \neq 0$ the factor $(1 + \delta\tilde{a}_i)$ can be regarded upon as a relaxation factor (cf. section 2.2). Hence, as long as the inner product computations are performed with an accuracy guaranteeing $|\delta\tilde{a}_i| \leq 1 - \delta$, for some $\delta \in (0,1]$, then these computations do not affect the monotonicity of the natural error but only the convergence speed. This explains why C_2 does not occur in (5). By similar arguments one can explain why C_2 does not occur in (19). □

3.4. *Numerical experiments*

In this section we report on the numerical experiments that have been carried out with the GM. Our main goal is to verify the validity of our analytical results deduced in sections 2 and 3. In addition we want to investigate whether and under which conditions the various estimates are best-possible or essentially best-possible in the sense that they contain the correct exponent of κ.

In section 1.6 we discussed three possible ways of constructing test problems and implementing a descent method: assembled implementation (AI), product form implementation (PFI) and artificial floating point implementation (AFI). Before reporting on the results of the tests, we first specify further how these three ways of implementation are employed.

Firstly, the matrix $\Lambda = \mathrm{diag}(\lambda_1,\ldots,\lambda_n)$, containing the eigenvalues of A, has to be chosen. In our tests we used two different distributions for λ_i: the *logarithmical distribution*, where the ratio λ_{i+1}/λ_i is constant for all $i = 1,\ldots,n-1$, and the *equidistant distribution*, where $\lambda_{i+1} - \lambda_i$ is constant for all $i = 1,\ldots,n-1$. However, only very few tests have been carried out with the latter distribution and they did not bring in essentially different insights (from our point of view). Therefore, we do not report on these tests. We always choose $\lambda_n = 1$ (and hence $\lambda_1 = \kappa^{-1}$).

Secondly, in the case of AI and PFI, the orthogonal matrix U of eigenvectors of A has to be chosen. In all our tests U was chosen as a product $U_m := H_m \cdots H_1$ of m Householder transformations

$$(1) \qquad H_i = I - 2\, h_i\, h_i^T / (h_i, h_i) \ ,$$

where the vectors $h_1, \ldots, h_m$ were chosen randomly in the sense that each component is a pseudorandom number from the interval $[-1,+1]$.

REMARK 1. It is only the direction of h_i that determines H_i. However, choosing each component of the vectors $h_1, \ldots, h_m$ randomly from the interval $[-1,+1]$ does not generate randomly directed vectors. If, for instance, $n = 2$, then in the square $\{\alpha e_1 + \beta e_2 \mid \alpha,\beta \in [-1,+1]\}$ there are "more" vectors with a direction angle between $\pi/4 - \delta$ and $\pi/4 + \delta$ than there are with a direction angle between $-\delta$ and δ. This implies that relatively fewer vectors will have a direction close to the one of e_1 or e_2 than to the one of $e_1 + e_2$ or $e_1 - e_2$. For larger values of n this effect becomes more pronounced. In order to generate randomly directed vectors one can proceed as follows: generate a vector h by choosing its components randomly between -1 and +1; compute $\|h\|$; if $\|h\| \leq 1$, then the vector is accepted, otherwise the procedure is repeated. In our tests this refinement is omitted. □

In the case of AI the matrix $A = fl(U_m \Lambda U_m^T)$ is computed from the relation $A = H_m \cdots H_1 \Lambda H_1 \cdots H_m$ in the way suggested by Wilkinson ([65], section 5.30), where full advantage is taken of symmetry.
In the case of PFI we do not use the computed matrix A for matrix by vector product calculations, but these products are computed straightforward (from right to left) from the relation $Ax = H_m \cdots H_1 \Lambda H_1 \cdots H_m x$.
The choice of U does not apply to the case of AFI, since then we have $U = I$.
Finally, the vectors b and x_0 have to be chosen. This is done in the way as mentioned in section 1.6. We choose machine vectors s and e and next, for AI and PFI, we compute $\bar{x} = fl(U_m s)$, $b = fl(A\bar{x})$, $x_0 = fl(\bar{x} - U_m e)$, where the matrix by vector products are based on PFI, whereas for AFI we compute $b = fl(\Lambda s)$, $x_0 = fl(s - e)$.

In all our discussions of numerical results in this section the symbol A stands for the matrix $fl(U_m \Lambda U_m^T)$. With respect to AI and PFI this is

the machine matrix computed in the way just mentioned; for AFI this is the machine matrix Λ. The symbol b stands for the machine vector b.

If we display the values of $F_{i,\alpha} := fl(\|A^{\alpha}(\hat{x} - x_i)\|)$ $(\alpha = 0, \frac{1}{2}, 1)$, when reporting on numerical results, then, for AI and PFI these values are computed according to the formula

$$F_{i,\alpha} := fl(\|fl(\Lambda^{\alpha-1} U_m (fl_2(b - Ax_i)))\|) \quad . \tag{2}$$

Here $v_i := fl_2(b - Ax_i)$ denotes the vector with components computed in double length precision (with 2t-digit mantissae) and rounded to single length precision. Note that consequently we also need the assembled A for PFI. The computation of $w_{i,\alpha} := fl(\Lambda^{\alpha-1} U_m v_i)$ $(\alpha = 0, \frac{1}{2})$ is based on product form implementation.
If we display the values of $F_{i,\alpha}$ $(\alpha = 0, \frac{1}{2}, 1)$ for AFI, then these values are computed according to the formula

$$F_{i,\alpha} := fl(\|fl(\Lambda^{\alpha-1}(fl(b - \Lambda x_i)))\|) \quad . \tag{3}$$

REMARK 2. One may ask how many significant figures one obtains if $\|A^{\alpha}(\hat{x} - x_i)\|$ is computed from (2). To this purpose one can show that, for some C_3 depending only on n,

$$v_i = fl_2(b - Ax_i) = \hat{f}_i + \delta r_i \quad , \tag{4}$$

where $\hat{f}_i := b - Ax_i$ and

$$\|\delta r_i\| \leq \varepsilon \|\hat{f}_i\| + \varepsilon^2 C_3 \|A\| \|x_i\| (1 + o(1)) \qquad [\varepsilon C_3 \to 0] \quad . \tag{5}$$

If we assume that for any machine vector v

$$fl(\Lambda^{\alpha-1} U_m v) = (U_m^T A^{\alpha-1} + E_\alpha) v \quad , \tag{6}$$

where $\alpha = 0, \frac{1}{2}$ and $\|E_\alpha\| \leq \varepsilon C_\alpha \|A^{\alpha-1}\|$, then the vector $w_{i,\alpha}$ satisfies

$$w_{i,\alpha} = (U_m^T A^{\alpha-1} + E_\alpha)(\hat{f}_i + \delta r_i) = U_m^T A^{\alpha-1} \hat{f}_i + \delta w_{i,\alpha} \quad , \tag{7}$$

where

$$\delta w_{i,\alpha} = E_\alpha \hat{f}_i + U_m^T A^{\alpha-1} (I + A^{1-\alpha} U_m) \delta r_i \quad . \tag{8}$$

Consequently,

$$\|w_{i,\alpha}\| = \|A^{\alpha-1} f_i\| \, (1 + \eta_{i,\alpha}) \ , \tag{9}$$

where

$$|\eta_{i,\alpha}| \le \|\delta w_{i,\alpha}\| \, / \|A^{\alpha-1} f_i\| \le \tag{10}$$

$$\le \{\varepsilon C_\alpha \|A^{\alpha-1}\| \|f_i\| + (\varepsilon \|A^{\alpha-1}\| \|f_i\| +$$

$$+ C_3 \varepsilon^2 \kappa^{1-\alpha} \|A^\alpha\| \|x_i\|) \, (1 + o(1))\} \, / \|A^{\alpha-1} f_i\| \le$$

$$\le \varepsilon ((C_\alpha + 1) \kappa^{1-\alpha} + C_3 \varepsilon \kappa^{1-\alpha} \|A^\alpha\| \|x_i\| \, / \|A^{\alpha-1} f_i\|) \, (1 + o(1))$$

under the restriction $\varepsilon (C_3 + C_\alpha \kappa^{1-\alpha}) \to 0$.
If the calculation of the norm is performed in such a way that the relative error does not exceed $C_4 (1 + o(1))$, under the restriction $C_4 \varepsilon \to 0$, then we finally obtain

$$F_{i,\alpha} = \|A^{\alpha-1} f_i\| \, (1 + \mu_{i,\alpha}) = \|A^\alpha (\hat{x} - x_i)\| \, (1 + \mu_{i,\alpha}) \ , \tag{11}$$

where

$$|\mu_{i,\alpha}| \le \varepsilon (C_4 + (C_\alpha + 1) \kappa^{1-\alpha} + C_3 \varepsilon \kappa^{1-\alpha} \|A^\alpha\| \|x_i\| \, / \|A^{\alpha-1} f_i\|) \, (1 + o(1)) \tag{12}$$

under the restriction

$$\varepsilon (C_3 + C_4 + C_\alpha \kappa^{1-\alpha}) \to 0 \ . \tag{13}$$

If we define $C_\alpha := 0$ for $\alpha = 1$, then (13) also holds for $\alpha = 1$. Now, if for some constant K, one has $\|f_i\| \ge \varepsilon K \|A\| \|x_i\|$, then $\|A^{\alpha-1} f_i\| \ge$ $\ge \varepsilon K \|A^\alpha\| \|x_i\|$ and hence in that case, under the restriction (13), it follows that

$$|\mu_{i,\alpha}| \le \varepsilon (C_4 + (C_\alpha + 1 + C_3 K^{-1}) \kappa^{1-\alpha}) \, (1 + o(1)) \ . \tag{14}$$

This inequality indicates that if the residual is not essentially less than the inherent residual, then, using (2), the error, the natural error and the residual are computed with a relative error of order $\varepsilon\kappa$, $\varepsilon\kappa^{\frac{1}{2}}$, ε, respectively. □

REMARK 3. Consider the case of AFI, where $F_{i,\alpha}$ is computed from (3). If $\|b - Ax_i\| \geq \delta K\|A\|\|x_i\|$, then one has for $\mu_{i,\alpha}$ as given in (11) the estimate

$$(15) \qquad |\mu_{i,\alpha}| \leq \varepsilon(C_4 + (4 + (K\delta)^{-1})\kappa^{1-\alpha})(1 + o(1)) ,$$

under the restriction $\varepsilon \to 0$. Consequently, if the residual is not essentially less than the inherent residual (corresponding to machine precision δ), then, using (3), the error, the natural error and the residual are computed with a relative error of order $(\varepsilon/\delta)\kappa$, $(\varepsilon/\delta)\kappa^{\frac{1}{2}}$, (ε/δ), respectively. □

3.4.1. *The true residual gradient method*

From proposition 3.3.1 it follows that, if at a certain iteration step we have $\|A^{\frac{1}{2}}(\hat{x} - x_{k+1})\| \geq \|A^{\frac{1}{2}}(\hat{x} - x_k)\|$, then the residual $\|A(\hat{x} - x_k)\|$ is of order $\varepsilon\|A\|\|x_k\|$. Of course, in the tests the exact value of the natural error is not known. However, we can compute the approximate value $F_{i,\frac{1}{2}}$ from (2) or (3). Therefore the TRGM iterations are stopped as soon as $F_{k+1,\frac{1}{2}} \geq F_{k,\frac{1}{2}}$. The values of $F_{k,0}$, $F_{k,\frac{1}{2}}$, $F_{k,1}$ are referred to as *pseudo minimal error*, *pseudo minimal natural error* and *pseudo minimal residual*, respectively.

REMARK 1. One may ask whether $F_{k+1,\frac{1}{2}} \geq F_{k,\frac{1}{2}}$ also implies that $\|f_k\| = \|A(\hat{x} - x_k)\|$ is of the order $\bar{\varepsilon}\|A\|\|x_k\|$ ($\bar{\varepsilon} = \varepsilon$ or $\bar{\varepsilon} = \delta$). If also $\|A^{-\frac{1}{2}} f_{k+1}\| \geq \|A^{-\frac{1}{2}} f_k\|$, then this is obviously true (cf. proposition 3.3.2). Now assume that $F_{k+1,\frac{1}{2}} \geq F_{k,\frac{1}{2}}$ and $\|A^{-\frac{1}{2}} f_{k+1}\| < \|A^{-\frac{1}{2}} f_k\|$. In the case of AI or PFI it then follows from the analysis of remark 3.4.2 that

$$(1) \qquad F_{i,\frac{1}{2}}^2 = \|A^{-\frac{1}{2}} f_i\|^2 (1 + \tau_i) ,$$

where

$$(2) \qquad |\tau_i| \leq 2\varepsilon\{C_4 + ((1 + C_{\frac{1}{2}})\|A^{\frac{1}{2}}\|\|\hat{x}_i\| + \varepsilon C_3\kappa^{\frac{1}{2}}\|A^{\frac{1}{2}}\|\|x_i\|) / \|A^{-\frac{1}{2}} f_i\|\}(1 + o(1)) ,$$

under the restriction $\varepsilon\{C_4 + \kappa^{\frac{1}{2}}(1 + C_{\frac{1}{2}} + C_3) + \varphi_i\} \to 0$. Here, according

to definition (2.4.8), $\varphi_i := \|A\|\|x_i\| / \|f_i\|$. From section 3.3 we know that

$$(3) \qquad \frac{\|A^{-\frac{1}{2}} f_{k+1}\|^2}{\|A^{-\frac{1}{2}} f_k\|^2} = 1 - \gamma_k^2(1+\nu_{k+1}) ,$$

where γ_k^{-1} is defined by (3.3.2) and ν_{k+1} satisfies (3.3.3) under the restriction $\varepsilon\{\kappa^{\frac{1}{2}}(1+C_2+C_1\kappa^{\frac{1}{2}}) + (1+C_1)\varphi_k\} \to 0$.
Combining (1) (for i = k,k+1) and (3), and assuming $\|A^{-\frac{1}{2}} f_{k+1}\| < \|A^{-\frac{1}{2}} f_k\|$ one can prove that

$$(4) \qquad \frac{F_{k+1,\frac{1}{2}}^2}{F_{k,\frac{1}{2}}^2} = 1 - \gamma_k^2(1+\nu_{k+1}+\theta_{k+1}) ,$$

where $|\theta_{k+1}| = O(1)$ under the restriction

$$(5) \qquad \varepsilon\{C_2\kappa^{\frac{1}{2}} + \kappa(1+C_{\frac{1}{2}}+C_1+C_3+C_4) + (1+C_1)\varphi_k\} \to 0 .$$

Consequently, $F_{k+1,\frac{1}{2}} \geq F_{k,\frac{1}{2}}$ implies $\nu_{k+1} \leq -1 + O(1)$ under the restriction (5). A similar reasoning as given in section 2.4 leads to the conclusion that $\|A(\tilde{x} - x_k)\|$ is of order $\varepsilon\|A\|\|x_k\|$ if $\varepsilon\{C_2\kappa^{\frac{1}{2}} + \kappa(1+C + C_1+C_3+C_4)\}$ is small and $F_{k+1,\frac{1}{2}} \geq F_{k,\frac{1}{2}}$.
In the case of AFI one can prove that $F_{i,\frac{1}{2}}$ satisfies (1) where

$$(6) \qquad |\tau_i| \leq 2\varepsilon\{C_4 + (4\|A^{-\frac{1}{2}}\|\|\tilde{f}_i\| + \kappa^{\frac{1}{2}}\|A^{\frac{1}{2}}\|\|x_i\|) / \|A^{-\frac{1}{2}} f_i\| (1+O(1)) ,$$

under the restriction $\varepsilon\{C_4 + \kappa^{\frac{1}{2}} + \kappa^{\frac{1}{2}}\varphi_i\} \to 0$.
Combining this result (for i = k,k+1) with (3) (where, in all expressions involved, ε has to be replaced by the artificial machine precision δ), one can prove that (4) holds with $\theta_{k+1} = O(1)$ under the restriction

$$(7) \qquad \delta\{\kappa^{\frac{1}{2}}(1+C_2+C_1\kappa^{\frac{1}{2}}) + (1+C_1)\varphi_k\} + \varepsilon\kappa\{1+C_4+\varphi_k\} \to 0 .$$

Hence, similar to the case of AI and PFI, the inequality $F_{k+1,\frac{1}{2}} \geq F_{k,\frac{1}{2}}$ leads to the conclusion that $\|A(\tilde{x} - x_k)\|$ is of order $\delta\|A\|\|x_k\|$ if $\varepsilon\kappa < \delta$ and if $\delta\{C_4 + \kappa^{\frac{1}{2}}(1+C_2+C_1\kappa^{\frac{1}{2}})\}$ is small.

In summary, for all three kinds of implementation the inequality $F_{k+1,\frac{1}{2}} \geq F_{k,\frac{1}{2}}$ leads to the qualitatively same conclusions as the inequality

$$\|A^{\frac{1}{2}}(\hat{x} - x_{k+1})\| \geq \|A^{\frac{1}{2}}(\hat{x} - x_k)\| \ . \qquad \square$$

The influence of m

In order to investigate the influence of the value of m on the pseudo minimal (natural) error and the pseudo minimal residual for AI and PFI, we performed several tests with fixed dimension n, fixed logarithmical eigenvalue distribution and fixed eigenvector components s and e of the solution vector $\hat{x}$ and the initial error vector $\hat{x} - x_0$. We only varied the number m of Householder transformations and the Householder vectors $h_1,\ldots,h_m$. In order to invoke different round-off errors we chose ten different sets of random vectors $\{h_1,\ldots,h_m\}$ for each value of m.

The results in table 1 are obtained for the case where n = 30, $k = 10^2$ ($\lambda_1 = 10^{-2}$, $\lambda_{j+1}/\lambda_j = 10^{2/29} \sim 1.17$), $e_j/e_{j+1} = s_j/s_{j+1} = 10^{-3}$ (j = 1,...,29), $\|s\| = 1$, $\|e\| = 10^{-6}$.

Each pair of columns gives the smallest and the largest observed values (of the ten test problems for each value of m) of the measured quantity indicated on top of the table and computed according to (3.4.2) or (3.4.3). In all cases at iteration step k (defined as being the first step for which $F_{k+1,\frac{1}{2}} \geq F_{k,\frac{1}{2}}$) there hold $\|x_k\| \sim 1$ ($= \|\hat{x}\|$).

m	k		$\|\hat{x} - x_k\|$		$\|A^{\frac{1}{2}}(\hat{x} - x_k)\|$		$\|A(\hat{x} - x_k)\|$		
1	13	139	6.1_{10}-12	1.5_{10}-10	1.2_{10}-12	1.8_{10}-11	5.6_{10}-13	3.9_{10}-12	
5	9	28	1.1_{10}-10	3.0_{10}-10	1.2_{10}-11	4.3_{10}-11	2.0_{10}-12	8.8_{10}-12	
10	8	28	1.3_{10}-10	3.0_{10}-10	1.9_{10}-11	4.4_{10}-11	4.7_{10}-12	9.8_{10}-12	AI
15	15	27	1.6_{10}-10	2.7_{10}-10	2.4_{10}-11	3.8_{10}-11	5.4_{10}-12	9.1_{10}-12	
30	13	26	2.0_{10}-10	3.0_{10}-10	2.5_{10}-11	4.1_{10}-11	5.5_{10}-12	1.1_{10}-11	
100	14	24	1.7_{10}-10	3.6_{10}-10	2.4_{10}-11	4.8_{10}-11	6.3_{10}-12	9.9_{10}-12	
1	6	93	1.2_{10}-11	2.5_{10}-10	1.7_{10}-12	3.2_{10}-11	9.5_{10}-13	5.9_{10}-12	
5	8	17	2.2_{10}-10	5.3_{10}-10	2.8_{10}-11	6.4_{10}-11	6.8_{10}-12	2.4_{10}-11	
10	8	18	4.6_{10}-10	8.3_{10}-10	6.0_{10}-12	1.1_{10}-10	1.3_{10}-11	2.3_{10}-11	PFI
15	5	13	5.7_{10}-10	1.3_{10}-9	9.7_{10}-11	1.7_{10}-10	2.2_{10}-11	4.8_{10}-11	
30	5	14	1.1_{10}-9	1.6_{10}-9	1.7_{10}-10	2.4_{10}-10	4.1_{10}-11	6.3_{10}-11	
100	3	7	1.9_{10}-9	4.3_{10}-9	4.6_{10}-10	6.5_{10}-10	1.6_{10}-10	2.4_{10}-10	

TABLE 1. *The influence of* m

From the upper half of table 1 we see that in the case of AI the results hardly depend on m if $m \geq 5$. Only the case $m = 1$ seems to be special in the sense that the range of k is essentially larger and the lower bounds for $\|A^{\alpha}(\hat{x} - x_k)\|$ $(\alpha = 0, \frac{1}{2}, 1)$ are significantly smaller. We believe that this difference occurs because the matrix U, based on only one Householder transformation, is in general diagonally dominant and in fact close to the identity (note that $U_{ij} = \delta_{ij} - 2\, h_i * h_j / (h,h)$, where δ_{ij} is the Kronecker delta). Consequently A is close to Λ which causes an atypical round-off behavior at the matrix by vector product computations.

Apparently, in all cases of table 1, based on AI, one has for $\alpha = 0, \frac{1}{2}, 1$

$$(8) \qquad \begin{cases} \|A^{\alpha}(\hat{x} - x_k)\| \leq g_{\alpha} \varepsilon \kappa^{1-\alpha} \|A^{\alpha}\| \|x_k\| \ , \\ g_0 = 0.5 \ , \quad g_{\frac{1}{2}} = 0.7 \ , \quad g_1 = 1.4 \ , \end{cases}$$

which agrees with the good-behavior of the TRGM.

From the lower half of table 1 we see that in the case of PFI the pseudo minimal values increase as m increases. Since an increasing number of Householder transformations involves an increasing number of arithmetical operations, the constant C_1 increases as m increases and therefore this result is not surprising. Another observation that can be made in connection with the value of m is that for small values of m the round-off vector Ex, defined by $fl(Ax) = (U_m A U_m^T + E)x$, lies more or less in a subspace of at most dimension 2m. The subspace only depends on A and not on x. Therefore the vector Ex certainly is not randomly directed for small values of m. In order to identify this 2m-dimensional subspace we first state the following lemma without proof.

LEMMA 2. *Let* $U_m := H_m \cdots H_1$, *where* H_i *are Householder transformations based on arbitrary machine vectors* h_i $(i = 1, \ldots, m)$, *and let* x *be an arbitrary machine vector. If* $fl(U_m x)$ *is computed by product form implementation, then we have*

$$(9) \qquad fl(U_m x) = (U_m + \tilde{E}_m)x - \sum_{\ell=1}^{m} \theta_m^{(\ell)} H_m \cdots H_{\ell+1} h_\ell / \|h_\ell\| \ ,$$

where

(10) $$\|\tilde{E}_m\| \le 3m\varepsilon(1+o(1)) ,$$

(11) $$|\theta_m^{(\ell)}| \le 4(n+1)\varepsilon\|x\|(1+o(1)) , \quad (\ell = 1,\dots,m) ,$$

under the restriction $(m+n)\varepsilon \to 0$. □

The vector $\tilde{E}_m x$ is more or less arbitrary and $\|\tilde{E}_m x\| \le 3m\varepsilon\|x\|(1+o(1))$. Each vector $\theta_m^{(\ell)} H_m \cdots H_{\ell+1} h_\ell / \|h_\ell\|$ points into the direction $H_m \cdots H_{\ell+1} h_\ell$ $(\ell = 1,\dots,m)$, and $\|\theta_m^{(\ell)} H_m \cdots h_\ell\| / \|h_\ell\| \le 4(n+1)\varepsilon\|x\|$. Consequently, if $m \ll n$, then the vector $fl(U_m x) - U_m x$ is more or less an arbitrary vector in the subspace spanned by $H_m \cdots H_2 h_1, H_m \cdots H_3 h_2, \dots, h_m$ of dimension m at most, which actually is identical with the subspace spanned by $h_1,\dots,h_m$. In particular, if $m = 1$, then $(fl(U_m x) - U_m x)$ is parallel to h_1. The assertion (9) can also be written in the more convenient form

(12) $$fl(U_m x) = (U_m + E_m)x ,$$

where by lemma 2 it follows that

(13) $$\|E_m\| \le m(4n+7)\varepsilon(1+o(1)) ,$$

under the restriction $(m+n)\varepsilon \to 0$.
The complete product form computation of fl(Ax) satisfies

(14) $$fl(Ax) = (U_m + E_m'')\Lambda(I+D)(U_m^T + E_m')x = (U_m \Lambda U_m^T + E)x ,$$

where E_m' and E_m'' satisfy (14), $|D| \le \varepsilon I$ and hence

(15) $$\|E\| \le m(8n+15)\varepsilon\|\Lambda\|(1+o(1)) , \quad [(m+n)\varepsilon \to 0] .$$

Similarly to the previous considerations it follows that, in case $m \ll n$, the vector Ex is more or less an arbitrary vector in the subspace spanned by $h_1,\dots,h_m,\Lambda h_1,\dots,\Lambda h_m$ of dimension at most 2m. In particular, if $m = 1$, then $Ex \in \mathrm{span}\{h_1,\Lambda h_1\}$. Hence for small values of m the product form implementation of $U_m \Lambda U_m^T x$ certainly not agrees with the real-world implementation where Ex is randomly directed. Apparently, in all cases of table 1, based on PFI, one has for $\alpha = 0, \frac{1}{2}, 1$

(16) $$\begin{cases} \|A^\alpha(\hat{x} - x_k)\| \le g_\alpha \varepsilon \kappa^{1-\alpha}\|A^\alpha\|\|x_k\| , \\ g_0 = 5.9 , \quad g_{\frac{1}{2}} = 9.0 , \quad g_1 = 33 , \end{cases}$$

which agrees with the good-behavior of the TRGM.

The influence of the eigenvector components of $\hat{x}$ *and* $\hat{x} - x_0$

Algebraically, the sequence $\{\hat{x} - x_i\}$ obtained by the GM only depends on the matrix A and the initial error vector $\hat{x} - x_0$, but not on $\hat{x}$. In order to investigate this for the numerical process, at least as far as $\hat{x} - x_0$ and $\hat{x}$ are concerned, we did several tests, only varying the eigenvector components of $\hat{x} - x_0$ and $\hat{x}$.
As mentioned before, these eigenvector components are controlled by machine vectors e and s, respectively. For both vectors we experimented with three different distributions, viz. s_j / s_{j+1}, e_j / e_{j+1} = $= 10^3, 1, 10^{-3}$ $(j = 1,\ldots,m-1)$. In case $s_j / s_{j+1} = 10^3$ the vector $\hat{x}$ points into the direction of the eigenvectors corresponding to the small eigenvalues and therefore this vector is called a *small-oriented vector*. Similarly, a vector that points into the direction of the eigenvectors corresponding to the large eigenvalues is called a *large-oriented vector*, and a vector with more or less equal eigenvector components is called an *un-oriented vector*. For each of the six combinations of distribution of the components of s and e, as indicated in table 2, we performed five different tests based on PFI. These tests are different in the sense that we chose different sets of random vectors $\{h_1,\ldots,h_m\}$ in order to invoke different round-off errors. In all cases $n = 20$, $m = 5$, $\kappa = 10^4$, $\|s\| = 1$, $\|\Lambda^{\frac{1}{2}} e\|$ $(= \|A^{\frac{1}{2}}(\hat{x} - x_0)\|) = 10^{-6}$. Just like in table 1 each pair of columns presents the smallest and the largest observed value in the five tests and furthermore in all cases it turned out that $\|x_k\| \sim 1$ $(= \|\hat{x}\|)$.

k		e_j/e_{j+1}	s_j/s_{j+1}	$\|\hat{x} - x_k\|$		$\|A^{\frac{1}{2}}(\hat{x} - x_k)\|$		$\|A(\hat{x} - x_k)\|$	
1180	6042	10^3	10^3	$1.1_{10}-8$	$5.6_{10}-8$	$1.4_{10}-10$	$6.0_{10}-10$	$2.5_{10}-12$	$9.7_{10}-12$
1252	4133	10^3	1	$4.9_{10}-8$	$1.2_{10}-7$	$5.2_{10}-20$	$1.2_{10}-9$	$7.0_{10}-12$	$1.4_{10}-11$
362	10451	10^3	10^{-3}	$6.5_{10}-8$	$4.2_{10}-7$	$4.2_{10}-10$	$1.8_{10}-9$	$9.0_{10}-12$	$5.1_{10}-11$
1180	6042	10^3	10^3	$1.1_{10}-8$	$5.6_{10}-8$	$1.4_{10}-10$	$6.0_{10}-10$	$2.5_{10}-12$	$9.7_{10}-12$
12859	14732	1	10^3	$1.0_{10}-8$	$3.4_{10}-8$	$1.1_{10}-10$	$6.0_{10}-10$	$2.3_{10}-12$	$6.8_{10}-12$
2	55	10^{-3}	10^3	$4.8_{10}-9$	$2.0_{10}-8$	$6.2_{10}-11$	$2.0_{10}-10$	$2.0_{10}-12$	$4.6_{10}-12$

TABLE 2. *The influence of the eigenvector components*

The difference between the values of $\|A^{\alpha}(\hat{x}-x_k)\|$, $(\alpha = 0,\frac{1}{2},1)$ in the upper half of table 2 where only the eigenvector components of the solution vector $\hat{x}$ are varied, are rather small. Hence we conclude that the direction of $\hat{x}$ does not affect the pseudo minimal values, although the values of $\|A^{\alpha}(\hat{x}-x_k)\|$, $(\alpha = 0,\frac{1}{2},1)$ seem to be slightly larger in case $\hat{x}$ is larger oriented. More pronounced is the wide range for the value of k in the case of large-oriented vectors $\hat{x}$. We have no satisfactory explanation for this phenomenon.

From the lower half of table 2, where only the eigenvector components of the initial error vector $\hat{x}-x_0$ are varied, we see that the values of $\|A^{\alpha}(\hat{x}-x_k)\|$, $(\alpha = 0,\frac{1}{2},1)$ do not depend on these components, but the number of steps needed to reach these values strongly depends on these components (note that in all cases initially $\|A^{\frac{1}{2}}(\hat{x}-x_0)\| = 10^{-6}$). For the cases where $\hat{x}-x_0$ is either large- or small oriented, the convergence appears to be faster than in the un-oriented case. This can be explained as follows. In the oriented cases the initial natural error $A^{\frac{1}{2}}(\hat{x}-x_0)$ essentially belongs to an invariant subspace of relatively small dimension, spanned by eigenvectors associated with either small or large eigenvalues. Hence it seems as if we are solving a linear equation of lower dimension and with a smaller condition number (the quotient of the extreme eigenvalues corresponding to the subspace involved), and this has a favourable influence on the initial convergence behavior.

As far as the two oriented cases are concerned, the convergence appears to be faster in the large-oriented case (l.o. case). From our experiments it turns out that this is mainly caused by the much stronger decrease of the natural error in the first step. If no round-off would occur one would have (cf. theorem 2.2.2)

$$\frac{\|A^{\frac{1}{2}}(\hat{x}-x_1)\|^2}{\|A^{\frac{1}{2}}(\hat{x}-x_0)\|^2} = 1 - \hat{\gamma}_0^2 , \tag{17}$$

where

$$\hat{\gamma}_0^2 := (\hat{r}_0,\hat{r}_0)^2 / (\|A^{-\frac{1}{2}}\hat{r}_0\|^2\|A^{\frac{1}{2}}\hat{r}_0\|^2) , \tag{18}$$

$$\hat{r}_0 := b - Ax_0 . \tag{19}$$

It can be shown that $(1-\hat{\gamma}_0^2)^{\frac{1}{2}} \sim 8_{10}\text{-}4$ for the small-oriented case (s.o. case) and $(1-\hat{\gamma}_0^2)^{\frac{1}{2}} \sim 3_{10}\text{-}4$ for the l.o. case. Consequently, the algebraic decrement of the natural error hardly differs for the two cases and hence the slower decrease for the s.o. case in the first step is due to round-off. In the presence of round-off the analogue of (17) reads

$$\frac{\|A^{\frac{1}{2}}(\hat{x}-x_1)\|^2}{\|A^{\frac{1}{2}}(\hat{x}-x_0)\|^2} = 1 - \gamma_0^2(1+\nu_1) \ , \tag{20}$$

where

$$\gamma_0^2 := (f_0,r_0)^2 / (\|A^{-\frac{1}{2}} f_0\|^2 \|A^{\frac{1}{2}} r_0\|^2) \ , \tag{21}$$

$$r_0 := fl(b - Ax_0) \ , \tag{22}$$

and where ν_1 satisfies (3.3.3) under the restriction (3.3.4) and f_0 is defined by (19). From lemma 2.4.1 it follows that

$$\begin{cases} \dfrac{(f_0,r_0)}{(f_0,f_0)} = 1 + o(1) & [\varepsilon(1+C_1\varphi_0) \to 0] \ , \\ \dfrac{\|A^{\frac{1}{2}} r_0\|}{\|A^{\frac{1}{2}} f_0\|} = 1 + o(1) & [\varepsilon(\kappa^{\frac{1}{2}} + C_1\chi_0) \to 0] \ , \end{cases} \tag{23}$$

where $\varphi_0 := \|A\|\|x_0\| / \|A(\hat{x}-x_0)\|$ and $\chi_0 := \|A^{3/2}\|\|x_0\| / \|A^{3/2}(\hat{x}-x_0)\|$. For the s.o. case we have $\varphi_0 = 10^{-8}$, $\chi_0 = 10^{-10}$ and as a more explicit version of (23) one can show that for this case one approximately has

$$\left|\frac{(f_0,r_0)}{(f_0,f_0)} - 1\right| \le C_1\, 10^{-3} \ , \quad \left|\frac{\|A^{\frac{1}{2}} r_0\|}{\|A^{\frac{1}{2}} f_0\|} - 1\right| \le C_1\, 10^{-1} \ . \tag{24}$$

For the l.o. case we have $\varphi_0 = \chi_0 = 10^{-6}$ and as a more explicit version of (23) one can show that for this case one approximately has

$$\left|\frac{(f_0,r_0)}{(f_0,f_0)} - 1\right| \le C_1\, 10^{-5} \ , \quad \left|\frac{\|A^{\frac{1}{2}} r_0\|}{\|A^{\frac{1}{2}} f_0\|} - 1\right| \le C_1\, 10^{-5} \ . \tag{25}$$

As far as ν_1 is concerned, the estimate (3.3.3) suggests that ν_1 is

larger for the s.o. case. However, the basic estimate (2.4.55), which still contains the parameter β_i, contradicts this expectation because, although φ_0 is a factor 10^2 larger, the parameter β_0 is a factor 10^2 smaller for the s.o. case. For both cases one can show that $|\nu_1| \leq 4_{10}\text{-}5$. Hence, for the s.o. case $(1-\gamma_0^2(1+\nu_1))$ can be of order $C_1\ 10\text{-}1$ and for the l.o. case $(1-\gamma_0^2(1+\nu_1))$ can be of order $C_1\ 10\text{-}5$, which explains the difference in decrement of the natural error in the first step for the two cases.

Of course, $A^{\frac{1}{2}}(\hat{x}-x_0)$ does not belong exactly to an invariant subspace of lower dimension. In the s.o. case the eigenvector components of $A^{\frac{1}{2}}(\hat{x}-x_0)$ associated with the large eigenvalues are reactivated by the GM. Therefore the difference in speed of convergence between the s.o. case and the u.o. case is restricted to the first steps. This is illustrated in figure 1 where the values of $\|A^{\alpha}(\hat{x}-x_i)\|$ $(\alpha = 0, \frac{1}{2}, 1)$ are plotted for a test problem with $e_j / e_{j+1} = 10^3$ and a test problem with $e_j / e_{j+1} = 1$ (in both tests $s_j / s_{j+1} = 10^3$). The initial values $\|(\hat{x}-x_0)\|$, $\|A^{\frac{1}{2}}(\hat{x}-x_0)\|$, $\|A(\hat{x}-x_0)\|$ are $_{10}\text{-}4$, $_{10}\text{-}6$, $_{10}\text{-}8$ and $2.8_{10}\text{-}6$. $_{10}\text{-}6$, $8.0_{10}\text{-}7$, respectively, whereas k is 5942 and 12859, respectively. We see that after the first 3000 steps the rate of convergence of $\|A^{\frac{1}{2}}(\hat{x}-x_i)\|$ is approximately the same for both tests. In both cases there holds for all $i \geq 3000$

$$0.9997 \leq \frac{\|A^{\frac{1}{2}}(\hat{x}-x_{i+1})\|}{\|A^{\frac{1}{2}}(\hat{x}-x_i)\|} \leq 0.9998 ,$$

and hence the convergence ratio varies between $1-3/\kappa$ and $1-2/\kappa$. Algebraically we know that the convergence ratio is no greater than $(\kappa-1)/(\kappa+1) \sim 1-2/\kappa$ (cf. (3.1.7)). The proof of this result is based on the Kantorovich inequality applied to the residual vector r_i, viz.

$$(26) \qquad \frac{\|A^{\frac{1}{2}}(\hat{x}-x_{i+1})\|^2}{\|A^{\frac{1}{2}}(\hat{x}-x_i)\|^2} = 1 - \frac{\|r_i\|^4}{\|A^{-\frac{1}{2}}r_i\|^2\|A^{\frac{1}{2}}r_i\|^2} \leq$$

$$\leq 1 - \frac{4\kappa}{(\kappa+1)^2} = \left(\frac{\kappa-1}{\kappa+1}\right)^2 ,$$

with equality iff r_i is a multiple of the vector $u_1 + u_n$ or $u_1 - u_n$, where u_ℓ are eigenvectors of A. In the two tests the residuals r_i

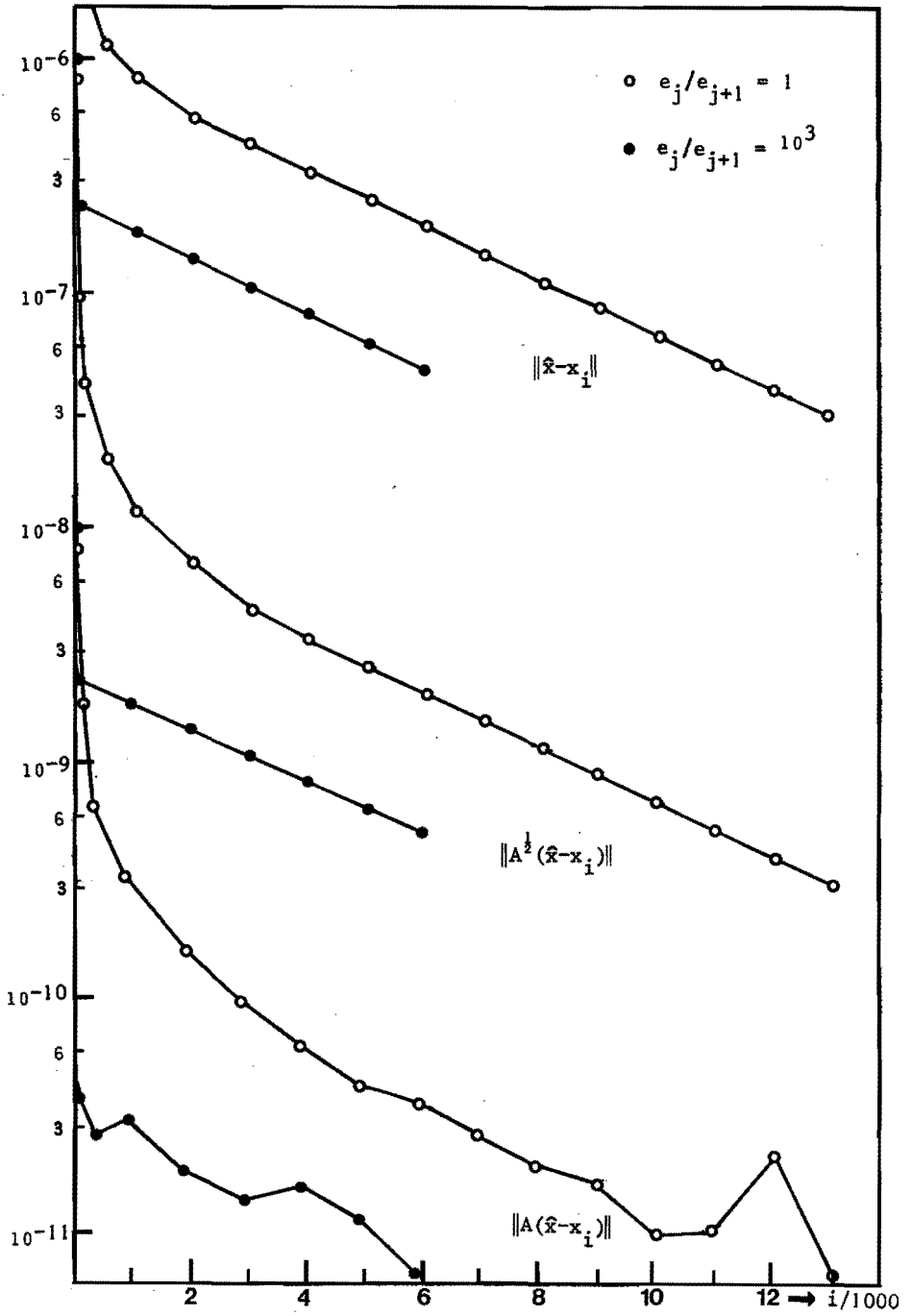


FIGURE 1. *The influence of the eigenvectorcomponents of* $\hat{x}-x_0$

($i \geq 3000$) appeared to be more or less un-oriented. Therefore it is not surprising that the convergence ratio in the tests is slightly better as indicated by the algebraic upper bound.
In the case of an exactly un-oriented residual vector, r_i being a multiple of the vector $\Sigma_{\ell=1}^{n} u_\ell$, the Kantorovich quotient $\|r_i\|^4 / (\|A^{-\frac{1}{2}} r_i\|^2 \|A^{\frac{1}{2}} r_i\|^2)$ would be approximately $60/\kappa$ for the test matrix A with logarithmical eigenvalue distribution and $\kappa = 10^4$. This would give rise to a convergence ratio $1 - 30/\kappa$. Hence, there seems to be somewhat more structure in the distribution of the eigenvector components of r_i, namely such that the Kantorovich quotient is approximately $60/\kappa$
Apparently, for all cases of table 2 one has

$$(27) \qquad \begin{cases} \|A^\alpha(\hat{x} - x_k)\| \leq g_\alpha \varepsilon \kappa^{1-\alpha} \|A^\alpha\| \|x_k\| , \\ g_0 = 5.8 , \quad g_{\frac{1}{2}} = 2.5 , \quad g_1 = 7.0 , \end{cases}$$

which agrees with the good-behavior of the TRGM.
Figure 1 also confirms our result (cf. proposition 3.3.5) stating that the error $\|\hat{x} - x_i\|$ converges step-wise linearly with a convergence ratio no greater than $(1 - 1/\kappa(1 + o(1))$ as long as the natural error has not attained the level of the inherent natural error.

The influence of κ

A well-behaved method has the property that $\|A^\alpha(\hat{x} - x_k)\| \leq$
$\leq g_\alpha \varepsilon \kappa^{1-\alpha} \|A^\alpha\| \|x_k\|$ ($\alpha = 0, \frac{1}{2}, 1$), where g_α only depends on α, ε and the dimension of the system. In (8), (16) and (27) we presented already the values for g_α following from the previous tests. The differences between these values of g_α (for the same value of α) are due to the implementation, the values of m and n, and the choices for s and e. In order to eliminate these influences we performed 25 tests based on AFI, with fixed n, s and e, but with a variable condition number κ, viz. $\kappa = 10^p$ ($p = 2, 2.5, 3, 3.5, 4$). We performed five different tests for each condition number; different in the sense that we invoked different artificial round-off errors (cf. section 1.6). The results in table 3 are obtained for the case where $n = 20$, $s_j / s_{j+1} = e_j / e_{j+1} = 10^3$ ($j = 1, \ldots, 19$), $\|s\| = 1$, $\|e\| = \kappa * 10^{-2}$, with artificial relative machine precision $\delta = 10^{-7}$. The displayed values have

to be interpreted in the same way as in the two previous tables, each line corresponding to five tests.

κ	k		$\|\hat{x} - x_k\|$		$\|A^{\frac{1}{2}}(\hat{x} - x_k)\|$		$\|A(\hat{x} - x_k)\|$	
10^2	105	120	9.9_{10}-6	1.4_{10}-5	1.2_{10}-6	1.6_{10}-6	2.3_{10}-7	3.0_{10}-7
$10^{2.5}$	333	360	3.9_{10}-5	5.2_{10}-5	2.6_{10}-6	3.4_{10}-6	2.4_{10}-7	2.8_{10}-7
10^3	847	1100	1.3_{10}-4	1.9_{10}-4	5.0_{10}-6	7.3_{10}-6	2.4_{10}-7	3.1_{10}-7
$10^{3.5}$	3131	4717	5.1_{10}-4	7.0_{10}-4	1.0_{10}-5	1.3_{10}-5	2.6_{10}-7	2.7_{10}-7
10^4	20768	26608	2.3_{10}-3	2.3_{10}-3	2.3_{10}-5	2.7_{10}-5	2.7_{10}-7	3.0_{10}-7

TABLE 3. *The influence of* κ

Table 3 shows that in all tests for $\alpha = 0, \frac{1}{2}, 1$ we have

$$(28) \qquad g_\alpha^{(1)} \delta \kappa^{1-\alpha} \|A^\alpha\| \|x_k\| \leq \|A^\alpha(\hat{x} - x_k)\| \leq g_\alpha^{(2)} \delta \kappa^{1-\alpha} \|A^\alpha\| \|x_k\| ,$$

where $g_\alpha^{(1)} = 0.99$ and $g_\alpha^{(2)} = 3.1$, which agrees with the good-behavior of the TRGM. Somewhat striking is the fact that the constants $g_\alpha^{(1)}$, $g_\alpha^{(2)}$ both systematically slightly increase as κ increases ($\alpha = 0, \frac{1}{2}$).

The influence of the basic arithmetical operations

From proposition 3.3.2 and remark 3.4.2 it follows that, if ε, C_1, C_2 and κ are sufficiently small, then the pseudo minimal residual is at most of order $\varepsilon(1 + C_1)\|A\|\|x_k\|$, and consequently the pseudo minimal (natural) error is at most of order $\varepsilon(1 + C_1)\kappa^{1-\alpha}\|A^\alpha\|\|x_k\|$ ($\alpha = 0, \frac{1}{2}$). This implies that the pseudo minimal values do not depend on the constant C_2, corresponding to the round-off errors occurring at the inner product computations. In order to verify this we performed 40 tests, based on AFI, with $n = 20$, $s_j / s_{j+1} = e_j / e_{j+1} = 10^3$ ($j = 1, \ldots, 19$), $\|e\| = \|s\| = 1$, $\kappa = 10^4$ fixed for all tests. We only varied the (artificial) relative machine precision of the various arithmetical operations. Furthermore, we distinguished between three types of arithmetical operations like we did in the round-off error analysis, viz. the dyadic arithmetical operations +, - (both for vectors), * (for scalar by vector), / (for scalars), the matrix by vector product operations (C_1) and the inner product operations (C_2). Each of these three types is performed either with artificial relative

machine precision $\delta = 10^{-7}$ (implemented in the way as described in section 1.6) or with artificial relative machine precision $\varepsilon = 10^{-11}$. If some type of arithmetical operation is performed with precision δ, then the arithmetical operations performed with precision ε can be regarded upon as being performed exactly (which means $C_1 = 0$ or $C_2 = 0$ in the appropriate cases). The results are written down in table 4. The first three columns indicate the relative machine precision for each of the three types of arithmetical operations, whereas the other columns indicate the smallest and largest observed values (of the five tests for each case), as in the previous tables. Again it turned out that in all cases $\|x_k\| \sim 1$.

A*v	+ − * /	(v,w)	k		$\|\hat{x} - x_k\|$		$\|A^{\frac{1}{2}}(\hat{x} - x_k)\|$		$\|A(\hat{x} - x_k)\|$	
δ	δ	δ	3955	12942	2.6_{10}-3	3.1_{10}-3	2.6_{10}-5	3.1_{10}-5	2.8_{10}-7	4.0_{10}-7
δ	δ	ε	5683	9728	2.4_{10}-3	3.0_{10}-3	2.4_{10}-5	3.0_{10}-5	2.7_{10}-7	3.3_{10}-7
δ	ε	δ	6284	10707	2.4_{10}-3	2.6_{10}-3	2.4_{10}-5	2.6_{10}-5	2.7_{10}-7	3.1_{10}-7
δ	ε	ε	6675	12628	2.3_{10}-3	2.7_{10}-3	2.3_{10}-5	2.7_{10}-5	2.6_{10}-7	2.9_{10}-7
ε	δ	δ	4	219	5.5_{10}-4	6.2_{10}-4	7.0_{10}-6	7.9_{10}-6	1.2_{10}-7	1.6_{10}-7
ε	δ	ε	26	240	5.4_{10}-4	6.2_{10}-4	6.9_{10}-6	7.8_{10}-6	1.3_{10}-7	1.5_{10}-7
ε	ε	δ	20506	23104	1.5_{10}-7	1.8_{10}-7	2.0_{10}-9	2.3_{10}-9	2.8_{10}-11	3.6_{10}-11
ε	ε	ε	21516	23500	1.5_{10}-7	1.8_{10}-7	1.9_{10}-9	2.3_{10}-9	2.8_{10}-11	3.6_{10}-11

TABLE 4. *The influence of the basic arithmetical operations*

We see that the results of each successive pairs of lines, where the relative precision of the inner product computations is either δ or ε, hardly differ. This confirms our analytical result that C_2 does not affect the pseudo minimal values.

The very small differences between the first four lines where the relative precision of the matrix by vector product computations equals δ, agrees with the analytical result that C_1 has a main influence on the pseudo minimal valus.

Comparing lines 5 and 6 (where the matrix by vector product operations are carried out with relative precision $\varepsilon = 10^{-11}$ and the dyadic arithmetical operations are carried out with relative precision $\delta = 10^{-7}$) with lines 3 and 4 (where we have the opposite case), we see

that the pseudo minimal values are slightly smaller in the first cases. This can be explained by the fact that round-off due to matrix by vector product operations is proportional to $C_1\delta$ ($C_1 \sim n^{\frac{1}{2}}$), whereas round-off due to the dyadic arithmetical operations is proportional to δ. Somewhat surprising is the extremely fast convergence in the first cases. Similar to the tests of table 2, this is mainly due to the much stronger decrease of the natural error in the first step, which is a consequence of the fact that $\hat{\gamma}_0$ defined by (18) and γ_0 defined by (21) differ less in the first cases.
For a more detailed explanation see the discussion concerning (17) and (20).
The results presented in the last two lines of table 4 differ by a factor of the order 10^{-4} from the results presented on the first two lines; this is just as one would expect, since the corresponding relative machine precisions also differ by a factor of the order 10^{-4}. Since in all tests $\|A^{\frac{1}{2}}(\tilde{x} - x_0)\| = 10^{-2}$, the natural error decreased by a factor 10^3 in the tests of the first two lines and by a factor 10^7 in the tests of the last two lines. This explains why more steps are needed in the tests of the last two lines.

3.4.2. *The recursive residual gradient method*

As far as our analytical results for the numerical behavior of the RRGM are concerned, the most striking result is the step-wise linear convergence to zero of the natural error $\|A^{-\frac{1}{2}} r_i\|$. We performed several tests with the RRGM, based on PFI ($m = 5$) and AFI ($\delta = 10^{-8}$), varying the dimension n ($20 \leq n \leq 50$), the eigenvector components s and e (s_j / s_{j+1}, $e_j / e_{j+1} = 10^3, 1, 10^{-3}$, $\|s\| = 1$, $10^{-1} \leq \|e\| \leq 10^{-6}$) and the condition number κ ($10^2 \leq \kappa \leq 10^4$). In order to avoid underflow, the iterations were stopped as soon as $\|r_i\| \leq 10^{-22}$. In all cases this level was attained.
From our tests it was hard to conclude anything about the influence of the dimension n on the numerical behavior of r_i.
Varying the eigenvector components e of the initial error vector only affected the convergence ratio of $\|A^{-\frac{1}{2}} r_i\|$ in the first (hundreds of) steps. Varying the eigenvector components s of the solution vector hardly caused any difference in the numerical behavior as far as r_i is

concerned. Varying the condition number affected the convergence ratio during all iterations, as was to be expected. In all cases, after some hundreds of steps, the convergence ratio of the natural error $\|A^{-\frac{1}{2}} r_i\|$ varied between $1 - 4/\kappa$ and $1 - 2/\kappa$, indicating that the initial orientation of the recursive residuals dies out and changes to an orientation that induces Kantorovich quotients varying between $8/\kappa$ and $4/\kappa$ (see the discussion concerning (3.4.1.26)).
As far as the approximations $\{x_i\}$ are concerned we have the analytical result formulated in proposition 3.2.3. It states that at iteration step $N := \text{ent}(2\kappa \log 1/\varepsilon + 1)$ there holds

$$(1) \qquad \|A^{\frac{1}{2}}(\hat{x} - x_N)\| \leq$$

$$\leq \varepsilon\{1 + \kappa^{\frac{1}{2}} + C_1\kappa + 2(4\kappa + (2 + C_1)\kappa^{3/2}) \log \frac{1}{\varepsilon}\} \|A^{\frac{1}{2}}(\hat{x} - x_0)\| (1 + o(1)) +$$

$$+ \varepsilon\{1 + C_1\kappa^{\frac{1}{2}} + 2\kappa \log \frac{1}{\varepsilon}\} \|A^{\frac{1}{2}}\| \|\hat{x}\| (1 + o(1)) \ ,$$

under the restriction (3.2.7). In order to verify whether (1) is sharp, in the sense that it contains the correct maximal exponent of κ, one has to test with large values of κ to be able to distinguish between $\kappa^{3/2}$, κ, etc. However, the number N of iteration steps is proportional to κ. The tests would cost a considerable amount of computing time. Therefore these tests were omitted.
The result (1) has been obtained from the intermediate analytical result formulated in theorem 2.3.2.2. The inequality holds for all $i \geq 0$ and is expressed in terms of the residual vector $f := b - Ax_i$ and the recursively computed residual vector r_i.
In the case of the RRGM the inequality reads

$$(2) \quad \|A^{-\frac{1}{2}}(f_i - r_i)\| \leq$$

$$\leq \varepsilon(1 + \varepsilon\kappa^{\frac{1}{2}})^i\{\kappa^{\frac{1}{2}} + C_1\kappa + \sqrt{i}\ (4\kappa^{\frac{1}{2}} + (2 + C_1)\kappa)\} \|A^{\frac{1}{2}}(\hat{x} - x_0)\| (1 + o(1)) +$$

$$+ \varepsilon(1 + \varepsilon\kappa^{\frac{1}{2}})^i\{i + C_1\kappa^{\frac{1}{2}} + \varepsilon\sqrt{i}\ C_1(4\kappa + (2 + C_1)\kappa^{3/2})\} \|A^{\frac{1}{2}}\| \|\hat{x}\| (1 + o(1)) \ ,$$

under the restriction $\varepsilon\kappa^{3/2}(C_2 + (1 + C_1)\kappa^{\frac{1}{2}}) \to 0$. For the natural error $\|A^{\frac{1}{2}}(\hat{x} - x_i)\|$ we have the estimate (cf. (2.3.2.30))

$$(3) \qquad \|A^{\frac{1}{2}}(\hat{x} - x_i)\| \leq \|A^{-\frac{1}{2}}(f_i - r_i)\| + \|A^{-\frac{1}{2}} r_i\| \ .$$

Consequently, for small $\|A^{-\frac{1}{2}} r_i\|$ estimate (2) also holds with $\|A^{-\frac{1}{2}}(\tilde{r}_i - r_i)\|$ replaced by $\|A^{\frac{1}{2}}(\tilde{x} - x_i)\|$. In order to verify whether (2) is sharp we performed a few tests based on AFI with artificial relative machine precision $\delta = 10^{-8}$ in the way described in section 1.6. Similar to the TRGM the iteration steps were stopped as soon as $F_{k+1,\frac{1}{2}} \geq F_{k,\frac{1}{2}}$. We report on the results for only one typical test. These results are displayed in table 1 and were obtained for the case where $n = 20$, $s_j / s_{j+1} = p_j / p_{j+1} = 10^3$, $\|s\| = 1$, $\|p\| = 10^3$, $\kappa = 10^4$. The iterations stopped at step $k = 14244$.

i	$\|\tilde{x} - x_i\|$	$\|A^{\frac{1}{2}}(\tilde{x}-x_i)\|$	$\|A(\tilde{x}-x_i)\|$	$\|A^{-\frac{1}{2}} r_i\|$
0	$1.00_{10}+3$	$1.00_{10}+1$	$1.00_{10}-1$	$1.00_{10}+1$
1	$6.43_{10}-1$	$7.60_{10}-2$	$5.11_{10}-2$	$7.59_{10}-2$
2	$6.33_{10}-1$	$3.35_{10}-2$	$1.13_{10}-2$	$3.34_{10}-2$
3	$6.28_{10}-1$	$2.55_{10}-2$	$8.40_{10}-3$	$2.53_{10}-2$
12000	$3.80_{10}-2$	$5.76_{10}-4$	$2.82_{10}-5$	$1.59_{10}-4$
13000	$3.52_{10}-2$	$5.47_{10}-4$	$2.81_{10}-5$	$1.16_{10}-4$
14000	$3.31_{10}-2$	$5.27_{10}-4$	$2.81_{10}-5$	$8.56_{10}-5$
14244	$3.27_{10}-2$	$5.24_{10}-4$	$2.81_{10}-5$	$7.95_{10}-5$

TABLE 1. *The RRGM*

For $i = 14244$ and $\kappa = 10^4$ the right-hand side of (2) is of order $(2 + C_1)10^{-2}\|A^{\frac{1}{2}}(\tilde{x} - x_0)\| + 10^{-4}\|A^{\frac{1}{2}}\|\|x\|$ and consequently, from (2) and (3) we obtain (approximately) the estimate

$$\|A^{\frac{1}{2}}(\tilde{x} - x_{14244})\| \leq (2 + C_1)10^{-1} \tag{4}$$

for the test of table 1. In view of the observed value of $\|A^{\frac{1}{2}}(\tilde{x} - x_{14244})\|$ this seems to be unsharp by at least a factor $\kappa^{\frac{1}{2}}$. This can be explained as follows. In section 2.3.1, we derived from the step-wise linear convergence of the natural error (cf. corollary 2.3.1.8)

$$\|A^{-\frac{1}{2}} r_{i+1}\|^2 \leq (1 - \gamma^2(1 + o(1)))\|A^{-\frac{1}{2}} r_i\|^2 , \tag{5}$$

the estimate (cf. (2.3.1.80))

$$(6)\qquad \Big(\sum_{\ell=0}^{\infty} \|A^{-\frac{1}{2}} r_\ell\|^2\Big)^{\frac{1}{2}} \le \|A^{-\frac{1}{2}} r_0\| \Big(\sum_{\ell=0}^{\infty} (1-\gamma^2(1+o(1)))^\ell\Big)^{\frac{1}{2}} =$$

$$= \gamma^{-1} \|A^{-\frac{1}{2}} r_0\| (1+o(1)) \ ,$$

which next was used in section 2.3.2, in order to prove (2) (see e.g. (2.3.2.26)). For the RRGM we have $\gamma^{-1} \le (\kappa+1)/(2\kappa^{\frac{1}{2}})$ (cf. (3.2.1)) so that from (6) it follows that approximately

$$(7)\qquad \Big(\sum_{\ell=0}^{\infty} \|A^{-\frac{1}{2}} r_\ell\|^2\Big)^{\frac{1}{2}} \le \tfrac{1}{2}\kappa^{\frac{1}{2}} \|A^{-\frac{1}{2}} r_0\| \ .$$

However, from the upper half of table 1 we see that in the first steps the natural error converges step-wise linearly with a convergence ratio that is much smaller than $(1-\gamma^2)^{\frac{1}{2}} \sim 1 - 2/\kappa$ and therefore (5) and consequently (7) is not sharp. For instance, if we only take into account the strong decrease of the natural error in the first step, then the following relations approximately hold

$$(8)\qquad \sum_{\ell=0}^{\infty} \|A^{-\frac{1}{2}} r_\ell\|^2 = \|A^{-\frac{1}{2}} r_0\|^2 + \sum_{\ell=1}^{\infty} \|A^{-\frac{1}{2}} r_\ell\|^2 \le$$

$$\le \|A^{-\frac{1}{2}} r_0\|^2 + \gamma^{-2} \|A^{-\frac{1}{2}} r_1\|^2 (1+o(1)) \le$$

$$\le \|A^{-\frac{1}{2}} r_0\|^2 (1 + \tfrac{1}{4}\kappa \|A^{-\frac{1}{2}} r_1\|^2 / \|A^{-\frac{1}{2}} r_0\|^2) =$$

$$= 1.1 \|A^{-\frac{1}{2}} r_0\|^2 \ .$$

This implies that we gain a factor of order $\kappa^{\frac{1}{2}}$ in comparison with estimate (7). It is easy to verify that the main term in the right-hand side of (2) can be lowered by a factor $\kappa^{\frac{1}{2}}$ when using estimate (8). As we can see from table 1, during the last thousands of iterations the convergence ratio of $\|A^{-\frac{1}{2}} r_1\|$ varies between $1 - 4/\kappa$ and $1 - 3/\kappa$ so that these natural errors hardly have an adverse effect on estimate (7). Since estimate (1) is based on estimate (2), the estimate (1) will also be unsharp by at least a factor $\kappa^{\frac{1}{2}}$ in cases of much faster convergence in the first steps.

REMARK 1. If in the test problem of table 1 we would restart the RRGM after the first step, which means that r_1 is not computed recursively but from $r_1 = fl(b - Ax_1)$, then we expect the further results to be only slightly different from the results in table 1. Then applying the analytical result (2) to the case with initial vector x_1 instead of x_0 would yield a rather sharp estimate for $\|A^{\frac{1}{2}}(\hat{x} - x_{14244})\|$, since then the strong decrease of the natural error $\|A^{-\frac{1}{2}} r_1\|$ in the first step is evaded. □

The residual $\|A(\hat{x} - x_i)\|$ seems to stagnate at the level 2.8_{10}-5 and hence it follows from this example that the RRGM is not well-behaved.

CHAPTER 4

THE CONJUGATE GRADIENT METHOD (CGM)

4.1. Introduction

The CGM was first described by Hestenes and Stiefel [52] and proposed as an iterative method for solving a definite linear system. Almost immediately the technique was extended to more general problems in the nonlinear programming field, where it proved to be extremely effective in dealing with general objective functions.

Algebraically, for linear problems, the CGM produces the solution $\hat{x} = A^{-1} b$ after at most n steps, but it is only slightly more complicated then the GM. In the presence of round-off, however, the n-th computed vector x_n generally is not even a reasonable approximation to $\hat{x}$ if the system is ill-conditioned. This is caused by the fact that the algebraic orthogonality relations are disturbed by round-off errors. For this reason the method saw little use as a method for solving linear systems until 1970, when it was shown by Reid [70] to be highly effective on some large, well-conditioned sparse systems. The most recent application of the CGM in connection with large sparse systems is first to transform the original system into another equivalent system that has a smaller condition and a more suitable spectrum, and next solve this preconditioned system by some version of the CGM (see e.g. Meijerink and van der Vorst [77], Kershaw [78], Manteuffel [80]).

Until now a few theoretical analyses have been carried out, explaining the numerical behavior of the CGM, but they are limited to an indication of some of the factors that influence the growth of round-off errors. As far as attainable accuracy is concerned, the only complete error analysis known to us at this time is given by Woźniakowski [80]. However, it analyzes a new version of the CGM, not contained in the paper of Hestenes and Stiefel [52], which is very closely related to the GM.

The CGM is defined by the following statements.

Conjugate Gradient Method (CGM)

Choose an initial point x_0;

$p_0 := r_0 := b - Ax_0$; $i := 0$;

while $r_i \neq 0 \wedge p_i \neq 0$ do

begin

(1) $\quad a_i := (r_i, p_i) / (p_i, Ap_i)$;

(2) $\quad x_{i+1} := x_i + a_i p_i$;

(3) $\quad r_{i+1} := \begin{cases} \text{either } b - Ax_{i+1} \\ \text{or} \quad r_i - A_i Ap_i \end{cases}$;

(4)

(5) $\quad b_i := - (r_{i+1}, Ap_i) / (p_i, Ap_i)$;

(6) $\quad p_{i+1} := r_{i+1} + b_i p_i$;

$\quad i := i + 1$

end.

We use either (3) at all steps or (4) at all steps and hence we disregard the mixed conjugate gradient method (MCGM) (cf. remark 2.2.1). In section 2.2 we stated already the basic idea behind the CGM: it is the conjugate direction method where the conjugate directions are chosen as an A-orthogonal version of the successive gradients. From the definition of b_i it follows immediately that p_{i+1}, computed from (6), satisfies $(p_{i+1}, Ap_i) = 0$. The fact that p_{i+1} and the other previous direction vectors p_ℓ ($\ell = 0,\ldots,i-1$) are conjugate with respect to A is stated in the following well-known theorem. The proof is straightforward, based on proving (i) and (ii) simultaneously by induction on i.

THEOREM 1. *If* $\{r_i\}$, $\{p_i\}$ *are generated by the CGM, then we have*

(i) $\quad \text{span}\{p_0,\ldots,p_i\} = \text{span}\{r_0,\ldots,r_i\} = \text{span}\{r_0,\ldots,A^i r_0\}$,

(ii) $\quad (p_i, Ap_\ell) = 0 \quad (\ell = 0,\ldots,i-1)$. □

In section 2.2 we established the following special proporties of the CGM.

THEOREM 2. *If* $\{r_i\}$, $\{p_i\}$ *are generated by the CGM, then we have*

(i) $\quad (r_{i+1},p_{i+1}) = (r_{i+1},r_{i+1}) \ ,$

(ii) $\quad \|A^{\frac{1}{2}} p_{i+1}\|^2 + \|b_i A^{\frac{1}{2}} p_i\|^2 = \|A^{\frac{1}{2}} r_{i+1}\|^2 \ ,$

(iii) $\quad \|p_i\|^2 = \|r_i\|^2 + b_{i-1}^2 \|p_{i-1}\|^2 \ ,$

(iv) $\quad a_i = (r_i,r_i) \,/\, (p_i,Ap_i) \ ,$

(v) $\quad b_i = (r_{i+1},r_{i+1}) \,/\, (r_i,r_i) \ .$ □

For the parameters α_i, β_i and γ_i, defined by (2.2.16), (2.2.17) and (2.2.8) we found (cf. (2.2.44), (2.2.45), (2.2.46))

$$\alpha_i = \frac{\|p_i\|}{\|r_i\|} \le \kappa^{\frac{1}{2}} \ , \qquad \beta_i = \frac{\|A^{\frac{1}{2}} p_i\|}{\|A^{\frac{1}{2}}\| \|r_i\|} \le 1 \ ,$$

$$\gamma_i^2 = \frac{\|r_i\|^4}{\|A^{-\frac{1}{2}} r_i\|^2 \|A^{\frac{1}{2}} p_i\|^2} \ge \frac{4\kappa}{(\kappa+1)^2} \ .$$

Consequently, from corollary 2.2.4 we obtain

$$(8) \qquad \frac{\|A^{\frac{1}{2}}(\hat{x} - x_{i+1})\|^2}{\|A^{\frac{1}{2}}(\hat{x} - x_i)\|^2} \le 1 - \frac{4\kappa}{(\kappa+1)^2} = \left(\frac{\kappa-1}{\kappa+1}\right)^2 \ ,$$

which reflects the step-wise linear convergence to zero of the natural error with a convergence ratio no greater than $(\kappa - 1) \,/\, (\kappa + 1)$.

From the next well-known considerations it will follow that the average convergence ratio of the natural error is no greater than $(\kappa^{\frac{1}{2}} - 1) \,/\, (\kappa^{\frac{1}{2}} + 1)$, which is essentially less than the convergence ratio.

Since $x_{i+1} = x_i + a_i p_i = x_0 + \Sigma_{\ell=0}^{i} a_\ell p_\ell$ and $\mathrm{span}\{p_0,p_1,\ldots,p_i\} = \mathrm{span}\{r_0,Ar_0,\ldots,A^i r_0\}$, there exists a polynomial Q_i of degree i such that $x_{i+1} = x_0 + Q_i(A) r_0$. As we observed already in section 2.2,

the residual r_{i+1} and the conjugate direction vectors $p_0,\dots,p_i$ satisfy $(r_{i+1},p_j) = 0$ $(j = 0,\dots,i)$ for every conjugate direction method. Since the objective function $F(x) := \|A^{\frac{1}{2}}(\hat{x} - x)\|^2$ is a strictly convex function and $\nabla F(x_{i+1}) = -r_{i+1}$, this implies that x_{i+1} not only minimizes $F(x)$ along the line $x = x_i + ap_i$ but on the whole affine set passing through x_0 and spanned by $p_0,p_1,\dots,p_i$.
Since $\operatorname{span}\{p_0,p_1,\dots,p_i\} = \operatorname{span}\{r_0,Ar_0,\dots,A^i r_0\}$, it follows that for any polynomial P_i of degree i there holds

$$\|A^{\frac{1}{2}}(\hat{x} - x_{i+1})\|^2 \le \|A^{\frac{1}{2}}(\hat{x} - (x_0 + P_i(A)r_0))\|^2 . \tag{9}$$

Expanding $\hat{x} - x_0$ in eigenvector components $\hat{x} - x_0 = \sum_{\ell=1}^{n} \xi_\ell u_\ell$, where $u_0,\dots,u_n$ are the eigenvectors of A, we obtain

$$\|A^{\frac{1}{2}}(\hat{x} - x_{i+1})\|^2 \le \sum_{\ell=1}^{n} \{1 - \lambda_\ell P_\ell(\lambda_\ell)\}^2 \lambda_\ell \xi_\ell^2 \le \tag{10}$$

$$\le \max_{\lambda_j} \{1 - \lambda_j P_i(\lambda_j)\}^2 \|A^{\frac{1}{2}}(\hat{x} - x_0)\|^2$$

for any polynomial P_i of degree i. Now select $P_i(\lambda)$ so that

$$1 - \lambda P_i(\lambda) = T_{i+1}\left(\frac{\lambda_n + \lambda_1 - 2\lambda}{\lambda_n - \lambda_1}\right) / T_{i+1}\left(\frac{\lambda_n + \lambda_1}{\lambda_n - \lambda_1}\right) , \tag{11}$$

where

$$T_{i+1}(z) := \tfrac{1}{2}\{(z + (z^2 - 1)^{\frac{1}{2}})^{i+1} + (z + (z^2 - 1)^{\frac{1}{2}})^{-i-1}\}$$

is the (i+1)-th Chebyshev polynomial (the formula for $T_{i+1}(z)$ can be used for all z even though in some cases the intermediate quantities may be complex). For this choice of P_i we obtain

$$\max_{\lambda_j} \{1 + \lambda_j P_i(\lambda_j)\}^2 = \left\{T_{i+1}\left(\frac{\lambda_n + \lambda_1}{\lambda_n - \lambda_1}\right)\right\}^{-2} \le 4\left(\frac{\kappa^{\frac{1}{2}} - 1}{\kappa^{\frac{1}{2}} + 1}\right)^{2(i+1)} . \tag{12}$$

In view of this we have

THEOREM 3. *If $\{x_i\}$ is generated by the CGM, then the natural error satisfies*

$$\|A^{\frac{1}{2}}(\hat{x} - x_i)\| \leq 2\left(\frac{\kappa^{\frac{1}{2}} - 1}{\kappa^{\frac{1}{2}} + 1}\right)^i \|A^{\frac{1}{2}}(\hat{x} - x_0)\| \ . \tag{13}$$ □

Another well-known algebraic property of the CGM is the monotonic decrement of the error $\|\hat{x} - x_i\|$.

THEOREM 4. *If* $\{x_i\}$ *is generated by the CGM, then the error satisfies*

$$\frac{\|\hat{x} - x_{i+1}\|^2}{\|\hat{x} - x_i\|^2} = 1 - \frac{\|p_i\|^2}{\|A^{-1} r_i\|^2 \|A^{\frac{1}{2}} p_i\|^2} (\|A^{-\frac{1}{2}} r_i\|^2 + \|A^{-\frac{1}{2}} r_{i+1}\|^2) \ . \tag{14}$$ □

Hestenes and Stiefel [52] gave a proof of (14) using backward induction based on the fact that $x_{i+1} = \hat{x}$ for some $i < n$. Kammerer and Nashed [72] gave a proof by forward induction, that is also valid in the Hilbert space case. Since $\|p_i\| \geq \|r_i\|$ and $\|A^{\frac{1}{2}} p_i\| \leq \|A^{\frac{1}{2}} r_i\|$ we have the following corollary of theorem 4.

COROLLARY 5. *If* $\{x_i\}$ *is generated by the CGM, then the error converges step-wise linearly to zero and*

$$\frac{\|\hat{x} - x_{i+1}\|^2}{\|\hat{x} - x_i\|^2} \leq 1 - \frac{1}{\kappa} \ . \tag{15}$$ □

We note that, just like the GM, the CGM is invariant relative to orthogonal basis transformations (cf. remark 3.1.3). Furthermore, we observe that it follows from theorem 2(i) that $r_{i+1} \neq 0$ implies $p_{i+1} \neq 0$ and hence the CGM terminates because of the fact that $r_{i+1} = 0$ as well as $p_{i+1} = 0$. Consequently, the condition $p_{i+1} \neq 0$ could be left out in the stopping criterion.

We conclude this section with another remarkable algebraic property of the CGM. Suppose we choose an arbitrary initial vector x_0, set $r_0 := b - Ax_0$ and instead of setting $p_0 := r_0$ we choose an arbitrary initial vector $p_0 \neq 0$ to start the CGM-iterations. Retracing the proofs in section 2.2 of the relations stated in theorem 2, it turns out that the relations (i), (ii) still hold for $i \geq 0$ and that the

relations (iii), (iv) and (v) still hold for $i \geq 1$. Consequently, this so-called *independent start conjugate gradient method* (ISCGM) ($p_0 \neq r_0$) is a DM for which the parameters α_i, β_i and γ_i satisfy the inequalities stated in (7) for $i \geq 1$. Hence, apart from the first step, the natural error converges step-wise linearly to zero with a convergence ratio no greater than $(\kappa - 1) / (\kappa + 1)$. As far as the first step is concerned, we have $\gamma_0 := |(r_0,p_0)| / (\|A^{-\frac{1}{2}} r_0\| \|A^{\frac{1}{2}} p_0\|)$ and consequently (cf. (i) of theorem 2.2.2) $\|A^{\frac{1}{2}}(\hat{x} - x_1)\| \leq \|A^{\frac{1}{2}}(\hat{x} - x_0)\|$, with equality iff $(r_0,p_0) = 0$. On the other hand, the results of theorem 3 and theorem 4 do certainly not hold for the ISCGM, since their proofs are strongly based on orthogonality relations like $(r_i,p_j) = 0$ $(i > j)$, $(p_i,Ap_j) = 0$ $(i \neq j)$, and these orthogonality relations are based on the fact that $p_0 = r_0$.

The numerical importance of the ISCGM can be explained as follows. Suppose we have performed k steps of the CGM in the presence of round-off. It is obvious that we may not expect the relations of theorem 2 to hold exactly. Without round-off, continuing the CGM after these k steps is equivalent to starting the ISCGM with initial vectors x_k and p_k. Consequently, we may conclude from the previous considerations that at all these later steps the natural error converges step-wise linearly to zero. This can be considered as a stability property of the CGM: if after some iteration step the occurrence of round-off is excluded permanently, then from this step on the natural error converges step-wise linearly to zero.

In Chapter 5 we deduce some more properties of the ISCGM and we also consider independent start versions of other variants of the CGM.

In the next two sections we discuss the numerical analogues of the inequalities given in (7) which will next be used to prove a numerical analogue of the step-wise linear convergence reflected by (8). In section 2 this is done for the RRCGM; in section 3 this is done for the TRCGM. In section 4 we report on numerical experiments that have been carried out with the CGM where also remarks concerning the validity of theorem 3 and corollary 5 in the presence of round-off are included.

4.2. *The recursive residual conjugate gradient method (RRCGM)*

The results deduced in this section will be based on the results of section 2.3 for general RRDM's. The results there were expressed in terms of the parameters α_i, β_i and γ_i, defined by (2.3.1.43), (2.3.1.44) and (2.3.1.4). Therefore, we have to estimate these parameters for the RRGM.

We first repeat the estimation of these parameters in the algebraic case, since the estimation in the presence of round-off proceeds along the same lines.

Since, algebraically, $A^{\frac{1}{2}} p_{i+1} - b_i A^{\frac{1}{2}} p_i = A^{\frac{1}{2}} r_{i+1}$, where b_i is chosen such that $(p_{i+1}, Ap_i) = 0$, it follows by taking squared norms at both sides that

$$\|A^{\frac{1}{2}} p_{i+1}\|^2 + \|b_i A^{\frac{1}{2}} p_i\|^2 = \|A^{\frac{1}{2}} r_{i+1}\|^2 , \tag{1}$$

or equivalently,

$$\frac{\|A^{\frac{1}{2}} p_{i+1}\|^2}{\|A^{\frac{1}{2}} r_{i+1}\|^2} = 1 - \theta_i^2 , \tag{2}$$

where

$$\theta_i := |(r_{i+1}, Ap_i)| / (\|A^{\frac{1}{2}} r_{i+1}\| \|A^{\frac{1}{2}} p_i\|) . \tag{3}$$

In particular we have

$$\|A^{\frac{1}{2}} p_{i+1}\| \leq \|A^{\frac{1}{2}} r_{i+1}\| . \tag{4}$$

Algebraically we have for any DM (cf. theorem 2.2.2(iii))

$$(r_{i+1}, p_i) = 0 , \tag{5}$$

and since in the CGM $p_{i+1} = r_{i+1} + b_i p_i$, we obtain

$$(r_{i+1}, p_{i+1}) = \|r_{i+1}\|^2 . \tag{6}$$

The formulas (4) and (6) are used to estimate the three parameters α_i, β_i and γ_i in the algebraic case. According to the definitions we obtain

(7) $$\alpha_i := \|r_i\|\|p_i\| / |(r_i,p_i)| = \|p_i\| / \|r_i\| \leq$$

$$\leq \kappa^{\frac{1}{2}}\|A^{\frac{1}{2}} p_i\| / \|A^{\frac{1}{2}} r_i\| \leq \kappa^{\frac{1}{2}} ,$$

(8) $$\beta_i := \|r_i\| \|A^{\frac{1}{2}} p_i\| / (\|A^{\frac{1}{2}}\| |(r_i,p_i)|) = \|A^{\frac{1}{2}} p_i\| / (\|A^{\frac{1}{2}}\|\|r_i\|) \leq 1 ,$$

(9) $$\gamma_i^{-1} := \|A^{-\frac{1}{2}} r_i\|\|A^{\frac{1}{2}} p_i\| / |(r_i,p_i)| \leq \|A^{-\frac{1}{2}} r_i\|\|A^{\frac{1}{2}} r_i\| / \|r_i\|^2 \leq$$

$$\leq (\kappa + 1) / (2\kappa^{\frac{1}{2}}) .$$

In order to estimate the parameters in the presence of round-off it seems necessary to investigate first the analogues of the formulas (4) and (6) and intermediately the analogues of (2) and (5). Observe that the algebraic relation (1) is quite similar to the algebraic relation (i) of theorem 2.2.2 whose numerical analogues are stated in theorem 2.3.1, theorem 2.3.4 and remark 2.3.2. Therefore it is not surprising that the proof of the numerical analogue of (1) is similar to the proofs of these theorems.

THEOREM 1. *Let* r_{i+1}, p_i *be two arbitrary nonzero machine vectors and let* p_{i+1} *be computed according to (4.1.5) and (4.1.6). Furthermore, let*

(10) $$\theta_i := |(r_{i+1},Ap_i)| / (\|A^{\frac{1}{2}} r_{i+1}\|\|A^{\frac{1}{2}} p_i\|) .$$

Then we have

(11) $$\frac{\|A^{\frac{1}{2}} p_{i+1}\|^2}{\|A^{\frac{1}{2}} r_{i+1}\|^2} = 1 - \theta_i^2 + \mu_{i+1} ,$$

where

(12) $$|\mu_{i+1}| \leq 12\varepsilon\kappa^{\frac{1}{2}}(1 + o(1)) + \varepsilon\kappa^{\frac{1}{2}}(C_2 + C_1\kappa^{\frac{1}{2}})o(1) ,$$

under the restriction

(13) $$\varepsilon\kappa^{\frac{1}{2}}(1 + C_2 + C_1\kappa^{\frac{1}{2}}) \to 0 .$$

PROOF. The proof is entirely analogous to the proof of theorem 2.3.1. We only state the main intermediate results, that will also be used later on.

We obtain for the computation of $b_i = fl(-(r_{i+1},Ap_i) / (p_i,Ap_i))$

$$b_i = \hat{b}_i + \delta b_i \ , \tag{14}$$

$$\hat{b}_i := -(r_{i+1},Ap_i) / (p_i,Ap_i) \ , \tag{15}$$

$$|\delta b_i'| \leq \varepsilon(1 + 2C_2\kappa^{\frac{1}{2}} + 2C_1\kappa)\|A^{\frac{1}{2}} r_{i+1}\| / \|A^{\frac{1}{2}} p_i\|(1 + o(1)) \ , \tag{16}$$

under the restriction

$$\varepsilon(1 + C_2\kappa^{\frac{1}{2}} + C_1\kappa) \to 0 \ . \tag{17}$$

The computation of $p_{i+1} = fl(r_{i+1} + b_i p_i)$ satisfies

$$p_{i+1} = r_{i+1} + b_i p_i + \delta p_{i+1}' \ , \tag{18}$$

$$\delta p_{i+1}' := F_i'' r_{i+1} + b_i V_i' p_i \ , \tag{19}$$

$$\|F_i''\| \leq \varepsilon \ , \quad \|V_i'\| \leq 2\varepsilon(1 + o(1)) \ , \quad [\varepsilon \to 0] \ , \tag{20}$$

and also

$$p_{i+1} = r_{i+1} + \hat{b}_i p_i + \delta p_{i+1}'' \ , \tag{21}$$

$$\delta p_{i+1}'' := F_i'' r_{i+1} + \delta b_i' p_i + I_i' p_i \ , \tag{22}$$

$$\|I_i'\| \leq 2\varepsilon\|A^{\frac{1}{2}} r_{i+1}\| / \|A^{\frac{1}{2}} p_i\|(1 + o(1)) \ , \tag{23}$$

under the restriction (17).
From (18) we derive the basic formula

$$\frac{\|A^{\frac{1}{2}} p_{i+1}\|^2}{\|A^{\frac{1}{2}} r_{i+1}\|^2} = 1 - \theta_i^2 + \mu_{i+1} \ , \tag{24}$$

where θ_i is defined by (10) and

$$\mu_{i+1} := \{2(A(r_{i+1} + \hat{b}_i p_i), \delta p_{i+1}'') + \|A^{\frac{1}{2}} \delta p_{i+1}''\|^2\} / \|A^{\frac{1}{2}} r_{i+1}\|^2 \ . \tag{25}$$

Using the former inequalities one can prove that (12) holds under the restriction (13). □

REMARK 2. Theorem 1 can also be written in a form closer related to expression (1). We have

(26) $$\theta_i^2 = (r_{i+1},Ap_i)^2 / (\|A^{\frac{1}{2}} r_{i+1}\| \|A^{\frac{1}{2}} p_i\|)^2 = \|\hat{b}_i A^{\frac{1}{2}} p_i\|^2 / \|A^{\frac{1}{2}} r_{i+1}\|^2 ,$$

where, according to (15), $\hat{b}_i := - (r_{i+1},Ap_i) / (p_i,Ap_i)$.
Consequently, (11) can be written as

$$\|A^{\frac{1}{2}} p_{i+1}\|^2 + \|\hat{b}_i A^{\frac{1}{2}} p_i\|^2 = (1+\mu_{i+1}) \|A^{\frac{1}{2}} r_{i+1}\|^2 .$$

It follows in particular that, under the restriction (13),

(27) $$\begin{cases} \|A^{\frac{1}{2}} p_{i+1}\| \leq (1+o(1)) \|A^{\frac{1}{2}} r_{i+1}\| , \\ \|\hat{b}_i A^{\frac{1}{2}} p_i\| \leq (1+o(1)) \|A^{\frac{1}{2}} r_{i+1}\| . \end{cases}$$

Using (14) and (16) this yields, under the restriction (13),

(28) $$\|b_i A^{\frac{1}{2}} p_i\| \leq (1+o(1)) \|A^{\frac{1}{2}} r_{i+1}\| .$$

Retracing the proof of theorem 1 and replacing all o-symbols by definite estimates involving explicit numerical constants, one can prove that $|\mu_{i+1}| \leq 7/40$ if

(29) $$\varepsilon \kappa^{\frac{1}{2}} (3 + C_2 + C_1 \kappa^{\frac{1}{2}}) \leq \frac{1}{40} .$$

Hence

(30) $$\begin{cases} \|A^{\frac{1}{2}} p_{i+1}\| \leq (1 + \frac{1}{10}) \|A^{\frac{1}{2}} r_{i+1}\| , \\ \|\hat{b}_i A^{\frac{1}{2}} p_i\| \leq (1 + \frac{1}{10}) \|A^{\frac{1}{2}} r_{i+1}\| . \end{cases}$$

Furthermore, it follows that

(31) $$\|b_i A^{\frac{1}{2}} p_i\| \leq (1 + \frac{2}{10}) \|A^{\frac{1}{2}} r_{i+1}\| .$$ □

We now derive the numerical analogue of (6). In remark 2.3.1.16 we saw already that the orthogonality of r_{i+1} and p_i can be seriously disturbed by round-off if $\|A^{-\frac{1}{2}} r_i\| / \|A^{-\frac{1}{2}} r_{i+1}\|$ is large. It is obvious that this loss of orthogonality influences the approximate validity of (6).

THEOREM 3. *Let r_i, p_i be two arbitrary nonzero machine vectors and let r_{i+1}, p_{i+1} be computed from one step RRCGM. Then we have*

(32) $$(r_{i+1},p_{i+1}) = (1+\lambda_{i+1})\|r_{i+1}\|^2 ,$$

where

(33) $$|\lambda_{i+1}| \le \left\{\varepsilon(1+2\kappa^{\frac12}) + \varepsilon\kappa^{\frac12}(1+3\kappa^{\frac12}+2C_2\kappa^{\frac12}+2C_1\kappa)\frac{\|A^{-\frac12}r_i\|}{\|A^{-\frac12}r_{i+1}\|}\right\}(1+o(1)) ,$$

under the restriction

(34) $$\varepsilon\kappa^{\frac12}(1+C_2+C_1\kappa^{\frac12}) \to 0 .$$

PROOF. From (2.3.1.14) and (2.3.1.24) we obtain

(35) $$(r_{i+1},p_i) = (r_i,p_i) - a_i(p_i,Ap_i) + (\delta r'_{i+1},p_i) =$$

$$= -\delta a'_i(p_i,Ap_i) + (\delta r'_{i+1},p_i) .$$

Hence, using (15), we get

(36) $$(r_{i+1},p_{i+1}) = (r_{i+1},r_{i+1}) + b_i(r_{i+1},p_i) + (r_{i+1},\delta p'_{i+1}) =$$

$$= (r_{i+1},r_{i+1}) - \delta a'_i b_i(p_i,Ap_i) + b_i(\delta r'_{i+1},p_i) + (r_{i+1},\delta p'_{i+1}) .$$

Consequently,

(37) $$(r_{i+1},p_{i+1}) = (1+\lambda_{i+1})\|r_{i+1}\|^2 ,$$

where

(38) $$\lambda_{i+1} := \{b_i(\delta r'_{i+1},p_i) + (r_{i+1},\delta p'_{i+1}) - \delta a'_i b_i(p_i,Ap_i)\} / \|r_{i+1}\|^2 .$$

From (25), (26), (27) and remark 2 of section 2.3.1 and from (28) of this section we obtain, under the restriction (34)

(39) $$|b_i(\delta r'_{i+1},p_i)| / \|r_{i+1}\|^2 \le \|b_i A^{\frac12}p_i\| \, \|A^{-\frac12}\delta r'_{i+1}\| / \|r_{i+1}\|^2 \le$$

$$\le \|b_i A^{\frac12}p_i\|\{\varepsilon\kappa^{\frac12}\|A^{-\frac12}r_i\| +$$

$$+ (2\varepsilon\kappa^{\frac12}\|a_i A^{\frac12}p_i\| + \varepsilon C_1\kappa\|a_i A^{\frac12}p_i\|)(1+o(1))\} / \|r_{i+1}\|^2 \le$$

$$\le \varepsilon\kappa^{\frac12}(3+C_1\kappa^{\frac12})\|A^{\frac12}r_{i+1}\|\|A^{-\frac12}r_i\| / \|r_{i+1}\|^2(1+o(1)) \le$$

$$\le \varepsilon\kappa(3+C_1\kappa^{\frac12})\|A^{-\frac12}r_i\| / \|A^{-\frac12}r_{i+1}\|(1+o(1)) .$$

From (19), (20) and (28) we obtain, under the restriction (34)

$$(40)\qquad |(r_{i+1},\delta p'_{i+1})| / \|r_{i+1}\|^2 \le \|\delta p'_{i+1}\| / \|r_{i+1}\| \le$$

$$\le (\varepsilon\|r_{i+1}\| + 2\varepsilon\|A^{-\frac{1}{2}}\|\|b_i A^{\frac{1}{2}} p_i\|) / \|r_{i+1}\| \le$$

$$\le \varepsilon(\|r_{i+1}\| + 2\|A^{-\frac{1}{2}}\|\|A^{\frac{1}{2}} r_{i+1}\| (1 + o(1)) / \|r_{i+1}\|^2 \le$$

$$\le \varepsilon(1 + 2\kappa^{\frac{1}{2}})(1 + o(1)) \ .$$

From (2.3.1.21) and (28) we obtain, under the restriction (34)

$$(41)\qquad |\delta a'_i\, b_i(p_i, Ap_i)| / \|r_{i+1}\|^2 \le$$

$$\le \varepsilon(1 + 2C_2\kappa^{\frac{1}{2}} + C_1\kappa)\, \|A^{-\frac{1}{2}} r_i\|\|b_i A^{\frac{1}{2}} p_i\| / \|r_{i+1}\|^2 (1 + o(1)) \le$$

$$\le \varepsilon(1 + 2C_2\kappa^{\frac{1}{2}} + C_1\kappa)\|A^{-\frac{1}{2}} r_i\|\|A^{\frac{1}{2}} r_{i+1}\| / \|r_{i+1}\|^2 (1 + o(1)) \le$$

$$\le \varepsilon\kappa^{\frac{1}{2}}(1 + 2C_2\kappa^{\frac{1}{2}} + C_1\kappa)\|A^{-\frac{1}{2}} r_i\| / \|A^{-\frac{1}{2}} r_{i+1}\| (1 + o(1)) \ .$$

Inequality (33) now follows from (38), (39), (40) and (41). □

A more explicit version of the result just derived is given without proof. We have

PROPOSITION 4. *Let* r_i, p_i *be two arbitrary nonzero machine vectors and let* r_{i+1}, p_{i+1} *be computed from one step RRCGM.*
If

$$(42)\qquad \varepsilon\kappa^{\frac{1}{2}}(3 + C_2 + C_1\kappa^{\frac{1}{2}}) \le \frac{1}{40} \ ,$$

then, in equality (33), certainly

$$(43)\quad |\lambda_{i+1}| \le \left\{\varepsilon(1 + 2\kappa^{\frac{1}{2}}) + \varepsilon\kappa^{\frac{1}{2}}(1 + 3\kappa^{\frac{1}{2}} + 2C_2\kappa^{\frac{1}{2}} + 2C_1\kappa)\, \frac{\|A^{-\frac{1}{2}} r_i\|}{\|A^{-\frac{1}{2}} r_{i+1}\|}\right\}(1 + \frac{3}{10}) \ .$$

□

Once we have derived the numerical analogues of (4) and (6), it is easily seen that in the presence of round-off the parameters α_i, β_i and γ_i satisfy

$$(44)\qquad \begin{cases} \alpha_i \le \kappa^{\frac{1}{2}}(1+o(1)) \ , \quad \beta_i \le 1+o(1)) \ , \\ \gamma_i^{-1} \le (\kappa+1)/(2\kappa^{\frac{1}{2}})(1+o(1)) \ , \end{cases}$$

under the restriction

$$(45)\qquad \varepsilon\kappa^{\frac{1}{2}}(1+C_2+C_1\kappa^{\frac{1}{2}}) + \varepsilon\kappa(1+C_2+C_1\kappa^{\frac{1}{2}}) \frac{\|A^{-\frac{1}{2}} r_{i-1}\|}{\|A^{-\frac{1}{2}} r_i\|} \to 0 \ .$$

This implies that if $\|A^{-\frac{1}{2}} r_i\|$ is not very small relative to $\|A^{-\frac{1}{2}} r_{i-1}\|$ (and $\varepsilon\kappa(1+C_2+C_1\kappa^{\frac{1}{2}})$ is appreciably less than unity), then the bounds for α_i, β_i and γ_i in the presence of round-off are close to the bounds for α_i, β_i and γ_i in the algebraic case. Consequently, it then follows from proposition 2.3.1.12 that $\|A^{-\frac{1}{2}} r_{i+1}\| < \|A^{-\frac{1}{2}} r_i\|$. Stated differently, if $\varepsilon\kappa(1+C_2+C_1\kappa^{\frac{1}{2}})$ is appreciably less than unity, then $\|A^{-\frac{1}{2}} r_{i+1}\| < \|A^{-\frac{1}{2}} r_i\|$ unless $\|A^{-\frac{1}{2}} r_i\| << \|A^{-\frac{1}{2}} r_{i-1}\|$. This justifies the expectation that eventually in all cases $\|A^{-\frac{1}{2}} r_{i+1}\| < \|A^{-\frac{1}{2}} r_{i-1}\|$. Concrete form is given to this expectation in the following proposition. As in the case of a general DM we are not primarily interested in the fact that the numerical convergence ratio is close to the algebraic ratio if (45) is small, but we want to know under what explicit conditions the natural error $\|A^{-\frac{1}{2}} r_i\|$ tends to zero. Because of this a version of proposition 5 using o-symbols is omitted. We only state an explicit version.

PROPOSITION 5. *Let* r_{i-1}, p_{i-1} *be two arbitrary nonzero machine vectors and let*

$$(46)\qquad L := \left\{1 - \frac{5}{16}\frac{\kappa}{(\kappa+1)^2}\right\}^{\frac{1}{2}} \ .$$

Consider two successive steps of the RRCGM. If

$$(47)\qquad \varepsilon\kappa(3+C_2+C_1\kappa^{\frac{1}{2}}) \le \frac{1}{40} \ ,$$

then at least one of the following two inequalities

$$(48)\qquad \frac{\|A^{-\frac{1}{2}} r_{i+1}\|^2}{\|A^{-\frac{1}{2}} r_i\|^2} \le L^2 \ ,$$

(49) $$\frac{\|A^{-\frac{1}{2}} r_{i+1}\|^2}{\|A^{-\frac{1}{2}} r_{i-1}\|^2} \leq L^4$$

holds.

PROOF. Note that $L^2 \geq 1-5/64 = 59/64$. If $\|A^{-\frac{1}{2}} r_{i+1}\|^2 \leq L^4 \|A^{-\frac{1}{2}} r_{i-1}\|^2$, then we are ready.
Let us now assume that $\|A^{-\frac{1}{2}} r_{i+1}\|^2 > L^4 \|A^{-\frac{1}{2}} r_{i-1}\|^2$. If (46) is satisfied, then (cf. remark 2.3.1.2) one certainly has $\|A^{-\frac{1}{2}} r_{i+1}\| \leq$ $\leq (11/10) \|A^{-\frac{1}{2}} r_i\|$ and consequently

$$\frac{\|A^{-\frac{1}{2}} r_{i+1}\|}{\|A^{-\frac{1}{2}} r_i\|} = \frac{\|A^{-\frac{1}{2}} r_{i-1}\|}{\|A^{-\frac{1}{2}} r_{i+1}\|} \frac{\|A^{-\frac{1}{2}} r_{i+1}\|}{\|A^{-\frac{1}{2}} r_i\|} < \frac{11}{10L^2} < \frac{4}{3} .$$

Consequently, from proposition 4 and assumption (46) it follows that

(50) $$(r_i, p_i) = (1+\lambda_i)\|r_i\|^2 ,$$

where

(51) $$|\lambda_i| \leq (\tfrac{13}{10})\{\varepsilon(2+\kappa^{\frac{1}{2}}) + (\tfrac{4}{3})\varepsilon\kappa^{\frac{1}{2}}(1+3\kappa^{\frac{1}{2}} + 2C_2\kappa^{\frac{1}{2}} + 2C_1\kappa)\} \leq$$

$$\leq 4\varepsilon\kappa(3 + C_2 + C_1\kappa^{\frac{1}{2}}) \leq \frac{1}{10} .$$

In view of (30) we furthermore have

(52) $$\|A^{\frac{1}{2}} p_i\| \leq (\tfrac{11}{10})\|A^{\frac{1}{2}} r_i\| .$$

Hence we obtain the explicit bounds

(53) $$\alpha_i \leq \kappa^{\frac{1}{2}}(\tfrac{11}{10})(1-\tfrac{1}{10})^{-1} < 2\kappa^{\frac{1}{2}} , \quad \beta_i < 2 , \quad \gamma_i^{-1} \leq (\kappa+1)/\kappa^{\frac{1}{2}} .$$

Consequently, the assumption of proposition 2.3.1.12 certainly is satisfied and we conclude

(54) $$\frac{\|A^{-\frac{1}{2}} r_{i+1}\|^2}{\|A^{-\frac{1}{2}} r_i\|^2} \leq 1 - \frac{5}{16}\gamma_i^2 \leq 1 - \frac{5}{16}\frac{\kappa}{(\kappa+1)^2} = L^2 .$$ □

REMARK 6. Since in proposition 5, r_{i-1} and p_{i-1} are arbitrary, the assertion given in (48) and (49) holds for all $i \geq 1$ not only for the

RRCGM, but also for the RRISCGM, as introduced in section 4.1. Proposition 5 indicates that if in a certain step the natural error does not decrease by a factor L, then it did decrease by at least a factor L^2 in the last two steps together. It is easily seen that the assertion given in (48) and (49) implies the assertion that for every $i \geq 2$

$$(55) \qquad \frac{\|A^{-\frac{1}{2}} r_i\|}{\|A^{-\frac{1}{2}} r_{i-2}\|} \leq L^2 \quad \text{or} \quad \frac{\|A^{-\frac{1}{2}} r_{i+1}\|}{\|A^{-\frac{1}{2}} r_{i-1}\|} \leq L^2 .$$

Therefore we have a kind of *bi-step-wise linear convergence* to zero of the natural error $\|A^{-\frac{1}{2}} r_i\|$ with a convergence ratio no greater than L. □

The linear convergence to zero on the average of the natural error is expressed in the following corollary of proposition 5.

COROLLARY 7. *Consider the RRCGM with arbitrary initial vector* x_0 *and assume that*

$$(56) \qquad \varepsilon\kappa(3 + C_2 + C_1\kappa^{\frac{1}{2}}) \leq \frac{1}{40} .$$

Then we have for $i \geq 0$

$$(57) \qquad \|A^{-\frac{1}{2}} r_i\| \leq L^2 \|A^{-\frac{1}{2}} r_0\| ,$$

where L *is defined by (46).*

PROOF. For i = 0 inequality (57) is trivially satisfied. For i = 1 (57) follows immediately from proposition 3.2.2 since the first step of the RRCGM is identical to the first step of the RRGM.
Now let $i \geq 2$ and suppose (57) holds for all $0 \leq k \leq i-1$. If $\|A^{-\frac{1}{2}} r_i\| / \|A^{-\frac{1}{2}} r_{i-1}\| \leq L$, then

$$(58) \qquad \|A^{-\frac{1}{2}} r_i\| / \|A^{-\frac{1}{2}} r_0\| \leq$$

$$\leq (\|A^{-\frac{1}{2}} r_i\| / \|A^{-\frac{1}{2}} r_{i-1}\|)(\|A^{-\frac{1}{2}} r_{i-1}\| / \|A^{-\frac{1}{2}} r_0\|) \leq L \cdot L^{i-1} = L^i .$$

If $\|A^{-\frac{1}{2}} r_i\| / \|A^{-\frac{1}{2}} r_{i-1}\| > L$, then, by proposition 5, one certainly has $\|A^{-\frac{1}{2}} r_i\| / \|A^{-\frac{1}{2}} r_{i-2}\| \leq L^2$, and therefore

(59) $$\|A^{-\frac{1}{2}} r_i\| / \|A^{-\frac{1}{2}} r_0\| \leq$$

$$\leq (\|A^{-\frac{1}{2}} r_i\| / \|A^{-\frac{1}{2}} r_{i-1}\|)(\|A^{-\frac{1}{2}} r_{i-2}\| / \|A^{-\frac{1}{2}} r_0\|) \leq L^2 \cdot L^{i-2} = L^i .$$

Hence, in either case (57) also holds for i and (50) follows by induction. □

REMARK 8. For the RRISCGM we cannot apply proposition 3.2.2 (which applies to the RRCGM) for the first step. We can apply, however, remark 2.3.1.2 which guarantees that $\|A^{-\frac{1}{2}} r_1\| \leq (\frac{11}{10})\|A^{-\frac{1}{2}} r_0\|$. From this and proposition 5 it follows, similar to corollary 7, that for the RRISCGM one has for all $i \geq 0$

(60) $$\|A^{-\frac{1}{2}} r_i\| \leq (\frac{11}{10}) L^{i-1} \|A^{-\frac{1}{2}} r_0\| .$$ □

REMARK 9. The restriction for the RRGM in proposition 3.2.2 and the restriction for RRCGM in proposition 2.4.5 essentially differ by a factor $\kappa^{\frac{1}{2}}$ in favour of the RRGM. The basic restriction (2.3.1.70) for general DM's, under which these proposition were derived, contains the term $\alpha(1 + C_2 + C_1\kappa)$ and since (even algebraically) the bounds for α differ by a factor $\kappa^{\frac{1}{2}}$ in the case of RRGM and RRCGM, the difference is not surprising. □

We do not present a numerical analogue of theorem 4.1.3 which states that algebraically the average convergence ratio is no greater than $(\kappa^{\frac{1}{2}} - 1) / (\kappa^{\frac{1}{2}} + 1)$. In practice, this algebraic average convergence ratio is observed (see section 4.4). The proof of theorem 4.1.3 is based on the whole history of the RRCGM, particularly on the orthogonality relations $(r_{i+1}, p_j) = 0$ $(j < i)$. In section 2.3.1 we considered only one step RRDM and in remark 2.3.1.16 we discussed the numerical analogue of only one orthogonality relation $(r_{i+1}, p_i) = 0$. In order to obtain a numerical analogue of theorem 4.1.3 it seems necessary to investigate the approximate validity of all orthogonality relations, but this is outside the scope of this thesis. However, we do realize that it is because of the smaller average linear convergence ratio that in practice the CGM is far superior to the GM. For the same reasons we do not present a numerical analogue of theorem 4.1.4.

Recall that, using recursive residuals, $\|A^{-\frac{1}{2}} r_i\| \to 0$ does not imply that $\|x_i - \hat{x}\| \to 0$. However, we cannot apply the results of section 3.2.2 without more ado. This is due to the fact that in the general case we assumed step-wise linear convergence of $\{A^{-\frac{1}{2}} r_i\}$ whereas for the RRCGM we only proved bi-step-wise linear convergence in the sense of (55). However, it is still possible to derive similar results (see Bollen [79] for a rather weak result), but we refrain from stating them here.

4.3. *The true residual conjugate gradient method* (TRCGM)

The results deduced in this section will be based on the results of section 2.4 for general TRDM's. The results there were expressed in terms of the parameters α_i, β_i and γ_i defined by (2.4.20), (2.4.51) and (2.4.52), respectively. Therefore, we have to estimate these parameters for the TRCGM. Of course, as far as the algebraic processes TRCGM and RRCGM are concerned, there is no difference between the estimates. Consequently, we proceed along the same lines as in section 4.2 and investigate the numerical analogues of (4.2.4) and (4.2.6). With respect to the numerical analogue of (4.2.4) we can use the results of theorem 4.2.1 and remark 4.2.2, on the understanding that now r_{i+1} stands for the computed true residual $r_{i+1} := fl(b - Ax_i)$. The results deduced in section 4.2 cannot be used to obtain the numerical analogue of (4.2.6) but nevertheless the proof of the numerical analogue of (4.2.6) for the TRCGM is very similar to the proof of theorem 4.2.3.
We first establish an auxiliary result.

LEMMA 1. *Let* x_i, p_i *be two arbitrary machine vectors and let* x_{i+1} *be computed according to (4.1.1), (4.1.2) and (4.1.5). Furthermore, let according to the definitions (2.4.19) and (2.4.21)*

(1) $$\hat{r}_i := b - Ax_i \ ,$$

(2) $$\psi_i := \|A^{\frac{1}{2}}\| \|x_i\| \, / \|A^{-\frac{1}{2}} \hat{r}_i\| \ .$$

Then we have

(3) $$\|x_i\| \le (\|x_{i+1}\| + \|A^{-\frac{1}{2}}\|\|A^{-\frac{1}{2}} f_i\|)\,(1+o(1)) ,$$

under the restriction

(4) $$\varepsilon\{\kappa^{\frac{1}{2}}(1+C_2+C_1\kappa^{\frac{1}{2}}) + (1+C_1\kappa^{\frac{1}{2}}\psi_i)\} \to 0 .$$

PROOF. All o-symbols are assumed to hold under the restriction (4). From (2.4.38) we obtain

(5) $$\|x_i\| \le \|x_{i+1}\| + \|\hat{a}_i p_i\| + \|\delta x_{i+1}\| .$$

From (2.4.39), (2.4.41) and (2.4.37) we obtain

(6) $$\|\delta x_{i+1}\| \le \|F_i''' x_i\| + \|\delta\bar{a}_i p_i\| + \|V_i p_i\| \le$$

$$\le (\|x_i\| + \|A^{-\frac{1}{2}}\|\|A^{-\frac{1}{2}} f_i\|)\, o(1) .$$

In view of remark 2.4.4 we have

(7) $$\|\hat{a}_i p_i\| \le \|A^{-\frac{1}{2}}\|\|\hat{a}_i A^{\frac{1}{2}} p_i\| \le \|A^{-\frac{1}{2}}\|\|A^{-\frac{1}{2}} f_i\|\,(1+o(1)) .$$

Substitution of (6) and (7) into (5) proves (3). □

We now arrive at the numerical analogue of (4.2.6) for the TRCGM.

THEOREM 2. *Let* x_i, p_i *be two arbitrary machine vectors and let* x_{i+1}, p_{i+1} *be computed from one step TRCGM. Furthermore, let according to previous definitions, for* $\ell = i, i+1$

(8) $$f_\ell := b - Ax_\ell ,$$

(9) $$\varphi_\ell := \|A\|\|x_\ell\| / \|f_\ell\| , \quad \psi_\ell := \|A^{\frac{1}{2}}\|\|x_\ell\| / \|A^{-\frac{1}{2}} f_\ell\| .$$

Then we have

(10) $$(f_{i+1}, p_{i+1}) = (1+\lambda_{i+1})\|f_{i+1}\|^2 ,$$

where

(11) $$|\lambda_{i+1}| \le \varepsilon\Big\{2(1+\kappa^{\frac{1}{2}}) + (1+C_1(1+\kappa^{\frac{1}{2}}))\varphi_{i+1} +$$

$$+ \kappa^{\frac{1}{2}}(1+2\kappa^{\frac{1}{2}}(2+C_2+C_1\kappa^{\frac{1}{2}}))\,\frac{\|A^{-\frac{1}{2}} f_i\|}{\|A^{-\frac{1}{2}} f_{i+1}\|}\Big\}(1+o(1))$$

under the restriction

$$\varepsilon\{\kappa^{\frac{1}{2}}(1+C_2+C_1\kappa^{\frac{1}{2}}) + (1+C_1\kappa^{\frac{1}{2}})(\psi_i+\psi_{i+1})\} \to 0 \ . \tag{12}$$

PROOF. All O-symbols are assumed to hold under the restriction (12). Expressing (2.4.38) in terms of $\hat{r}_i$, $\hat{r}_{i+1}$ we obtain

$$\hat{r}_{i+1} = \hat{r}_i - \hat{a}_i A\hat{p}_i - A\delta x_{i+1}$$

and consequently, from the definition of $\hat{a}_i$ it follows that

$$(\hat{r}_{i+1},\hat{p}_i) = (\hat{r}_i,\hat{p}_i) - \hat{a}_i(\hat{p}_i,A\hat{p}_i) - (A\delta x_{i+1},\hat{p}_i) = \tag{13}$$

$$= -(A\delta x_{i+1},\hat{p}_i) \ .$$

Using (4.2.18) and (2.4.25) this yields

$$(\hat{r}_{i+1},\hat{p}_i) = (\hat{r}_{i+1},r_{i+1}) + b_i(\hat{r}_{i+1},\hat{p}_i) + (\hat{r}_{i+1},\delta p'_{i+1}) = \tag{14}$$

$$= (\hat{r}_{i+1},\hat{r}_{i+1}) + (\hat{r}_{i+1},\delta r_{i+1}) + (\hat{r}_{i+1},\delta p'_{i+1}) - b_i(A\delta x_{i+1},\hat{p}_i) =$$

$$= (1+\lambda_{i+1})\|\hat{r}_{i+1}\|^2 \ ,$$

where

$$\lambda_{i+1} := \{(\hat{r}_{i+1},\delta r_{i+1}) + (\hat{r}_{i+1},\delta p'_{i+1}) - b_i(A\delta x_{i+1},\hat{p}_i)\} / \|\hat{r}_{i+1}\|^2 \ . \tag{15}$$

It remains to be estimated each of the terms in (15).
From (2.4.12) it follows that

$$|(\hat{r}_{i+1},\delta r_{i+1})| / \|\hat{r}_{i+1}\|^2 \le \|\delta r_{i+1}\| / \|\hat{r}_{i+1}\| \le \tag{16}$$

$$\le \varepsilon(1+C_1\varphi_{i+1}(1+O(1))) \ .$$

From (4.2.19), (4.2.20), (4.2.28) and (2.4.5) it follows that

$$|(\hat{r}_{i+1},\delta p'_{i+1})| / \|\hat{r}_{i+1}\|^2 \le \|\delta p'_{i+1}\| / \|\hat{r}_{i+1}\| \le \tag{17}$$

$$\le (\varepsilon\|r_{i+1}\| + 2\varepsilon\|A^{-\frac{1}{2}}\|\|b_i A^{\frac{1}{2}}\hat{p}_i\|(1+O(1))) / \|\hat{r}_{i+1}\| \le$$

$$\le (\varepsilon\|r_{i+1}\| + 2\varepsilon\|A^{-\frac{1}{2}}\|\|A^{\frac{1}{2}} r_{i+1}\|(1+O(1)) / \|\hat{r}_{i+1}\| \le$$

$$\le \varepsilon(1+2\kappa^{\frac{1}{2}})(1+O(1)) \ .$$

From (2.4.39), (2.4.34), (4.2.28), (3), (2.4.28), (2.4.11) and (2.4.41) it follows that

$$(18)\qquad |b_i(A\,\delta x_{i+1}, p_i)| \,/\, \|\hat{f}_{i+1}\|^2 \le \|A^{\frac{1}{2}}\delta x_{i+1}\|\,\|b_i A^{\frac{1}{2}} p_i\| \,/\, \|\hat{f}_{i+1}\|^2 \le$$

$$\le (\|A^{\frac{1}{2}}\|\|F_i'''x_i\| + \|\delta a_i' A^{\frac{1}{2}} p_i\| + \|A^{-\frac{1}{2}}\delta r_i\| +$$

$$+ \|A^{\frac{1}{2}}\|\|V_i p_i\|)\,\|A^{\frac{1}{2}} r_{i+1}\| \,/\, \|\hat{f}_{i+1}\|^2 (1 + o(1)) \le$$

$$\le \{\varepsilon\|A^{\frac{1}{2}}\|(\|x_{i+1}\| + \|A^{-\frac{1}{2}}\|\,\|A^{-\frac{1}{2}}\hat{f}_i\|) + \varepsilon(1 + 2C_2\kappa^{\frac{1}{2}} + C_1\kappa)\|A^{-\frac{1}{2}}\hat{f}_i\| +$$

$$+ \|A^{-\frac{1}{2}}\|(\varepsilon\|\hat{f}_i\| + \varepsilon C_1\|A\|(\|x_{i+1}\| + \|A^{-\frac{1}{2}}\|\|A^{-\frac{1}{2}}\hat{f}_i\|)) +$$

$$+ 2\varepsilon\kappa^{\frac{1}{2}}\|A^{-\frac{1}{2}}\hat{f}_i\|\}\|A^{\frac{1}{2}}\| \,/\, \|\hat{f}_{i+1}\|(1 + o(1)) =$$

$$= \{\varepsilon(1 + C_1\kappa^{\frac{1}{2}})\|A\|\|x_{i+1}\| +$$

$$+ \varepsilon(1 + 2\kappa^{\frac{1}{2}}(2 + C_2 + C_1\kappa^{\frac{1}{2}}))\|A^{\frac{1}{2}}\|\,\|A^{-\frac{1}{2}}\hat{f}_i\|\} \,/\, \|\hat{f}_{i+1}\|(1 + o(1)) \le$$

$$\le \varepsilon(1 + C_1\kappa^{\frac{1}{2}})\varphi_{i+1} +$$

$$+ \varepsilon\kappa^{\frac{1}{2}}(1 + 2\kappa^{\frac{1}{2}}(2 + C_2 + C_1\kappa^{\frac{1}{2}}))\|A^{-\frac{1}{2}}\hat{f}_i\| \,/\, \|A^{-\frac{1}{2}}\hat{f}_{i+1}\|(1 + o(1)) \,.$$

Inequality (11) now follows from (15), (16), (17) and (18). □

A more explicit version of this result is stated without proof. We have

PROPOSITION 3. *Let* x_i, p_i *be two arbitrary machine vectors and let* x_{i+1}, p_{i+1} *be computed from one step TRCGM. If*

$$(19)\qquad \varepsilon\kappa^{\frac{1}{2}}(3 + C_2 + C_1\kappa^{\frac{1}{2}}) \le \frac{1}{40}\,,$$

and if for $\ell = i, i+1$:

$$(20)\qquad \varepsilon(1 + C_1\kappa^{\frac{1}{2}})\psi_\ell \le \frac{1}{16}\,,$$

then, in equality (11), certainly

$$(21)\qquad |\lambda_{i+1}| \le \varepsilon\Big\{2(1+\kappa^{\frac12}) + (1+C_1(1+\kappa^{\frac12}))\varphi_{i+1} + \\ + \kappa^{\frac12}(1+2\kappa^{\frac12}(2+C_2+C_1\kappa^{\frac12}))\,\frac{\|A^{-\frac12}f_i\|}{\|A^{-\frac12}f_{i+1}\|}\Big\}(1+\tfrac{9}{10})\ . \ \square$$

Now that we have derived the numerical analogue of (4.2.6) we can estimate the parameters α_i, β_i and γ_i in the presence of round-off. From (11) and (12), replacing $i+1$ by i, it follows that under the restriction

$$(22)\qquad \varepsilon\Big\{\kappa^{\frac12}(1+C_2+C_1\kappa^{\frac12}) + (1+C_1\kappa^{\frac12})\psi_{i-1} + (1+C_1\kappa^{\frac12})\varphi_i + \\ + \kappa^{\frac12}(1+2\kappa^{\frac12}(2+C_2+C_1\kappa^{\frac12}))\,\frac{\|A^{-\frac12}f_{i-1}\|}{\|A^{-\frac12}f_i\|}\Big\} \to 0\ ,$$

certainly $|\lambda_i| = O(1)$. Using (2.4.5), (2.4.6) and (4.2.27) it follows that, under the restriction (22),

$$(23)\qquad \alpha_i := \frac{\|f_i\|\,\|p_i\|}{|(f_i,p_i)|} \le \frac{\|f_i\|\|A^{-\frac12}\|\|A^{\frac12}p_i\|}{\|f_i\|^2(1-|\lambda_i|)} \le \frac{\|A^{\frac12}\|\,\|A^{\frac12}r_i\|}{\|f_i\|}(1+O(1)) \le$$

$$\le \kappa^{\frac12}\frac{\|r_i\|}{\|f_i\|}(1+O(1)) \le \kappa^{\frac12}(1+O(1))\ ,$$

$$(24)\qquad \beta_i := \frac{\|f_i\|\|A^{\frac12}p_i\|}{\|A^{\frac12}\|\,|(f_i,p_i)|} \le \frac{\|f_i\|\|A^{\frac12}r_i\|}{\|A^{\frac12}\|\|f_i\|^2(1-|\lambda_i|)}(1+O(1)) \le$$

$$\le \frac{\|r_i\|}{\|f_i\|}(1+O(1)) \le 1+O(1)\ ,$$

$$(25)\qquad \gamma_i^{-1} := \frac{\|A^{-\frac12}f_i\|\|A^{\frac12}p_i\|}{|(f_i,p_i)|} \le \frac{\|A^{-\frac12}r_i\|\|A^{\frac12}r_i\|}{\|f_i\|^2(1-|\lambda_i|)}(1+O(1)) \le$$

$$\le \frac{\|A^{-\frac12}r_i\|\|A^{\frac12}r_i\|}{\|r_i\|^2}(1+O(1)) \le \frac{\kappa+1}{2\kappa^{\frac12}}(1+O(1))\ .$$

This completes the set of basic relations that are needed to prove the main theorem of this section. We only formulate an explicit version (cf. proposition 4.2.5).

PROPOSITION 4. *Let* x_{i-1}, p_{i-1} *be two arbitrary machine vectors and let*

$$L := \left\{1 - \frac{1}{5}\,\frac{\kappa}{(\kappa+1)^2}\right\}^{\frac{1}{2}} .$$

Consider two steps of the TRCGM. If

$$\varepsilon\kappa(3 + 3C_2 + C_1\kappa^{\frac{1}{2}}) \le \frac{1}{40} , \tag{26}$$

$$\varepsilon(3 + 2C_1\kappa^{\frac{1}{2}})\varphi_i \le \frac{1}{16} , \tag{27}$$

and if

$$\varepsilon(1 + C_1\kappa^{\frac{1}{2}})\psi_{i-1} \le \frac{1}{16} , \tag{28}$$

then at least one of the following two inequalities

$$\frac{\|A^{\frac{1}{2}}(\hat{x} - x_{i+1})\|^2}{\|A^{\frac{1}{2}}(\hat{x} - x_i)\|^2} \le L^2 , \tag{29}$$

$$\frac{\|A^{\frac{1}{2}}(\hat{x} - x_{i+1})\|^2}{\|A^{\frac{1}{2}}(\hat{x} - x_{i-1})\|^2} \le L^4 \tag{30}$$

holds.

PROOF. First we derive some explicit bounds concerning r_i, $\hat{r}_i$. From (2.4.9), (2.4.12) and the assumptions (26) and (27) it follows that

$$\|r_i\| / \|\hat{r}_i\| \le 1 + \|\delta r_i\| / \|\hat{r}_i\| \le 1 + \varepsilon(1 + C_1(1+\varepsilon)\kappa^{\frac{1}{2}}\psi_i) \le \tag{31}$$

$$\le 1 + \frac{1}{120} + \left(\frac{1}{32}\right)\left(\frac{121}{120}\right) < 1 + \frac{1}{10} .$$

From (2.4.9), (2.4.13) and the assumptions (26) and (27) it follows that

$$\|A^{-\frac{1}{2}}\hat{r}_i\| / \|A^{-\frac{1}{2}}r_i\| \le 1 + (\|A^{-\frac{1}{2}}\delta r_i\| / \|A^{-\frac{1}{2}}\hat{r}_i\|)(\|A^{-\frac{1}{2}}\hat{r}_i\| / \|A^{-\frac{1}{2}}r_i\|) \le \tag{32}$$

$$\le 1 + \varepsilon\kappa^{\frac{1}{2}}(1 + C_1(1+\varepsilon))\psi_i\|A^{-\frac{1}{2}}\hat{r}_i\| / \|A^{-\frac{1}{2}}r_i\| \le$$

$$\le 1 + \frac{1}{10}\,\|A^{-\frac{1}{2}}\hat{r}_i\| / \|A^{-\frac{1}{2}}r_i\| .$$

Hence

$$\|A^{-\frac{1}{2}} f_i\| / \|A^{-\frac{1}{2}} r_i\| \leq 1 + \frac{1}{9} . \tag{33}$$

The remaining part of the proof is very similar to the proof of proposition 4.2.5.

Note that $L^2 \geq 1/20$. If $\|A^{-\frac{1}{2}} f_{i+1}\|^2 \leq L^4 \|A^{-\frac{1}{2}} f_{i-1}\|^2$, then we are ready. Let us now assume that $\|A^{-\frac{1}{2}} f_{i+1}\|^2 > L^4 \|A^{-\frac{1}{2}} f_{i-1}\|^2$. From the assumptions (26) and (28) it follows that $\|A^{-\frac{1}{2}} f_{i+1}\| \leq (13/10)\|A^{-\frac{1}{2}} f_i\|$ and (cf. remark 2.4.4) consequently

$$\frac{\|A^{-\frac{1}{2}} f_{i-1}\|}{\|A^{-\frac{1}{2}} f_i\|} = \frac{\|A^{-\frac{1}{2}} f_{i-1}\|}{\|A^{-\frac{1}{2}} f_{i+1}\|} \frac{\|A^{-\frac{1}{2}} f_{i+1}\|}{\|A^{-\frac{1}{2}} f_i\|} \leq \frac{13}{10L^2} < \frac{7}{5} . \tag{34}$$

Hence, from proposition 3 and the assumptions (26), (27) and (28) we obtain

$$\begin{aligned} |\lambda_i| \leq (\tfrac{19}{10})\{\varepsilon(2+\kappa^{\frac{1}{2}}) + \varepsilon(1+C_1(1+\kappa^{\frac{1}{2}}))\varphi_i + \\ + (\tfrac{7}{5})\varepsilon\kappa^{\frac{1}{2}}(1+2\kappa^{\frac{1}{2}}(2+C_2+C_1\kappa^{\frac{1}{2}}))\} \leq \\ \leq 7\varepsilon\kappa(3+3C_2+C_1\kappa^{\frac{1}{2}}) + 2\varepsilon(3+2C_1\kappa^{\frac{1}{2}})\varphi_i \leq \frac{12}{40} . \end{aligned} \tag{35}$$

In view of (4.2.30) one has

$$\|A^{\frac{1}{2}} p_i\| \leq (\tfrac{11}{10})\|A^{\frac{1}{2}} r_i\| . \tag{36}$$

The inequalities (31), (32), (34) and (35) enable us to replace the estimates (23), (24) and (25) by explicit bounds. We thus obtain

$$\alpha_i \leq \kappa^{\frac{1}{2}}\left(1 - \frac{12}{40}\right)^{-1}\left(\frac{11}{10}\right)^2 \leq 2\kappa^{\frac{1}{2}} , \tag{37}$$

$$\beta_i \leq 1\left(1 - \frac{12}{40}\right)^{-1}\left(\frac{11}{10}\right)^2 \leq 2 , \tag{38}$$

$$\gamma_i^{-1} \leq (\kappa+1) / (2\kappa^{\frac{1}{2}})\left(1 - \frac{12}{40}\right)^{-1}\left(\frac{10}{9}\right)\left(\frac{11}{10}\right)^2 \leq (\kappa+1)/\kappa^{\frac{1}{2}} . \tag{39}$$

Finally, from (26), (27), (37) and (38) it follows that the assumptions of proposition 2.4.8 are satisfied and consequently

$$(40)\qquad \frac{\|A^{-\frac{1}{2}} f_{i+1}\|^2}{\|A^{-\frac{1}{2}} f_i\|^2} \le 1 - \frac{1}{5}\gamma_i^2 \le 1 - \frac{\kappa}{5(\kappa+1)^2} = L^2 \ . \qquad \square$$

We now assume that (26) holds. It easily follows from proposition 4 that if (27) holds for all $0 \le i \le k$ and (28) holds for all $0 \le \ell \le k$, then for every $0 \le i \le k$

$$(41)\qquad \frac{\|A^{\frac{1}{2}}(\hat{x} - x_i)\|}{\|A^{\frac{1}{2}}(\hat{x} - x_{i-2})\|} \le L^2 \quad \text{or} \quad \frac{\|A^{\frac{1}{2}}(\hat{x} - x_{i+1})\|}{\|A^{\frac{1}{2}}(\hat{x} - x_{i-1})\|} \le L^2 \ .$$

Hence, as long as (27) and (28) are satisfied (cf. section 4.2) we have bi-step-wise linear convergence of the natural error $\|A^{\frac{1}{2}}(\hat{x} - x_i)\|$ and consequently also linear convergence on the average with a convergence ratio no greater than L. From this we can draw three conclusions similar to those we derived for the TRGSM from proposition 2.4.10, but now expressed in terms of bi-step-wise linear convergence. They are combined into the following proposition.

PROPOSITION 5. *If* $\{x_i\}$ *is generated by the TRCGM with arbitrary initial machine vector* x_0 *and if*

$$(42)\qquad \varepsilon\kappa(3 + 3C_2 + C_1\kappa^{\frac{1}{2}}) \le \frac{1}{40} \ ,$$

then the natural error $\|A^{\frac{1}{2}}(\hat{x} - x_i)\|$ *converges bi-step-wise linearly with a convergence ratio no greater than* $(1 - \kappa / (5(\kappa+1)^2))^{\frac{1}{2}}$, *at least until the iteration step where one of the following two inequalities is true:*

$$(43)\qquad \|A(\hat{x} - x_i)\| \le 16\varepsilon(3 + 2C_1\kappa^{\frac{1}{2}})\|A\|\|x_i\| \ ,$$

$$(44)\qquad \|A^{\frac{1}{2}}(\hat{x} - x_{i-1})\| \le 16\varepsilon(1 + C_1\kappa^{\frac{1}{2}})\|A^{\frac{1}{2}}\|\|x_i\| \ . \qquad \square$$

If inequality (43) is essentially sharp and if C_1 is of order unity (or order $n^{3/2}$ as in straightforward implementation), then the residual $\|b - Ax_i\|$ is a factor $\kappa^{\frac{1}{2}}$ too large to have good-behavior; if C_1 is of the order $\varepsilon^{\frac{1}{2}}$ (e.g. by using double length precision), then the residual is small enough to guarantee good-behavior.

If inequality (44) is satisfied, then the natural error $\|A^{\frac{1}{2}}(\hat{x} - x_i)\|$ is at most of the order of the inherent natural error and hence we have A-numerical stability and consequently also numerical stability. Combining the two inequalities, and using (1.4.3) and (1.4.4), we may conclude that if (42) is satisfied, then TRCGM generates at least one approximation x_i for which

$$(45) \qquad \|A^{\alpha}(\hat{x} - x_i)\| \leq 16\varepsilon(3 + 2C_1\kappa^{\frac{1}{2}})\kappa^{1-\alpha}\|A^{\alpha}\|\|x_i\| \qquad (\alpha = 0, \tfrac{1}{2}, 1) \ .$$

Hence we cannot conclude that in general the TRCGM is well-behaved or (A)-numerically stable. If, however, C_1 is of the order $\varepsilon^{\frac{1}{2}}$, then it follows that TRCGM is well-behaved.

REMARK 6. In view of proposition 2.4.8 it is not surprising that (43) contains a term of order $\varepsilon C_1\kappa^{\frac{1}{2}}\|A\|\|x_i\|$ (which precludes the proof of good-behavior), since the underlying restriction (2.4.98) contains a term $\varepsilon C_1\alpha_i\varphi_i$ and we only have the a priori bound $\alpha_i \leq \kappa^{\frac{1}{2}}(1 + o(1))$ (cf. remark 4.2.9). Note, however, that it follows from proposition 2.4.8 that if (42) is satisfied and if α_i and β_i are of order unity at a step where the monotonicity of the natural error breaks down, then the residual $\|b - Ax_i\|$ is of the order $\varepsilon\|A\|\|x_i\|$ and hence we have good-behavior. □

Note that all results derived in this section also hold for the TRISCGM as defined in section 4.1.

4.4. Numerical experiments

In this section we report on some of the numerical experiments that have been carried out with the CGM in order to verify our analytical results deduced in sections 2 and 3. Our tests are based on the three different ways of implementation, viz. assembled implementation (AI), profuct form implementation (PFI) and artificial floating point implementation (AFI), as described in section 1.6. As far as the distribution of eigenvalues, the choice of the orthogonal matrix U, the assemblage of U (both in case of AI or PFI) and the eigenvector

components of b and x_0 are concerned, we refer to the introductory part of section 3.4, which deals with similar tests for the GM. If we display the values of $F_{i,\alpha} := fl(\|A^{\alpha}(\hat{x} - x_i)\|)$, $(\alpha = 0, \frac{1}{2}, 1)$, when reporting on numerical results, then these values are computed according to the formulas (3.4.2) (for AI and PFI) and (3.4.3) (for AFI). For a discussion on significant figures we refer to remark 3.4.2 and remark 3.4.3, respectively.

4.4.1. The true residual conjugate gradient method

For the numerical performance of the TRCGM we have deduced two basic analytical results. Firstly, the (explicit) result stated in proposition 4.3.5, from which it follows that, if $\varepsilon\kappa(3 + 3C_2 + C_1\kappa^{\frac{1}{2}}) \leq 1/40$, then the TRCGM generates at least one approximation x_i for which

$$(1) \qquad \|A^{\alpha}(\hat{x} - x_i)\| \leq 16\varepsilon(3 + 2C_1\kappa^{\frac{1}{2}})\kappa^{1-\alpha}\|A^{\alpha}\|\|x_i\| , \quad (\alpha = 0, \tfrac{1}{2}, 1);$$

this will be called the *reachable level*. Secondly, the (explicit) result stated in proposition 2.4.8 and holding for every TRDM, from which it follows that, if $\|A^{\frac{1}{2}}(\hat{x} - x_{k+1})\| \geq \|A^{\frac{1}{2}}(\hat{x} - x_k)\|$ for some k and if $\varepsilon\{\kappa^{\frac{1}{2}}(1 + C_2 + C_1\kappa^{\frac{1}{2}}) + \alpha_k(1 + C_2)\} \leq 1/40$, then

$$(2) \qquad \|A^{\alpha}(\hat{x} - x_k)\| \leq 4\varepsilon(2\beta_k(1 + \beta_k) + C_1\alpha_k)\kappa^{1-\alpha}\|A^{\alpha}\|\|x_k\| \quad (\alpha = 0, \tfrac{1}{2}, 1) .$$

Here

$$(3) \qquad \alpha_k := \|\hat{r}_k\|\|p_k\| / |(\hat{r}_k, p_k)| ,$$

$$(4) \qquad \beta_k := \|\hat{r}_k\| \, \|A^{\frac{1}{2}}p_k\| / (\|A^{\frac{1}{2}}\| \, |(\hat{r}_k, p_k)|) ,$$

$$(5) \qquad \hat{r}_k := b - Ax_k .$$

We recall that algebraically $\alpha_k \leq \kappa^{\frac{1}{2}}$ and $\beta_k \leq 1$.
Our tests are mainly designed for the verification of these two results. In all experiments the TRCGM iterations are terminated as soon as $F_{k+1,\frac{1}{2}} \geq F_{k,\frac{1}{2}}$ (see remark 3.4.1.1).
We also performed tests similar to those performed for the GM in order to investigate the influence of the value of m (for AI and PFI), the influence of the eigenvector components of $\hat{x}$ and $\hat{x} - x_0$ and the influence of the basic arithmetical operations. The conclusions that can

be drawn from these tests do not essentially differ from the conclusions for the GM and therefore the results are deleted.

The influence of κ

We performed several tests based on AFI with fixed dimension n, fixed eigenvector components s and e for $\hat{x}$ and $\hat{x} - x_0$, and with logarithmical eigenvalue distribution, only varying the condition number. A set of representative results is given in table 1, where n = 20, $s_j / s_{j+1} = e_j / e_{j+1} = 10^3$, $\|s\| = 1$, $\|e\| = \kappa_{10}-1$ (hence $\|A^{\frac{1}{2}}(\hat{x} - x_0)\| = \kappa^{\frac{1}{2}}{}_{10}-1$) and the artificial relative precision is $\delta = 10^{-6}$. We performed five different tests for each condition number; different in the sense that we invoked different artificial round-off errors (cf. section 1.6). In all cases at iteration step k, defined as being the first step for which $F_{k+1,\frac{1}{2}} \geq F_{k,\frac{1}{2}}$, there holds $\|x_k\| \sim 1$ $(= \|\hat{x}\|)$. The smallest and largest observed values of k of each set of five tests with the same condition number are given in the table. In the columns headed $\|A^{\alpha}(\hat{x} - x_k)\|, (\alpha = 0, \frac{1}{2}, 1)$, we display the largest and smallest observed value of $F_{k,\alpha}$. The column denoted by α_{max} indicates the largest observed value of α_i (cf. (3)) throughout all iteration steps and all five tests together. The column denoted by α_k indicates the largest observed value of α_k.

κ	k		$\|\hat{x} - x_0\|$		$\|A^{\frac{1}{2}}(\hat{x} - x_0)\|$		$\|A(\hat{x} - x_0)\|$		α_{max}	α_k
10^2	24	26	$2.8_{10}-5$	$5.9_{10}-5$	$7.5_{10}-6$	$8.7_{10}-6$	$3.1_{10}-6$	$4.0_{10}-6$	2.3	1.6
$10^{2.5}$	42	58	$1.2_{10}-4$	$2.6_{10}-4$	$1.0_{10}-5$	$1.6_{10}-5$	$2.3_{10}-6$	$3.7_{10}-6$	3.0	2.0
10^3	77	97	$5.3_{10}-4$	$9.2_{10}-4$	$1.8_{10}-5$	$3.7_{10}-5$	$2.3_{10}-6$	$8.1_{10}-6$	7.6	2.0
$10^{3.5}$	123	218	$1.7_{10}-3$	$3.3_{10}-3$	$3.9_{10}-5$	$6.4_{10}-5$	$3.3_{10}-6$	$4.4_{10}-6$	9.5	3.1
10^4	231	337	$4.8_{10}-3$	$1.2_{10}-2$	$5.6_{10}-5$	$1.2_{10}-4$	$3.2_{10}-6$	$6.1_{10}-6$	48	2.9

TABLE 1. *The influence of* κ

We see that at step k certainly (1) is satisfied in all cases, even if we replace the factor $C_1\kappa^{\frac{1}{2}}$ by C_1. Consequently the non-well-behavior of the TRCGM, suggested by (1), is not confirmed by the results of table 1. We recall that (1) is in fact deduced from (2), using the a priori

bound $\alpha_i \leq \kappa^{\frac{1}{2}}(1+o(1))$ (cf. (4.3.23)). As we see from table 1, at iteration step k, where the monotonicity breaks down, α_k is of order 1 in all cases. Hence the estimate for α_i as used in (1) is unsharp by a factor $\kappa^{\frac{1}{2}}$ and therefore (1) itself is also unsharp by a factor $\kappa^{\frac{1}{2}}$. On the other hand, the data of the column headed α_{max} indicate that it is doubtful whether the a priori bound $\alpha_i \leq \kappa^{\frac{1}{2}}(1+o(1))$ (for all $0 \leq i \leq k-1$) can be sharpened as far as the exponent of κ is concerned. In all 25 tests, throughout all iteration steps, there holds $\beta_i \leq 1$, as was to be expected in view of (4.3.24), Therefore using the actual values of α_k and β_k, we conclude from (2) that for the tests of table 1 the TRCGM is well-behaved. This corresponds to the numerical results obtained. The results of table 1 also agree with the algebraic property (for i = k) stated in theorem 4.1.3, i.e.,

$$\|A^{\frac{1}{2}}(\hat{x}-x_i)\| \leq 2\left(\frac{\kappa^{\frac{1}{2}}-1}{\kappa^{\frac{1}{2}}+1}\right)^i \|A^{\frac{1}{2}}(\hat{x}-x_0)\| \quad , \tag{6}$$

and indicating the faster linear convergence on the average with an average convergence ratio no greater than $(\kappa^{\frac{1}{2}}-1)/(\kappa^{\frac{1}{2}}+1)$ $(\sim 1-2/\kappa^{\frac{1}{2}})$ for the TRCGM in comparison with the TRGM. As far as the step-wise convergence of the natural error is concerned, we proved (cf. proposition 4.3.4) that the convergence ratio is no greater than $(\kappa-1)/(\kappa+1)$ $(\sim 1-2/\kappa)$. This difference between average convergence ratio and (step-wise) convergence ratio is also revealed by our tests. For instance, in a particular test with $\kappa = 10^4$ the natural error decreased by a factor 0.5504 in the steps from 200 to 300. Since $(0.5504)^{1/100} = 0.9940$, there is at least one step for which $\|A^{\frac{1}{2}}(\hat{x}-x_{i+1})\| / \|A^{\frac{1}{2}}(\hat{x}-x_i)\| \geq 0.9940$, whereas $1-2/\kappa^{\frac{1}{2}} = 0.980$. It also follows that the average convergence ratio, based on these hundred steps, is greater than $1-2/\kappa^{\frac{1}{2}}$. One may ask whether it is possible to construct a test problem (with a large initial natural error $\|A^{\frac{1}{2}}(\hat{x}-x_0)\|$ and slow convergence in the first steps in order to accomplish the need of much more than n iterations before attaining the reachable level), which ultimately contradicts the algebra average convergence ratio. Note that the proof of (6) is based on algebraic orthogonality relations holding for all $0 \leq i \leq n$. Of course, these relations cannot hold anymore after n iterations and therefore it is doubtful whether (6) still holds for the numerical process in case

$i \geq n$. We performed a rather limited set of trials to justify this doubt, but in all tests (6) held.

REMARK 1. In fact, our propositions on the numerical speed of convergence for the TRCGM are expressed in terms of bi-step-wise linear convergence (cf. proposition 4.3.4) with a convergence ratio no greater than $(\kappa - 1) / (\kappa + 1)$. We were not able to prove step-wise linear convergence since the estimate for the parameter λ_i in the relation $(\tilde{r}_i, p_i) = (1 + \lambda_i)\|\tilde{r}_i\|$ contains a term with a factor $\|A^{-\frac{1}{2}} \tilde{r}_{i-1}\| / \|A^{-\frac{1}{2}} \tilde{r}_i\|$ (cf. theorem 4.3.2) and there is no a priori upper bound available for this factor. Consequently, no a priori upper bound for the parameters α_i, β_i and γ_i is available (the restriction under which the inequalities (23), (24) and (25) of section 4.3 hold contains the same factor). Therefore, if $\|A^{-\frac{1}{2}} \tilde{r}_{i-1}\| / \|A^{-\frac{1}{2}} \tilde{r}_i\|$ is extremely small, then it might be possible that in the next step the ratio $\|A^{-\frac{1}{2}} \tilde{r}_i\| / \|A^{-\frac{1}{2}} \tilde{r}_{i+1}\|$ is slightly less than one, even if the residual $\|\tilde{r}_i\|$ has not yet achieved the reachable level $\varepsilon(1 + C_1 \kappa^{\frac{1}{2}})\|A\|\|x_i\|$. However, from the results of the tests of table 1 we see that in all steps, apart from the very first steps, one has $\|A^{-\frac{1}{2}} \tilde{r}_{i-1}\| / \|A^{-\frac{1}{2}} \tilde{r}_i\| \leq 4$. Hence, in view of estimate (4.3.21) it follows that in all cases the parameter λ_i is fairly small as long as the residual is not of the order of the reachable level. Consequently, after the very first steps it is impossible to encounter the situation that the natural error increases in some step, whereas at the same step the residual is not of the order of the reachable level. Hence, using this bit of a posteriori information, we may state that for all our tests our analytical results yield step-wise linear convergence as long as the reachable level has not yet been attained. □

As we mentioned already at the end of section 4.3, all results deduced in that section also hold for the TRISCGM as defined in section 4.1 (with not necessarily $p_0 = r_0$). We carried out several tests with $\kappa = 10^4$ and the same parameter setting as in the tests of table 1. The only difference consisted of the (independent) choice of p_0. All tests confirmed the analytical results of section 4.3. For the five tests, with initial p_0 vector with components p_j satisfying $p_j / p_{j+1} = 10^3$ and $\|p\| = 1$, the observed values (ordered and to be interpreted in the same way as the lines of table 1) read

10^4|624|868|6.0_{10}-3|1.3_{10}-2|2.7_{10}-5|1.3_{10}-4|2.1_{10}-6|1.4_{10}-5|160|160|

The computed values of $\|A^{\alpha}(\hat{x} - x_k)\|$, $(\alpha = 0, \frac{1}{2})$, agree with the corresponding values for the TRCGM with $\kappa = 10^4$ (cf. last line of table 1). The values of $\|A(\hat{x} - x_k)\|$ seem to be slightly larger. The test for which (the value) $\|A(\hat{x} - x_k)\| = 1.4_{10}$-5 is attained, is also the test with $\alpha_{max} = \alpha_k = 160$. In view of (2) it is not surprising that the very test with a relatively large value of α_k also has a relatively large value of $\|A(\hat{x} - x_k)\|$. Note that the algebraic inequality $|\alpha_k| \leq \kappa^{\frac{1}{2}}$ is seriously disturbed at this step. As far as the speed of convergence is concerned we see that, compared with the TRCGM case, twice as many steps are needed to achieve the reachable level.

As we mentioned already in section 4.1, the algebraic upper bound $(\kappa^{\frac{1}{2}} - 1) / (\kappa^{\frac{1}{2}} + 1)$ for the average convergence ratio of the TRCGM not necessarily holds for the TRISCGM, since in the latter case, at the point x_{i+1}, the objective function $F(x) := \|A^{\frac{1}{2}}(\hat{x} - x)\|$ is not necessarily minimized on the affine set passing through x_0 and spanned by $p_0, \ldots, p_i$. Therefore the slower convergence of TRISCGM is not surprising.

In figure 1 we plotted the values of $\|A^{\frac{1}{2}}(\hat{x} - x_i)\|$ and $\|A(\hat{x} - x_i)\|$ $(i = 0,50,100,\ldots,k)$, both for one test of the TRGM and one test of the TRISCGM; the values of k are 301 and 709, respectively.

We see that the faster convergence of $\|A^{\frac{1}{2}}(\hat{x} - x_i)\|$ for the TRCGM test is restricted to the first hundred steps; after these steps the speed of convergence hardly differs between the two tests and the convergence ratio varies between $1 - 6/\kappa^{\frac{1}{2}}$ and $1 - 2/\kappa$ for both tests (this observation is not only based on the plotted steps but on all steps). The average convergence ratio is 0.9869 ($\sim 1 - 1.3/\kappa^{\frac{1}{2}}$) for the TRISCGM test (based on all 709 steps) and 0.9817 ($\sim 1 - 1.8/\kappa^{\frac{1}{2}}$) for the TRCGM test (based on the last 200 steps). Hence, for both tests the algebraic upper bound $(\kappa^{\frac{1}{2}} - 1) / (\kappa^{\frac{1}{2}} + 1)$ ($\sim 1 - 2/\kappa^{\frac{1}{2}}$) for the average convergence ratio of the TRCGM does not quite hold for the numerical process during the last hundred of steps. We also see that initially the residual converges much faster for the TRCGM test. Although the figure suggests the monotonic decrement of the residual after the first 50 steps, this is not true. Both tests have the property that, incidentally, there are steps at which the residual slightly increased. The error $\|\hat{x} - x_i\|$ is not plotted. Also, initially, the error decreases

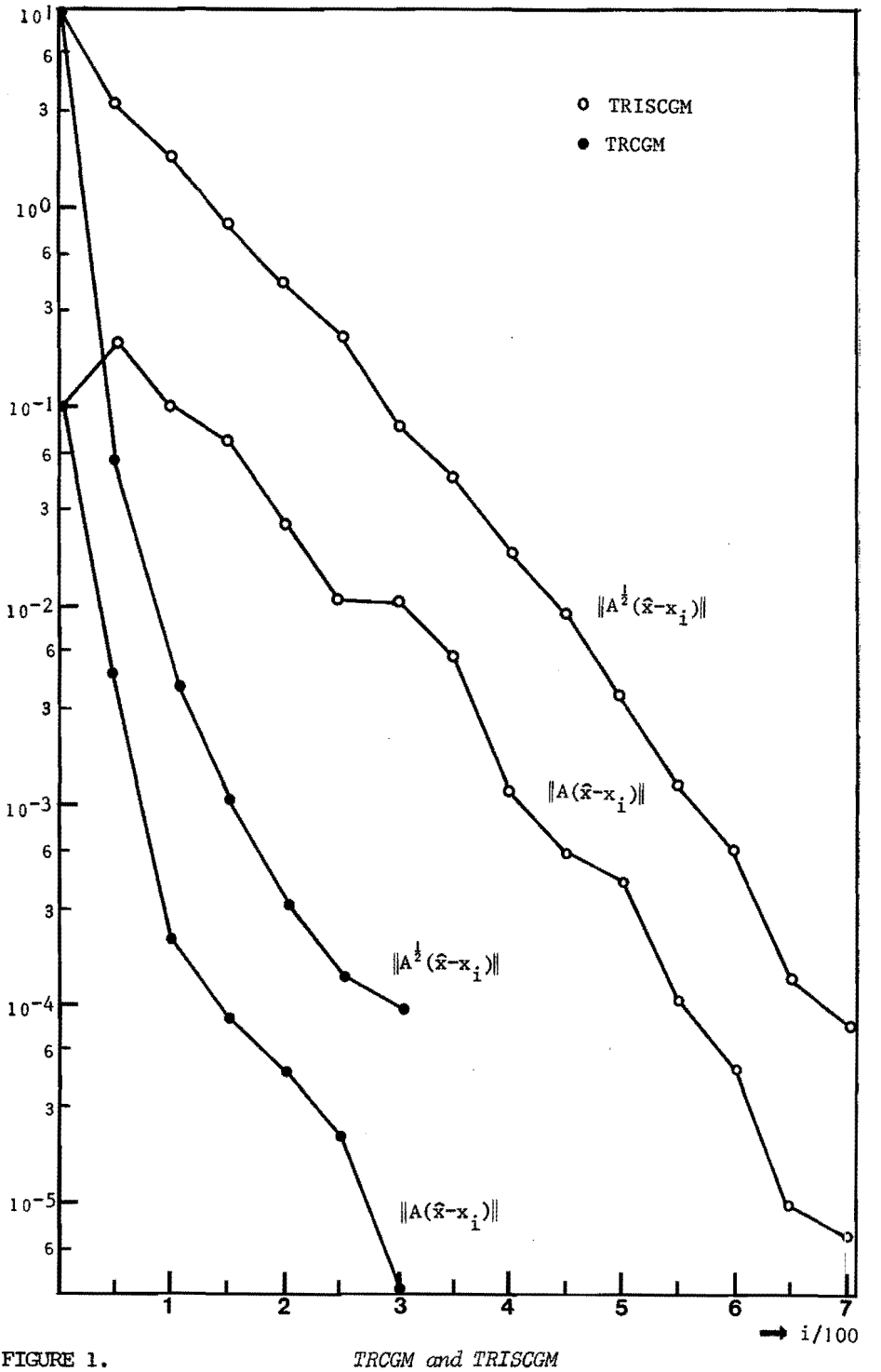


FIGURE 1. *TRCGM and TRISCGM*

faster for the TRCGM test and decreases monotonically at all steps. This last fact agrees with the algebraic property as stated in corollary 4.1.5. The TRISCGM test has the property that, incidentally, there are steps at which the error slightly increased.

4.4.2. The recursive residual conjugate gradient method

As far as the analytical results for the numerical performance of the RRCGM are concerned, the most striking result is the bi-step-wise linear convergence to zero of the natural error $\|A^{-\frac{1}{2}} r_i\|$ (cf. proposition 4.2.5). We performed several tests with the RRCGM, based on PFI ($n = 5$) and AFI ($\delta = 10^{-6}$), varying the dimension n ($20 \leq n \leq 50$), the eigenvector components s and e ($s_j / s_{j+1}, e_j / e_{j+1} = 10^3, 1, 10^{-3}$, $\|s\| = 1$, $10^{-1} \leq \|e\| \leq 10^{-6}$) and the condition number κ ($10^2 \leq \kappa \leq 10^4$). In order to avoid underflow the iterations were stopped as soon as $\|r_i\| \leq 10^{-22}$. In all cases this level has been reached. At all steps the natural error decreased at least by a factor $1 - 2/\kappa$. The reason that the situation, representative for bi-step-wise linear convergence, where $\|A^{-\frac{1}{2}} r_{i+1}\| / \|A^{-\frac{1}{2}} r_i\| > 1 - 2/\kappa$ and $\|A^{-\frac{1}{2}} r_{i+1}\| / \|A^{-\frac{1}{2}} r_{i-1}\| \leq (1 - 2/\kappa)^2$, does not occur is due to the fact that the factor $\|A^{-\frac{1}{2}} r_{i-1}\| / \|A^{-\frac{1}{2}} r_i\|$ never is extremely small in our test (see remark 4.4.1.1 for a more detailed explanation in a similar situation).
Since we do not have available an analytical result concerning the attainable accuracy of $\|A^{\alpha}(\hat{x} - x_i)\|$ ($\alpha = 0, \frac{1}{2}, 1$), for the RRCGM only limited attention to this aspect is paid and we do not present numerical results here.

CHAPTER 5

VARIANTS OF THE CONJUGATE GRADIENT METHOD

5.1. *Introduction*

In this chapter we discuss three other variants of the CGM. The only difference between these methods is the way in which the parameters a_i and b_i are computed.

We recall that in the CGM as defined in section 4.1 these parameters are determined by the relations

$$(1) \qquad a_i = (r_i, p_i) / (p_i, Ap_i) \ , \quad b_i = -(r_{i+1}, Ap_i) / (p_i, Ap_i) \ .$$

The formula for a_i follows directly from the fact that we want to minimize $\|A^{-\frac{1}{2}}(r_i - a\,Ap_i)\|^2$ (regarded as a function of a). Hence, evidently $(r_{i+1}, p_i) = (r_i - a_i\,Ap_i, p_i) = 0$. The formula for b_i follows immediately from the fact that we want $(r_{i+1} + bp_i, Ap_i) = 0$. From this point of view formulas (1) appear to be natural and therefore they will be referred to as *natural formulas* for a_i and b_i.

From (iv) and (v) of theorem 4.1.2 it follows that, algebraically, one has

$$(2) \qquad a_i = (r_i, r_i) / (p_i, Ap_i) \ , \quad b_i = (r_{i+1}, r_{i+1}) / (r_i, r_i) \ .$$

So, for both parameters we have an alternative formula.

The formulas (2) will be referred to as *unnatural formulas* for a_i and b_i.

Determining a_i from either the natural formula or the unnatural formula and determining b_i from either the natural or unnatural formula, we obtain four, algebraically equivalent, versions of the conjugate gradient method.

However, in the presence of round-off, these versions not necessarily have the same numerical properties.

In section 4.1 we introduced the so-called independent start conjugate gradient method (ISCGM), which is exactly the same as the CGM apart from the uncoupled choice of x_0 and p_0. Every step from x_i, p_i to x_{i+1}, p_{i+1} of the CGM can be considered as the first step of the ISCGM with initial vectors x_i, p_i. Algebraically, for every choice of x_0 and $p_0 \neq 0$, the natural error converges step-wise linearly to zero with a convergence ratio no greater than $(\kappa - 1) / (\kappa + 1)$ for the ISCGM (cf. section 4.1). From a numerical point of view this is interesting, since it implies that if in the CGM the occurrence of round-off is excluded permanently after some iteration step, then in the consecutive steps the natural error converges step-wise linearly to zero. This property holds for all DM's defined by the algorithm in section 2.2. No matter whether b_i is computed from (1) or from (2), if we compute a_i from (1), then the method is a DM. Therefore we may expect that the afore mentioned property also holds if we use formula (1) for a_i and (2) for b_i. One may ask whether the same property holds for the other two variants of the CGM, based on formula (2) for a_i.

We consider the independent start version of all four alternatives

Independent Start Conjugate Gradient Methods (ISCGM)

Choose an initial point x_0;

$r_0 := b - Ax_0$; $i := 0$;

Choose an initial direction vector p_0;

while $r_i \neq 0 \wedge p_i \neq 0$ do

begin

(3) $a_i := (r_i,p_i) / (p_i,Ap_i)$; (natural formula)

(4) or $a_i := (r_i,r_i) / (p_i,Ap_i)$; (unnatural formula)

(5) $x_{i+1} := x_i + a_i p_i$;

(6) $r_{i+1} := b - Ax_{i+1}$; (true residual)

(7) or $r_{i+1} := r_i - a_i Ap_i$; (recursive residual)

(8) $b_i := - (r_{i+1},Ap_i) / (p_i,Ap_i)$; (natural formula)

(9) or $b_i := (r_{i+1},r_{i+1}) / (r_i,r_i)$; (unnatural formula)

(10) $p_{i+1} := r_{i+1} + b_i p_i$;

```
        i := i + 1
end.
```

We use either (3) at all steps or (4) at all steps. The same applies to (6),(7) and to (8),(9), respectively; we disregard the mixed cases, where different alternative formulas are used at different steps. Note that from a numerical point of view we have 8 versions.
If the independent start conjugate gradient method is carried out using the natural formula for a_i, the true residual formula for r_{i+1} and the unnatural formula for b_i (in all steps!), then this method will be denoted by TRISCGNUM. The algorithms TRISCGNUM, RRISGUUM etc. are defined in a similar way. From now on ISCGM not only stands for the independent start version of our basic CGM of chapter IV, but for any of the four versions defined above.

REMARK 1. All four versions of the CGM are introduced in the paper of Hestenes and Stiefel [52]. In fact, they proposed the CGUUM as the basic one and mentioned the natural formulas as other possible choices. Reid [71] also considers all four versions and recommends the unnatural formulas and the recursive residuals on computational grounds, since he observes no significant difference as far as numerical convergence behavior is concerned.

In the next section we derive some algebraic properties of the ISCGM's. In the third section we carry out a one-round-off error analysis for the ISCGUNM to illustrate the numerical analogues of the algebraic properties and their implications.

5.2. *Algebraic properties of the independent start conjugate gradient methods (ISCGM)*

Obviously, we do not need to distinguish between the use of true and of recursive residuals as far as algebraic properties are concerned. Crowder and Wolfe [72] and Powell [76] considered the ISCGNNM, and some of their results will be mentioned later. As far as the other variants are concerned we believe our convergence results to be new. We start off with the following fundamental properties.

THEOREM 1. *Let* $\{r_0,\ldots,r_{\ell+1}\}$, $\{p_0,\ldots,p_{\ell+1}\}$ *($\ell \geq 1$) be computed by any version of the ISCGM's with arbitrary initial vectors* $r_0 \neq 0$, $p_0 \neq 0$. *Then we have for* $1 \leq i \leq \ell$

(i) $\quad a_i = (r_i,r_i) / (p_i,Ap_i)$,

(ii) $\quad b_i = (r_{i+1},r_{i+1}) / (r_i,r_i)$,

(iii) $\quad (r_{i+1},p_i) = b_{i-1}(r_i,p_{i-1})$,

(iv) $\quad (p_{i+1},Ap_i) = b_{i-1}(p_i,Ap_{i-1})$,

(v) $\quad (r_{i+1},p_{i+1})(r_i,r_i) = (r_i,p_i)(r_{i+1},r_{i+1})$.

PROOF. In the proof we frequently use the relations $r_{i+1} = r_i - a_i Ap_i$ and $p_{i+1} = r_{i+1} + b_i p_i$ $(1 \leq i \leq \ell)$.
Relation (i) is trivially satisfied if we use the unnatural formula for a_i. If we use the natural formula for a_i, then it follows immediately (cf. theorem 2.2.2) that $(r_i,p_{i-1}) = 0, (1 \leq i \leq \ell)$, and hence $(r_i,p_i) = (r_i,r_i) + b_{i-1}(r_i,p_{i-1}) = (r_i,r_i), (1 \leq i \leq \ell)$ which proves (i).
Relation (ii) is trivially satisfied if we use the unnatural formula for b_i. If we use the natural formula for b_i, then it follows immediately that $(p_i,Ap_{i-1}) = 0, (1 \leq i \leq \ell)$. Hence, in view of (i), $(r_{i+1},r_i) = (r_i,r_i) - a_i(Ap_i,r_i) = (r_i,r_i) - a_i((Ap_i,p_i) + - b_{i-1}(p_i,Ap_{i-1})) = 0, (1 \leq i \leq \ell)$, and consequently, again using (i), $(r_{i+1},Ap_i) = a_i^{-1}(r_{i+1},r_i - r_{i+1}) = - a_i^{-1}(r_{i+1},r_{i+1}) =$ $= - (r_{i+1},r_{i+1}) / (r_i,r_i)(p_i,Ap_i)$, $(1 \leq i \leq \ell)$, which proves (ii).
Using (i) we obtain $(r_{i+1},p_i) = (r_i,p_i) - a_i(p_i,Ap_i) =$ $= (r_i,p_i) - (r_i,r_i) = (r_i,p_i - r_i) = b_{i-1}(r_i,p_{i-1}), (1 \leq i \leq \ell)$, which proves (iii).
Using (i) and (ii) we obtain $(p_{i+1},Ap_i) = (r_{i+1},Ap_i) + b_i(p_i,Ap_i) =$ $= a_i^{-1}(r_{i+1},r_i - r_{i+1}) + a_i^{-1}(r_{i+1},r_{i+1}) = a_i^{-1}(r_{i+1},r_i) =$ $= a_i^{-1}((r_i,r_i) - a_i(Ap_i,r_i)) = (p_i,Ap_i) - (r_i,Ap_i) = (p_i - r_i,Ap_i) =$ $= b_{i-1}(p_{i-1},Ap_i), (1 \leq i \leq \ell)$, which proves (iv).
Using (ii) and (iii) we obtain $(r_{i+1},p_{i+1})(r_i,r_i) =$ $= (r_i,r_i)((r_{i+1},r_{i+1}) + b_i(r_{i+1},p_i)) = (r_{i+1},r_{i+1})((r_i,r_i) + (r_{i+1},p_i)) =$ $= (r_{i+1},r_{i+1})((r_i,r_i) + b_{i-1}(r_i,p_{i-1})) = (r_i,p_i)(r_{i+1},r_{i+1}), (1 \leq i \leq \ell)$ which proves (v).
Note that these properties hold after the first step. □

The propagation formulas (iii) and (iv) are already contained in the paper of Reid [71].
For natural a_i relations (iii) and (v) become trivial, since in that case $(r_{i+1},p_i) = 0, (0 \le i \le \ell)$, and $(r_{i+1},p_{i+1}) = (r_{i+1},r_{i+1}) + b_i(r_{i+1},p_i) = 0, (0 \le i \le \ell)$. For natural b_i relation (iv) is obvious, since then $(p_{i+1},Ap_i) = 0, (0 \le i \le \ell)$.
For the unnatural choices, (iii) and (iv) indicate the growth of the non-orthogonality of r_{i+1}, p_i and p_{i+1}, Ap_i, respectively, whereas (v) indicates the invariance of $(p_i,r_i) / (r_i,r_i), (1 \le i \le \ell)$.

With respect to the CGM we have seen that if the iterations terminate at step $\ell \ge 0$, then $r_\ell = 0$ as well as $p_\ell = 0$ and therefore the restriction $p_i \neq 0$ could in fact be left out in the stopping criterion. One may ask whether this also holds for the ISCGM's. A general result, valid for all variants, can be formulated if at least two steps are performed.

THEOREM 2. *If any version of the ISCGM's terminates at iteration step* $\ell \ge 2$ *and* $(r_1,p_1) \neq 0$, *then* $r_\ell = 0$ *as well as* $p_\ell = 0$.

PROOF. If $r_\ell = 0$, then $b_{\ell-1} = 0$ and hence $p_\ell = r_\ell + b_{\ell-1} p_{\ell-1} = 0$.
If $r_i \neq 0, (0 \le i \le \ell)$, then (v) can be written as $(r_{i+1},p_{i+1}) / (r_{i+1},r_{i+1}) = (r_i,p_i) / (r_i,r_i), (1 \le i \le \ell-1)$. Consequently, $(r_\ell,p_\ell) / (r_\ell,r_\ell) = (r_1,p_1) / (r_1,r_1) \neq 0$ and hence $p_\ell \neq 0$. □

If an ISCGM terminates after one step because $r_1 = 0$, then also $b_0 = 0$ and $p_1 = r_1 + b_0 p_0 = 0$.
For an ISCGM with the natural formula for a_i we have after one step $(r_1,p_1) = (r_1,r_1)$ and consequently, if $r_1 \neq 0$, then $p_1 \neq 0$ and $(r_1,p_1) \neq 0$. Thus, for the ISCGNNM and the ISCGNUM the assertion of theorem 2 also holds for $\ell = 1$ and the restriction $(r_1,p_1) \neq 0$ can be omitted.
For an ISCGM with the unnatural formula for a_i it is possible to have after one step $r_1 \neq 0$ and $p_1 = 0$. For example, if $A = \mathrm{diag}(1,2)$, $r_0^T = (-3,+3)$, $p_0^T = (1,1)$, then for both ISCGUNM and ISCGUUM one has $r_1^T = -(9,9)$, $p_1^T = (0,0)$. So for these two methods termination can occur at the first step, although $x_1 = \hat{x}$ does not hold. The case in which these two methods do not end after the first step and $(r_1,p_1) = 0$

is irrelevant, since from our results it follows that in that case $\|A^{-\frac{1}{2}} r_i\|$ increases monotonically as long as $p_i \neq 0$.

Another interesting question that arises is what can be said if the ISCGM's do not terminate. The following theorem expresses the relation between two successive natural errors.

THEOREM 3. *Let* $\{r_i\}$, $\{p_i\}$ *be generated by an ISCGM with arbitrary initial vectors* r_0, p_0. *Then we have for* $i \geq 1$

$$(1) \qquad \frac{\|A^{-\frac{1}{2}} r_{i+1}\|^2}{\|A^{-\frac{1}{2}} r_i\|^2} = 1 - \tau_i^2 \left(2\frac{(r_i,p_i)}{(r_i,r_i)} - 1\right),$$

where

$$(2) \qquad \tau_i := \|r_i\|^2 / (\|A^{-\frac{1}{2}} r_i\| \|A^{\frac{1}{2}} p_i\|) .$$

PROOF. From the relation $A^{-\frac{1}{2}} r_{i+1} = A^{-\frac{1}{2}} r_i - a_i A^{\frac{1}{2}} p_i$ we obtain, by taking squared norms at both sides,

$$(3) \qquad \|A^{-\frac{1}{2}} r_{i+1}\|^2 = \|A^{-\frac{1}{2}} r_i\|^2 - a_i (2(r_i,p_i) - a_i \|A^{\frac{1}{2}} p_i\|^2) .$$

Together with (i) of theorem 1 this yields for $i \geq 1$

$$\|A^{-\frac{1}{2}} r_{i+1}\|^2 = \|A^{-\frac{1}{2}} r_i\|^2 - \|r_i\|^2 / \|A^{\frac{1}{2}} p_i\|^2 (2(r_i,p_i) - (r_i,r_i)),$$

from which (1) readily follows. □

Note that for the CGM the parameter τ_i corresponds to the parameter γ_i defined by (2.2.8) and furthermore $2(r_i,p_i) / (r_i,r_i) - 1 = 1$. Hence, in that case theorem 3 coincides with theorem 2.2.2.

From theorem 3 we shall derive results concerning the step-wise linear convergence to zero of the natural error for ISCGM's, just like we did from theorem 2.2.2 for DM's.
We first determine a lower bound for τ_i.

LEMMA 4. *Let* $\{r_i\}$, $\{p_i\}$ *be generated by an ISCGM with arbitrary initial vectors* r_0 *and* p_0 *and let* τ_i *be defined by (2). Then we have for* $i \geq 1$

$$\tau_i \geq \frac{2\kappa^{\frac{1}{2}}}{(\kappa+1)(M+1)^{\frac{1}{2}}}, \tag{4}$$

where

$$M := (2\|A^{-1}\| / \|r_0\|^2) \max((p_1, Ap_0), 0) . \tag{5}$$

PROOF. From the relation $A^{\frac{1}{2}} p_{i+1} - b_i A^{\frac{1}{2}} p_i = A^{\frac{1}{2}} r_{i+1}$ we obtain, taking squared norms of both sides,

$$\|A^{\frac{1}{2}} p_{i+1}\|^2 - 2b_i (p_{i+1}, Ap_i) + \|b_i A^{\frac{1}{2}} p_i\|^2 = \|A^{\frac{1}{2}} r_{i+1}\|^2 , \ (i \geq 0) . \tag{6}$$

Using (iv) of theorem 1 we obtain recursively

$$b_i (p_{i+1}, Ap_i) = b_i b_{i-1} (p_i, Ap_{i-1}) = (b_i \cdots b_0)(p_1, Ap_0) , \ (i \geq 0) .$$

So, for unnatural b_i we conclude from (ii) of theorem 1 that one has $b_i (p_{i+1}, Ap_i) = \|r_{i+1}\|^2 (p_1, Ap_0) / \|r_0\|^2$ and consequently

$$2b_i (p_{i+1}, Ap_i) \leq M \|A^{\frac{1}{2}} r_{i+1}\|^2 , \ (i \geq 0) , \tag{7}$$

where

$$M := (2\|A^{-1}\| / \|r_0\|^2) \max((p_1, Ap_0), 0) . \tag{8}$$

For natural b_i we have $(p_{i+1}, Ap_i) = 0 = (p_1, Ap_0)$ and hence (7) also holds. By (6) and (7) we finally conclude

$$\|A^{\frac{1}{2}} p_{i+1}\| \leq (1+M)^{\frac{1}{2}} \|A^{\frac{1}{2}} r_{i+1}\| , \ (i \geq 0) , \tag{9}$$

and hence for all ISCGM's

$$\tau_i \geq \frac{\|r_i\|^2}{\|A^{-\frac{1}{2}} r_i\| \|A^{\frac{1}{2}} r_i\| (1+M)^{\frac{1}{2}}} \geq \frac{2\kappa^{\frac{1}{2}}}{(\kappa+1)(1+M)^{\frac{1}{2}}} \quad (i \geq 1) . \tag{10}$$ □

If we use the natural formula for a_i, then $(r_i, p_i) / (r_i, r_i) = 1$ $(i \geq 1)$. Combining this, (1) and (4) we obtain

THEOREM 5. *If* $\{r_i\}$ *is generated by the ISCGNNM or the ISCGNUM with arbitrary initial vectors* x_0, p_0, *then we have for all* $i \geq 1$

$$(11) \qquad \frac{\|A^{-\frac{1}{2}} r_{i+1}\|^2}{\|A^{-\frac{1}{2}} r_i\|^2} \leq 1 - \frac{4\kappa}{(M+1)(\kappa+1)^2},$$

where M *is defined by (8).* □

We thus conclude that for arbitrary initial vectors x_0, p_0, assuming the iterations do not terminate, the natural error converges step-wise linearly to zero with a convergence ratio no greater than $(1 - 4 / ((M+1)(\kappa+1)^2))^{\frac{1}{2}}$.
As far as the first step is concerned we have according to (3)

$$(12) \qquad \frac{\|A^{-\frac{1}{2}} r_1\|^2}{\|A^{-\frac{1}{2}} r_0\|^2} = 1 - \frac{(r_0,p_0)^2}{\|A^{-\frac{1}{2}} r_0\|^2 \|A^{\frac{1}{2}} p_0\|^2},$$

and hence $\|A^{-\frac{1}{2}} r_1\| \leq \|A^{-\frac{1}{2}} r_0\|$, with equality only if $(r_0,p_0) = 0$.

REMARK 6. Crowder and Wolfe [72] gave an example of the ISCGNNM in which the ratio $\|A^{-\frac{1}{2}} r_{i+1}\| / \|A^{-\frac{1}{2}} r_i\|$ is constant for all $i \geq 0$. Obviously, there exist initial vectors x_0, p_0 for which the convergence is only linear, so the finite termination property of the CGM does not hold in all cases for the ISCGNNM. □

If we use the unnatural formula for a_i, then, from (v) of theorem 1, it follows that $(r_i,p_i) / (r_i,r_i) = (r_1,p_1) / (r_1,r_1)$ $(i \geq 1)$. Combining this, (1) and (4) yields

THEOREM 7. *Let* $\{r_i\}$ *be generated by the ISCGUNM or the ISCGUUM with arbitrary initial vectors* x_0, p_0, *let* M *be defined by (8) and let* $K := 2(r_1,p_1) / (r_1,r_1) - 1$. *Then we have*

(i) *if* $K > 0$, *then for all* $i \geq 1$

$$\frac{\|A^{-\frac{1}{2}} r_{i+1}\|^2}{\|A^{-\frac{1}{2}} r_i\|^2} \leq 1 - \frac{4K\kappa}{(M+1)(\kappa+1)^2};$$

(ii) *if* $K = 0$, *then for all* $i \geq 1$

$$\frac{\|A^{-\frac{1}{2}} r_{i+1}\|^2}{\|A^{-\frac{1}{2}} r_i\|^2} = 1;$$

(iii) *If* $K < 0$, *then for all* $i \geq 1$

$$\frac{\|A^{-\frac{1}{2}} r_{i+1}\|^2}{\|A^{-\frac{1}{2}} r_i\|^2} \geq 1 - \frac{4K\kappa}{(M+1)(\kappa+1)^2} .$$

□

We conclude that for arbitrary initial vectors x_0, p_0 the natural error converges step-wise linearly after the first step with a convergence ratio no greater than $(1 - 4K\kappa / ((M+1)(\kappa+1)^2))^{\frac{1}{2}}$ if $K > 0$. If $K = 0$, then the natural error stays unaltered after the first step. If $K < 0$, then the natural error diverges at a linear rate after the first step.
For the first step we have, according to (3),

$$(13) \qquad \frac{\|A^{-\frac{1}{2}} r_1\|^2}{\|A^{-\frac{1}{2}} r_0\|^2} = 1 - \frac{\|r_0\|^4}{\|A^{-\frac{1}{2}} r_0\|^2 \|A^{\frac{1}{2}} p_0\|^2} \left(2 \frac{(r_0,p_0)}{(r_0,r_0)} - 1\right) ,$$

and hence $\|A^{-\frac{1}{2}} r_1\| \leq \|A^{-\frac{1}{2}} r_0\|$ if $2(r_0,p_0) / (r_0,r_0) - 1 \geq 0$.

REMARK 8. For the ISCGUUM we can derive a simple expression for K in terms of the initial vectors r_0 and p_0. We have

$$(14) \qquad (r_1,p_1) = (r_1,r_1) + b_0(r_1,p_0) =$$

$$= (r_1,r_1)(1 + (r_1,p_0) / (r_0,r_0)) =$$

$$= (r_1,r_1)(1 + \{(r_0,p_0) - a_0(p_0,Ap_0)\} / (r_0,r_0)) =$$

$$= (r_1,r_1)(r_0,p_0) / (r_0,r_0) ,$$

and hence $K = 2(r_0,p_0) / (r_0,r_0) - 1$. Therefore, $K \gtrless 0$ if $(r_0,p_0) \gtrless \frac{1}{2}(r_0,r_0)$.
In case one deals with the ISCGUNM, K is generally not equal to $2(r_0,p_0) / (r_0,r_0) - 1$. □

We end this section by mentioning a result of Powell [76] concerning the ISCGNNM.

THEOREM 9. *If* $\{r_i\}$ *is generated by the ISCGNNM with arbitrary initial vectors* x_0, p_0 *and if* $r_i \neq 0$ *for all* $0 \leq i \leq n+1$, *then one has*

(i) *There exists an* ℓ *satisfying* $2 \le \ell < n$ *such that* $p_1,\ldots,p_\ell$ *are mutually conjugate and* p_1 *and* $p_{\ell+1}$ *are not conjugate.*

(ii) *For all* $i \ge 0$ *the direction vectors* $p_{i+1},\ldots,p_{i+\ell+1}$ *are mutually conjugate, but* p_{i+1} *and* $p_{i+\ell+1}$ *are not conjugate.*

(iii) *Termination never occurs and convergence to the solution occurs at a linear rate.* □

REMARK 10. The condition $\ell \ge 2$ in theorem 9 immediately follows from the fact that for natural b_i one always has $(p_1, Ap_2) = 0$. It can be proved by induction that also for all $i \ge 0$ the vector $r_{i+\ell+1}$ is orthogonal to the vectors $p_{i+1},\ldots,p_{i+\ell}$. Consequently, if $p_1,\ldots,p_n$ are mutually conjugate (and hence linearly independent), then $r_{n+1} = 0$. This explains why $\ell < n$ holds in theorem 9. □

The most important conclusion that can be drawn from theorem 9 is that the ISCGNNM either terminates within $(n+1)$ iterations or convergence to the solution occurs at a linear rate. Powell [76] also shows that in the general case, where both r_0 and p_0 are arbitrary, the linear rate of convergence is usual. In our opinion this result did not get the attention in the literature it deserves. For instance, it implies that if during the CGM iterations r_i and p_i are computed exactly in all steps but one, then we may expect the convergence to be only linear.

5.3. *A one-round-off error analysis of the true residual independent start conjugate gradient method using the unnatural formula for* a_i *and the natural formula for* b_i (TRISCGUNM)

Examining the TRCGM in Chapter 4, we first derived some basic algebraic properties and next deduced the numerical analogues in order to obtain results for the numerical behavior. In this section we follow the same strategy and deduce numerical analogues of some algebraic properties derived in the foregoing section, in order to gain insight in the numerical behavior of other variants of the CGM.
We only consider the TRISCGUNM. The TRISCGNNM has been treated already in Chapter 4, whereas the TRISCGNUM is a special case of the DM's and

probably may be treated using the results of Chapter 2; the TRISCGUUM is believed to show the same peculiarities as the TRISCGUNM.
Furthermore, we restrict ourselves to a round-off error analysis where round-off only occurs at the computation of $\mathrm{fl}(Ax_i)$, since in all our previous analyses the errors occurring at that particular computation caused the largest discrepancy between algebraic and numerical properties.
Throughout this section we use, as before, the definitions

$$(1) \qquad \hat{r}_i := b - Ax_i \ , \quad r_i = b - \mathrm{fl}(Ax_i) \ , \quad \delta r_i = r_i - \hat{r}_i \ ,$$

$$(2) \qquad \varphi_i := \|A\|\|x_i\| \,/\, \|\hat{r}_i\| \ , \quad \psi_i := \|A^{\frac{1}{2}}\|\|x_i\| \,/\, \|A^{-\frac{1}{2}}\hat{r}_i\| \ .$$

The following theorem states the numerical analogues of (v) of theorem 5.2.1 and of theorem 5.2.3. These analogues form the basic results for the subsequent convergence considerations.

THEOREM 1. *Let* $\{x_i\}$, $\{p_i\}$ *be computed by the TRISCGUNM with arbitrary initial vectors* x_0, p_0 *and assume that only round-off occurs at the computation of* $\mathrm{fl}(Ax_i)$. *Then we have*

$$(i) \qquad \frac{\|A^{\frac{1}{2}}(\hat{x} - x_{i+1})\|^2}{\|A^{\frac{1}{2}}(\hat{x} - x_i)\|^2} = 1 - \tau_i^2\left(2\frac{(\hat{r}_i, p_i)}{(r_i, r_i)} - 1\right), \quad (i \geq 0) \ ,$$

where

$$\tau_i := \|r_i\|^2 \,/\, (\|A^{-\frac{1}{2}}\hat{r}_i\|\|A^{\frac{1}{2}}p_i\|) \ ;$$

$$(ii) \qquad \frac{(\hat{r}_{i+1}, p_{i+1})}{(r_{i+1}, r_{i+1})} = \frac{(\hat{r}_i, p_i)}{(r_i, r_i)}(1 + \mu_i) - \eta_i \ , \quad (i \geq 1) \ ,$$

where

$$\mu_i := (r_i + r_{i+1} \, , \, \delta r_{i+1} - \delta r_i) \,/\, \|r_{i+1}\|^2 \ ,$$

$$\eta_i := ((r_{i+1}, \delta r_{i+1}) + (r_i + r_{i+1} \, , \, \delta r_{i+1} - \delta r_i)) \,/\, \|r_{i+1}\|^2 \ .$$

PROOF. For the computation of r_i we have (cf. lemma 2.4.1)

$$(3) \qquad r_i = b - \mathrm{fl}(Ax_i) = b - Ax_i - E_i x_i = \hat{r}_i + \delta r_i \ ,$$

(4) $\qquad \tilde{r}_i = b - A x_i \ , \quad \delta r_i = - E_i x_i \ , \quad \|\delta r_i\| / \|\tilde{r}_i\| \le \varepsilon C_1 \varphi_i \ ,$

(5) $\qquad (\tilde{r}_i, r_i) = \|\tilde{r}_i\|^2 (1 + o(1)) = \|r_i\|^2 (1 + o(1)) \qquad [\varepsilon C_1 \varphi_i \to 0] \ .$

Further we have

(6) $\qquad a_i = \|r_i\|^2 / \|A^{\frac{1}{2}} p_i\|^2 \ , \quad x_{i+1} = x_i + a_i p_i \ .$

Consequently,

(7) $\qquad \|A^{\frac{1}{2}}(\hat{x} - x_{i+1})\|^2 = \|A^{\frac{1}{2}}(\hat{x} - x_i)\|^2 - 2a_i(\tilde{r}_i, p_i) + a_i^2 \|A^{\frac{1}{2}} p_i\|^2 =$

$$= \|A^{\frac{1}{2}}(\hat{x} - x_i)\|^2 - \|r_i\|^2 / \|A^{\frac{1}{2}} p_i\|^2 \, (2(\tilde{r}_i, p_i) - \|r_i\|^2) \ ,$$

from which (i) readily follows.
In proving (ii) we closely follow the lines of the proof of theorem 5.2.1. We have

(8) $$\begin{cases} b_i = - (r_{i+1}, Ap_i) / \|A^{\frac{1}{2}} p_i\|^2 \ , \quad p_{i+1} = r_{i+1} + b_i p_i \ , \\ (p_{i+1}, Ap_i) = 0 \ . \end{cases}$$

From (3) and (6) we obtain

(9) $\qquad \tilde{r}_{i+1} = \tilde{r}_i - a_i Ap_i \ , \quad r_{i+1} = r_i - a_i Ap_i + \delta r_i - \delta r_{i+1} \ .$

Consequently, for $i \ge 1$,

(10) $\qquad (r_{i+1}, r_i) = (r_i, r_i) - a_i (Ap_i, r_i) + (r_i, \delta r_i - \delta r_{i+1}) =$

$$= (r_i, r_i) - a_i ((Ap_i, p_i) - b_{i-1}(p_i, Ap_{i-1})) + (r_i, \delta r_i - \delta r_{i+1}) =$$

$$= (r_i, \delta r_i - \delta_{ri+1}) \ ,$$

and hence

(11) $\qquad (r_{i+1}, Ap_i) = a_i^{-1} (r_{i+1} \, , \, r_i - r_{i+1} + \delta r_i - \delta r_{i+1}) =$

$$= a_i^{-1} ((r_i + r_{i+1} \, , \, \delta r_i - \delta r_{i+1}) - \|r_{i+1}\|^2) \ .$$

From this we obtain the numerical analogue of relation (ii) of theorem 5.2.1, i.e.

$$(12)\qquad b_i = -\frac{(r_{i+1},Ap_i)}{(p_i,Ap_i)} = \frac{\|r_{i+1}\|^2}{\|r_i\|^2}(1+\mu_i) \qquad (i \geq 1)\ ,$$

where

$$(13)\qquad \mu_i := (r_i + r_{i+1}\ ,\ \delta r_i - \delta r_{i+1})\ /\ \|r_{i+1}\|^2\ .$$

From (9) we obtain the numerical analogue of relation (iii) of theorem 5.2.1, i.e.

$$(14)\qquad (f_{i+1},p_i) = (f_i,p_i) - a_i(p_i,Ap_i) = (f_i,p_i - r_i) - (r_i,\delta r_i) =$$

$$= b_{i-1}(f_i,p_{i-1}) - (r_i,\delta r_i)\ ,\quad (i \geq 1)\ .$$

Substituting the various equalities we finally find, for $i \geq 1$

$$(15)\qquad (f_{i+1},p_{i+1})(r_i,r_i) = \|r_i\|^2\{(f_{i+1},r_{i+1}) + b_i(f_{i+1},p_i)\} =$$

$$= \|r_i\|^2(\|r_{i+1}\|^2 - (\delta r_{i+1},r_i))$$

$$+ \|r_{i+1}\|^2(1+\mu_i)(b_{i-1}(f_i,p_{i-1}) - (r_i,\delta r_i)) =$$

$$= \|r_{i+1}\|^2\{(f_i,r_i) + (\delta r_i,r_i) + b_{i-1}(f_i,p_{i-1}) - (r_i,\delta r_i)\} +$$

$$- \|r_i\|^2(\delta r_{i+1},r_i) + \|r_{i+1}\|^2\mu_i((f_i,p_i - r_i) - (r_i,\delta r_i)) =$$

$$= \|r_{i+1}\|^2(f_i,p_i)(1+\mu_i) - \|r_i\|^2(\delta r_{i+1},r_i) - \|r_{i+1}\|^2\|r_i\|^2\mu_i\ .$$

This yields the numerical analogue of relation (v) of theorem 5.2.1, viz.

$$(16)\qquad \frac{(f_{i+1},p_i)}{(r_{i+1},r_{i+1})} = \frac{(f_i,p_i)}{(r_i,r_i)}(1+\mu_i) - \frac{(\delta r_{i+1},r_i)}{\|r_{i+1}\|^2} - \mu_i\ ,$$

which proves (ii). □

We finally discuss some conclusions that can be drawn from this theorem.

Since we assume that no round-off occurs at the computation of b_i and p_{i+1}, we have $(p_{i+1},Ap_i) = 0$ and (cf. formula (5.2.6)) $\|A^{\frac{1}{2}}p_{i+1}\|^2 + \|b_i A^{\frac{1}{2}}p_i\|^2 = \|A^{\frac{1}{2}}r_{i+1}\|^2$, and in particular $\|A^{\frac{1}{2}}p_{i+1}\| \leq \|A^{\frac{1}{2}}r_{i+1}\|$,

$(i \geq 0)$. Consequently, using (2.4.6) we obtain

$$(17) \qquad \tau_i \geq \frac{\|r_i\|^2}{\|A^{-\frac{1}{2}} \hat{r}_i\| \|A^{\frac{1}{2}} r_i\|} = \frac{\|r_i\|^2}{\|A^{-\frac{1}{2}} r_i\| \|A^{\frac{1}{2}} r_i\|} (1 + o(1)) \geq$$

$$\geq \frac{2\kappa^{\frac{1}{2}}}{(\kappa + 1)} (1 + o(1)) \qquad (i \geq 1) ,$$

under the restriction $\varepsilon C_1 \kappa^{\frac{1}{2}} \psi_i \to 0$. Hence, as far as the lower bound for τ_i is concerned, there are no complications as long as the natural error has not reached the level of the inherent natural error (cf. section 1.4). By (i) of theorem 1 it follows immediately that the natural error decreases as long as $(\hat{r}_i, p_i) / (r_i, r_i) > \frac{1}{2}$. In the case of exact computations, $(r_i, p_i) / (r_i, r_i)$ is constant for $i \geq 1$. In the presence of round-off at $fl(Ax_i)$ this invariance is disturbed as indicated by (ii) of theorem 1.

The parameter μ_i satisfies

$$(18) \qquad \mu_i = \frac{\|r_i\|}{\|r_{i+1}\|} \left(\frac{(r_i, \delta r_{i+1})}{\|r_i\| \|r_{i+1}\|} - \frac{(r_{i+1}, \delta r_i)}{\|r_{i+1}\| \|r_i\|} \right) +$$

$$- \frac{\|r_i\|^2}{\|r_{i+1}\|^2} \frac{(r_i, \delta r_i)}{\|r_i\|^2} + \frac{(r_{i+1}, \delta r_{i+1})}{\|r_{i+1}\|^2} .$$

Since we know that $\|\delta r_i\| / \|r_i\| \leq \varepsilon C_1 \varphi_i (1 + o(1))$, $[\varepsilon C_1 \varphi_i \to 0]$, we have the estimate

$$(19) \quad |\mu_i| \leq \left\{ \varepsilon C_1 (\varphi_{i+1} + \varphi_i) \frac{\|\hat{r}_i\|}{\|\hat{r}_{i+1}\|} + \varepsilon C_1 \varphi_i \frac{\|\hat{r}_i\|^2}{\|\hat{r}_{i+1}\|^2} + \varepsilon C_1 \varphi_{i+1} \right\} (1 + o(1)) ,$$

under the restriction $\varepsilon C_1 (\varphi_i + \varphi_{i+1}) \to 0$. A similar inequality holds for η_i. Consequently, as long as $\|\hat{r}_i\| / \|\hat{r}_{i+1}\|$ is of order unity and the residual has not reached the level corresponding to good-behavior, the algebraic invariance of $(\hat{r}_{i+1}, p_{i+1}) / (r_{i+1}, r_{i+1})$ is only slightly disturbed by round-off at each step. However, if $\|\hat{r}_i\| / \|\hat{r}_{i+1}\|$ is large or if the residual comes close to the level $\varepsilon \|A\| \|x_i\|$, then this invariance can be seriously affected by round-off. If $2(\hat{r}_i, p_i) / (r_i, r_i) - 1 > 0$, then in the next step possibly $2(\hat{r}_{i+1}, p_{i+1}) / (r_{i+1}, r_{i+1}) - 1 < 0$, which implies according to (i) of

theorem 1, that the natural error increases at the step from i + 1 to i + 2. Moreover, even if all round-off is excluded after step i + 1, then the natural error tends to infinity at a linear rate after that step. Thus, the stability property saying that if round-off is excluded after any iteration step, no matter how much round-off has occurred in the previous steps, then next the natural error converges step-wise linearly to zero, does certainly not hold for the TRISCGUNM. For the special case of the TRCGUNM ($p_0 := r_0$) we also have this non-stability phenomenon.

In the numerical experiments, reported by several authors, it turns out that for the TRCGUNM the quantity $\|\hat{r}_i\| / \|\hat{r}_{i+1}\|$ is in practice never extremely large, nor are these experiments continued until the natural error reaches the level of the inherent natural error. Therefore, $(\hat{r}_i, p_i) / (r_i, r_i)$ does not change very much during these experiments, which explains the satisfactory results.

REFERENCES

The figures between square brackets [AB] indicate the year 19AB of appearance of the paper or book on hand.

Axelsson, O. [74],
On preconditioning and convergence acceleration in sparse matrix problems.
CERN European Organization, Geneva (Report CERN 74-10).

Bollen, J.A.M. [79],
Round-off error analysis of the conjugate gradient algorithm.
Technological University Eindhoven (T.H.-Report 79-WSK-06).

Bruijn, N.G. de [61],
Asymptotic Methods in Analysis.
North-Holland, Amsterdam.

Crowder, H., R.S. Dembo and J.M. Mulvey [79],
On reporting computational experiments with mathematical software.
ACM Trans. Math. Software 5, 193-203.

Crowder, H. and P. Wolfe [72],
Linear convergence of the conjugate gradient method.
IBM J. Res. Dev. 16, 431-433.

Dekker, T.J. [79],
Correctness proofs and machine arithmetic.
Proc. IFIP TC2 Working Conference on Performance Evaluation of Numerical Software. North-Holland, Amsterdam; 31-43.

Hestenes, M.R. [80],
Conjugate Direction Methods in Optimization.
Springer-Verlag, New York.

Hestenes, M.R. and E. Stiefel [52],
Methods of conjugate gradients for solving linear systems.
NBS J. Res. 49, 409-436.

Jankowski, M. and H. Wózniakowski [77],
Iterative refinement implies numerical stability.
BIT 17, 303-311.

Kammerer, W.J. and M.Z. Nashed [72],
On the convergence of the conjugate gradient method for singular linear operator equations.
SIAM J. Numer. Anal. 9, 165-181.

Kershaw, D.S. [78],
The incomplete Cholesky-conjugate gradient method for the iterative solution of systems of linear equation.
J. Comp. Phys. 26, 43-65.

Luenberger, D.G. [73],
Introduction to Linear and Nonlinear Programming.
Addison-Wesley, Reading Mass.

Manteuffel, T.A. [80],
An incomplete factorization technique for positive definite linear systems.
Math. Comp. 34, 473-497.

Meijerink, J.A. and H.A. van der Vorst [77],
An iterative solution method for linear systems of which the coefficient matrix is a symmetric M-matrix.
Math. Comp. 31, 148-162.

Powell, M.J.D. [76],
Some convergence properties of the conjugate gradient method.
Math. Program. 11, 42-49.

Reich, E. [49],
On the convergence of the classical iterative method of solving linear simultaneous equations.
Ann. Math. Statist. 20, 448-451.

Reid, J.K. [71],
On the method of conjugate gradients for the solution of large sparse systems of linear equations.
Proc. Conference on Large Sparse Sets of Linear Equations.
Academic Press, New York; 231-254.

Stewart, G.W. [80],
The efficient generation of random orthogonal matrices with an application to condition estimators.
SIAM J. Numer. Anal. 17, 403-409.

Stoer, J. and R. Bulirsch [80],
Introduction to Numerical Analysis.
Springer-Verlag, New York.

Wilkinson, J.H. [65],
The Algebraic Eigenvalue Problem.
Clarendon Press, Oxford.

Wózniakowski, H. [77],
Numerical stability of the Chebyshev method for the solution of large linear systems.
Numer. Math. 28, 191-209.

Wózniakowski, H. [78],
Round-off error analysis of iterations for large linear systems.
Numer. Math. 30, 301-314.

Wózniakowski, H. [80],
Round-off error analysis of a new class of conjugate gradient algorithms.
Linear Algebra Appl. 29, 507-529.

INDEX

SAMENVATTING

Bij het numeriek oplossen van lineaire stelsels vergelijkingen $Ax = b$ maakt men onderscheid tussen direkte en iteratieve methoden. Een belangrijk verschil tussen direkte en iteratieve methoden is dat, in geval van exact rekenen, een direkte methode de oplossing $\hat{x} := A^{-1}b$ bepaalt in een eindig aantal rekenkundige bewerkingen, terwijl een iteratieve methode een oneindige rij benaderingen $\{x_i\}$ bepaalt die naar $\hat{x}$ convergeert en waarbij elke nieuwe benadering x_i uit zijn voorganger(s) wordt bepaald door een eindig aantal rekenkundige bewerkingen. Voor wat betreft het geheugengebruik van een rekenmachine zijn beide methoden sterk verschillend. Bij de meeste direkte methoden worden de elementen van de matrix A opgeslagen in een twee-dimensionale rij en de bij direkte methoden berekende decompositie matrices kunnen vervolgens (geheel of ten dele) worden opgeslagen in de geheugenplaatsen gebruikt voor de matrix A zelf. Bij iteratieve methoden kan het expliciet opslaan van de elementen van A worden vermeden; het is voldoende als men beschikt over een procedure om een gegeven vector x met de matrix A te vermenigvuldigen. Voor kleine stelsels is dit verschil van ondergeschikt belang maar voor grote ijle stelsels, waar de matrix A een relatief gering aantal niet-nul elementen bevat, zijn de decompositie matrices, corresponderend met direkte methoden, in het algemeen minder ijl dan A zelf, hetgeen leidt tot extra hoge eisen voor wat betreft de capaciteit van het rekenmachinegeheugen. Bij iteratieve methoden kan de ijlheid van de matrix A volledig worden uitgebuit door bij matrix maal vector vermenigvuldigingen de nul-elementen over te slaan.

Een ander aspect bij het oplossen van lineaire stelsels op de rekenmachine is de invloed van afrondfouten op de berekende oplossing. Voor de meeste gebruikte direkte methoden toonde Wilkinson [65] "goed-gedrag" en "numerieke stabiliteit" aan. Een methode met goed-gedrag berekent, indien uitgevoerd op een rekenmachine met relatieve machine-

precisie ε, een benadering x welke de exacte oplossing is van een lineair stelsel vergelijkingen met een gering verstoorde matrix A, d.w.z. van het stelsel $(A+E)x = b$ waarbij E van de orde $\varepsilon\|A\|$ is. Derhalve berekent een methode met goed-gedrag een benadering x waarvan de relatieve fout $\|\hat{x} - x\| / \|x\|$ hoogstens van de orde $\varepsilon\|A\|\|A^{-1}\| := \varepsilon\kappa$ is. Een methode die een benadering berekent met een relatieve fout hoogstens van de orde $\varepsilon\kappa$ noemen we numeriek stabiel. Een methode met goed-gedrag is dus numeriek stabiel maar het omgekeerde is niet noodzakelijk waar.

Voor iteratieve methoden bestaat er tot op heden slechts weinig literatuur betreffende de invloed van afrondfouten. Dit is ten dele te wijten aan het feit dat iteratieve methoden meer zelf-corrigerend leken te zijn dan direkte methoden en men verwachtte dat ze automatisch goed-gedrag zouden vertonen. Een andere reden is wellicht dat de gebruikers van iteratieve methoden in het algemeen meer geïnteresseerd zijn in het aantal iteraties dat nodig is om een benadering van de oplossing te berekenen met een bevredigende nauwkeurigheid dan in de maximaal haalbare nauwkeurigheid na eventueel vele iteraties. Desalniettemin is het opmerkelijk dat er nauwelijks iteratieve methoden zijn waarvoor een afrondfoutenanalyse bestaat. Wózniakowski is een van de eerste auteurs die zeer recent resultaten publiceerde omtrent goed-gedrag en numerieke stabiliteit van enkele iteratieve methoden.

Deze dissertatie levert een bijdrage aan dit nogal nieuwe onderzoekgebied. We bestuderen in Hoofdstuk 2 het gedrag van algemene descentmethoden in de aanwezigheid van afrondfouten. Binnen de verzameling van iteratieve methoden vormen de descentmethoden een belangrijke deelklasse; de meest gebruikte iteratieve methoden behoren hiertoe. In Hoofdstuk 3 en Hoofdstuk 4 besteden we speciale aandacht aan het numerieke gedrag van de gradiënt-methode en de geconjugeerde gradiëntmethode. De gradiënt-methode (ook wel methode van de steilste helling genoemd) is een descentmethode die, vooral vanuit theoretisch standpunt bezien, belangrijk is aangezien het een van de eenvoudigste niet lineaire iteratieve methoden is waarvoor een bevredigende analyse over het convergentiegedrag bestaat in het geval van exact rekenen. De geconjugeerde gradiënt-methode, die onafhankelijk door Hestenes en Stiefel [52] werd ontwikkeld, is zowel een direkte als een iteratieve methode. Het is een iteratieve methode omdat bij elke stap een betere benadering voor de oplossing wordt verkregen. Het is een direkte

methode omdat, bij exact rekenen, na hoogstens n stappen de oplossing $\hat{x}$ wordt bereikt, waarbij n de dimensie van het stelsel is. Het aanvankelijke enthousiasme over deze eindigheidseigenschap verdween al spoedig toen bleek dat in de aanwezigheid van afrondfouten de n-de iterand x_n vaak zelfs niet eens een redelijke benadering is voor de oplossing $\hat{x}$ van slecht geconditioneerde problemen. De methode werd nog slechts van academisch belang geacht, althans voor zover het het oplossen van lineaire stelsels betrof. Het artikel van Reid [71], waarin het iteratieve karakter van de methode werd benadrukt, bracht de methode opnieuw in de belangstelling en tegenwoordig staat de methode bekend als een iteratieve methode met zeer gunstige convergentie-eigenschappen voor sommige grote ijle stelsels met een niet te groot conditiegetal. Voor deze stelsels levert de methode vaak een redelijke benadering na veel minder dan n stappen.

Het toepassen van de algemene theorie uit Hoofdstuk 2 op de gradiënt-methode en de geconjugeerde gradiënt-methode levert een aantal resultaten voor wat betreft hun goed-gedrag en numerieke stabiliteit. Deze resultaten betreffen voornamelijk het uiteindelijke convergentiegedrag. Ter ondersteuning van de gedane uitspraken worden de uitkomsten van numerieke experimenten besproken.

In Hoofdstuk 5 komen varianten van de geconjugeerde gradiënt methode aan de orde.

CURRICULUM VITAE

De schrijver van dit proefschrift werd op 6 juni 1952 geboren te Geleen. In 1969 behaalde hij het diploma h.b.s.-B aan het Rombouts-college te Brunssum. Daarna studeerde hij voor wiskundig ingenieur aan de Technische Hogeschool te Eindhoven, waar hij onder leiding van Prof.dr.ir. M.L.J. Hautus zijn afstudeerwerk verrichtte op het gebied van de systeem- en regeltheorie. In 1975 behaalde hij (met lof) het doctoraal examen wiskunde. Na een half jaar wiskundeleraar te zijn geweest aan de Kantanshi Secondary School te Mufulira, Zambia, trad hij in 1976 als wetenschappelijk ambtenaar in dienst van de onderafdeling der Wiskunde van genoemde Hogeschool. Naast het geven van onderwijs in de numerieke wiskunde was zijn belangrijkste taak het doen van onderzoek op het gebied van de numerieke lineaire algebra, onder leiding van Prof.dr. G.W. Veltkamp, hetgeen leidde tot dit proefschrift.

STELLINGEN

1

De methode van de geconjugeerde richtingen voor het oplossen van een definiet stelsel $Ax = b$, waarbij de richtingvectoren $p_0,\ldots,p_{n-1}$ worden bepaald door Gram-Schmidt A-orthogonalisatie van de eenheidsvectoren $e_1,\ldots,e_n$, is niet alleen algebraïsch equivalent met de Cholesky methode maar levert ook numeriek even bevredigende resultaten.

2

Het voorwaarts sommeren van een reeks $\Sigma_{\ell=1}^{\infty} y_\ell$ van een linear convergente rij getallen ($|y_{\ell+1}| \leq L|y_\ell|$, $0 < L < 1$) is een numeriek instabiel proces. Deze instabiliteit kan worden opgeheven door voor elke optelling het algoritme van Møller te gebruiken.

Møller, O., Quasi-double precision in floating point addition. BIT, 5 (1965), 37-50, 251-255.

3

Zij A een $n \times n$ positief definiete matrix met orthonormale eigenvectoren $u_0,\ldots,u_{n-1}$ en positieve eigenwaarden $\lambda_0,\ldots,\lambda_{n-1}$. Als bij de methode van de geconjugeerde gradiënten de start residuvector (zie paragraaf 4.1 van dit proefschrift) voldoet aan

$$r_0 = \sum_{\ell=0}^{n-1} \alpha_\ell \varepsilon^\ell u_\ell ,$$

waarbij $\alpha_\ell \neq 0$ ($\ell = 0,\ldots,n-1$), dan geldt voor $k = 1,\ldots,n-1$ en $\varepsilon \to 0$, voor de exact berekende grootheden

$$a_{k-1} = \lambda_{k-1}^{-1} + O(\varepsilon^2) ,$$

$$r_k = \sum_{\ell=0}^{k-1} O(\varepsilon^{2k-\ell}) u_\ell + \sum_{\ell=k}^{n-1} \alpha_\ell [(1 - a_0\lambda_\ell) \cdots (1 - a_{k-1}\lambda_\ell) + O(\varepsilon^2)] \varepsilon^\ell u_\ell ,$$

$$b_{k-1} = O(\varepsilon^2) ,$$

$$p_k = \sum_{\ell=0}^{k-1} O(\varepsilon^{2k-\ell}) u_\ell +$$

$$+ \sum_{\ell=k}^{n-1} \alpha_\ell [(1 - a_0\lambda_\ell) \cdots (1 - a_{k-1}\lambda_\ell) + O(\varepsilon^2)] \varepsilon^\ell u_\ell \ .$$

4

Beschouw het volgende optimaliseringsprobleem: "Bepaal een stuksgewijs continue 2π-periodieke functie $u(t)$ met $|u(t)| \leq 1$ zodanig dat bij gegeven $\varepsilon > 0$ en gegeven tweemaal continu differentieerbare functie $f: \mathbb{R} \to \mathbb{R}$ de differentiaalvergelijking

$$\ddot{x} + \varepsilon f(x)\dot{x} + x = \varepsilon u$$

een 2π-periodieke oplossing heeft met maximale amplitude."
Als de functie f even is en

$$\int_0^x f(\zeta)d\zeta \to \infty \qquad (x \to \infty) \ ,$$

dan bestaat er voor iedere $\varepsilon > 0$ een oplossing $\bar{u}_\varepsilon(t)$ van bovenstaand probleem, waarbij $\bar{u}_\varepsilon(t)$ alleen de waarden $+1$ en -1 aanneemt. Bovendien nadert de afstand tussen de discontinuïteitspunten van $\bar{u}_\varepsilon(t)$ tot π als $\varepsilon \downarrow 0$.

5

Beschouw het optimaliseringsprobleem uit stelling 4 voor de differentiaalvergelijking

$$\ddot{x} + \varepsilon(f(x)\dot{x} + g(x)) + x = \varepsilon u \ ,$$

waarbij $\varepsilon > 0$ en $f,g: \mathbb{R} \to \mathbb{R}$ beide tweemaal continu differentieerbaar zijn.
Als de functie f even is en $f(x) \geq \delta > 0$ voor alle $x \in \mathbb{R}$ en als de functie g oneven is en $xg(x) \geq 0$ voor alle $x \in \mathbb{R}$ dan bestaat er voor iedere $\varepsilon > 0$ een oplossing $\bar{u}_\varepsilon(t)$ van het optimaliseringsprobleem waarbij $\bar{u}_\varepsilon(t)$ alleen de waarden $+1$ en -1 aanneemt. Bovendien nadert de afstand tussen de discontinuïteitspunten van $\bar{u}_\varepsilon(t)$ tot π als $\varepsilon \downarrow 0$.

Zij verder α,β: $\mathbb{R} \to \mathbb{R}$ gedefinieerd door

$$\alpha(r) := \int_0^{2\pi} \{g(r\cos t) - f(r\cos t)\, r \sin t\} \sin t \, dt ,$$

$$\beta(r) := \int_0^{2\pi} \{g(r\cos t) - f(r\cos t)\, r \sin t\} \cos t \, dt .$$

Indien $\bar{r} := \max \{r \in \mathbb{R} \mid \alpha^2(r) + \beta^2(r) \leq 16\}$ bestaat en $\frac{d}{dr}(\alpha^2(r) + \beta^2(r)) \neq 0$ voor $r = \bar{r}$, dan geldt voor de maximale amplitude A_ε

$$\lim_{\varepsilon \downarrow 0} A_\varepsilon = \bar{r} .$$

6

Het door Reid opgemerkte feit dat zijn numerieke resultaten bij toepassing van de methode van de geconjugeerde gradiënten voor het oplossen van uit discretisatie van partiële differentiaalvergelijkingen voortkomende definiete stelsels lineaire vergelijkingen nauwelijks verschillen bij het gebruik van recursieve of echte residuen geldt zeker niet in het algemeen voor slecht geconditioneerde stelsels. Het is af te raden om bij slecht geconditioneerde stelsels uitsluitend recursieve residuen te gebruiken.

Reid, J.K., On the method of conjugate gradients for the solution of large sparse systems of linear equations. Proc. Conference on Large Sparse Sets of Linear Equations. Academic Press, New York, 1971; 231-254.

Dit proefschrift, Hoofdstuk 4.

7

Het is ontoelaatbaar dat in een modern leerboek over numerieke lineaire algebra nauwelijks aandacht wordt besteed aan afrondfouten tengevolge van eindige machineprecisie.

Wait, R., The numerical solution of algebraic equations. John Wiley & Sons, New York, 1979.

8

Het is te betreuren dat de verzameling mensen die nadenkt over de economische gevolgen van werktijdverkorting slechts een kleine doorsnede heeft met de verzameling mensen die nadenkt over de maatschappelijke gevolgen van vrijetijdverlenging.

9

Een van de belangrijkste kenmerken van wetenschappelijk denken is dat het rekening houdt met het risico dat uitsluitend wetenschappelijk denken met zich meebrengt.

10

Er is een grote overeenkomst tussen politici en schepen: beiden maken het meeste lawaai als ze in de mist de koers kwijt zijn.

J.A.M. Bollen

2 december 1980

Wijze Boeken

als ik die dikke boeken zie van de geleerden
vol wijze dingen die zij allemaal beweerden
en ik zie de wereld om mij heen, dan moet ik vrezen
dat niemand ooit die wijze boeken heeft gelezen

Toon Hermans

Uit de gedichtenbundel "Fluiten naar de overkant" van Toon Hermans. Uitgeverij Elsevier Nederland B.V., Amsterdam/Brussel.